Quality over Quantity: Optimizing pulsar timing array analysis for stochastic and continuous gravitational wave signals

Lorenzo Speri[1], ★ Nataliya K. Porayko[2], Mikel Falxa[3], Siyuan Chen[4], Jonathan R. Gair[1], Alberto Sesana[5,6], Stephen R. Taylor[7]

[1]*Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Am Mühlenberg 1, 14476 Potsdam, Germany*
[2]*Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, 53121 Bonn, Germany*
[3]*Université de Paris, CNRS, Astroparticule et Cosmologie, 75013 Paris, France*
[4]*Kavli Institute for Astronomy and Astrophysics, Peking University, Beijing 100871, P. R. China*
[5] *Dipartimento di Fisica "G. Occhialini", Università degli Studi di Milano-Bicocca, Piazza della Scienza 3, I-20126 Milano, Italy*
[6] *INFN, Sezione di Milano-Bicocca, Piazza della Scienza 3, 20126 Milano, Italy*
[7]*Department of Physics & Astronomy, Vanderbilt University, 2301 Vanderbilt Place, Nashville, TN 37235, USA*

arXiv:2211.03201v1 [astro-ph.HE] 6 Nov 2022

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

The search for gravitational waves using Pulsar Timing Arrays (PTAs) is a computationally expensive complex analysis that involves source-specific noise studies. As more pulsars are added to the arrays, this stage of PTA analysis will become increasingly challenging. Therefore, optimizing the number of included pulsars is crucial to reduce the computational burden of data analysis. Here, we present a suite of methods to rank pulsars for use within the scope of PTA analysis. First, we use the maximization of the signal-to-noise ratio as a proxy to select pulsars. With this method, we target the detection of stochastic and continuous gravitational wave signals. Next, we present a ranking that minimizes the coupling between spatial correlation signatures, namely monopolar, dipolar, and Hellings & Downs correlations. Finally, we also explore how to combine these two methods. We test these approaches against mock data using frequentist and Bayesian hypothesis testing. For equal-noise pulsars, we find that an optimal selection leads to an increase in the log-Bayes factor two times steeper than a random selection for the hypothesis test of a gravitational wave background versus a common uncorrelated red noise process. For the same test but for a realistic EPTA dataset, a subset of 25 pulsars selected out of 40 can provide a log-likelihood ratio that is 89% of the total, implying that an optimally selected subset of pulsars can yield results comparable to those obtained from the whole array. We expect these selection methods to play a crucial role in future PTA data combinations.

Key words: Gravitational Waves – Data Analysis – Pulsars

1 INTRODUCTION

Pulsar Timing Array (PTA) experiments search for nanohertz-frequency gravitational waves (GWs) through induced shifts in radio-pulse arrival times from Galactic millisecond pulsars (Sazhin 1978; Detweiler 1979). The timing precision and regularity of the pulse times of arrival (TOAs) from these pulsars make them exquisite laboratories for studying a variety of astrophysical and fundamental physics phenomena (e.g., Verbiest et al. 2009). This includes GWs, which impart changes to the proper separation of Earth and the pulsar, causing pulses to arrive earlier or later than expected. These timing deviations are a function of the GW source characteristics, as well as the geometry of the GW source relative to the Earth-pulsar line-of-sight. Upon fitting a deterministic timing ephemeris (describing leading order behavior such as the rotational period, spindown rate, etc.) to a pulsar's TOAs, the remaining timing residuals can be analysed to search for the presence of GW signals amidst noise contributions. In a single pulsar's timing residuals, GW signals can easily be conflated with intrinsic pulsar noise effects (e.g., Shannon & Cordes 2010, and references therein) or even poorly understood artifacts of the ionized interstellar medium that radio pulses must traverse (e.g., Cordes & Shannon 2010, and references therein). But by constructing an array of pulsars, the fact that the GW-induced timing deviations are correlated between pulsars can be leveraged to distinguish it from uncorrelated astrophysical and instrumental noise processes (Foster & Backer 1990).

Several large collaborations have been monitoring ensembles of millisecond pulsars over long timing baselines in a bid to detect both a stochastic GW background (GWB) and individually-resolvable GW sources. These include the European Pulsar Timing Array (EPTA, Kramer & Champion 2013), the North American Nanohertz Observatory for Gravitational waves (NANOGrav, McLaughlin 2013), and the Parkes Pulsar Timing Array (PPTA, Manchester et al. 2013). Together with the more recently established Indian PTA (InPTA, Joshi et al. 2018), these collaborations constitute the International Pulsar Timing Array (IPTA, Verbiest et al. 2016; Perera et al. 2019a), which

★ lorenzo.speri@aei.mpg.de

© 2022 The Authors

aims to synthesize the aforementioned regional efforts to achieve more significant and rapid discoveries. Other recent timing efforts include the Chinese PTA (CPTA, Lee 2016), the MeerTIME program (Bailes et al. 2018) conducted at the MeerKAT telescope (Camilo et al. 2018), CHIME/Pulsar (Ng 2018), GMRT (Swarup 1990) and FAST (Jiang et al. 2019). Recent results from NANOGrav (Arzoumanian et al. 2020), the PPTA (Goncharov et al. 2021), the EPTA (Chen et al. 2021), and the IPTA (Antoniadis et al. 2022) all show strong evidence in favor of a common-spectrum process versus independent red-noise processes with Bayes factors of order $\sim 10^3 - 10^4$. These stochastic processes have similar spectral characteristics with estimated amplitudes around $A \sim 2 - 3 \times 10^{-15}$, and are all in broad agreement with expectations for a GWB generated by an astrophysical population of supermassive black-hole binaries (SMBHBs, e.g., Middleton et al. 2021). However, there is not yet significant evidence for the distinctive pattern of inter-pulsar correlations, known as the Hellings & Downs (HD) curve. In fact such evidence needs more time to emerge than the presence of a common process (Pol et al. 2021; Romano et al. 2021).

Building evidence for GW-induced inter-pulsar correlations requires many well-timed pulsars in order to forge effective pairings across different angular separations in order to trace out the HD pattern (Hellings & Downs 1983). This pattern is mostly quadrupolar in angular separation, with two zero crossings between 0° and 180°. Yet there are several issues associated with building an effective pulsar array for GW detection. (*i*) First, we are constrained by the Galactic distribution of millisecond pulsars, so there is little reason to consider array geometries that contradict this. (*ii*) Furthermore, if one were to only try to discover new pulsars that would maximize the significance of HD correlations, then the best strategy would be to survey close to the most sensitive pulsars. However, this would not trace the full pattern of this correlation curve, thereby severely inhibiting our ability to discriminate it from systematic noise processes that can also induce inter-pulsar correlations (Tiburzi et al. 2016). The latter include solar-system ephemeris errors that create dipolar correlations (Vallisneri et al. 2020; Champion et al. 2010; Guo et al. 2019; Caballero et al. 2018; Roebber 2019), and long-timescale systematics in time standards that create monopolar correlations (Hobbs et al. 2012, 2020). (*iii*) Finally, the next generation of radio facilities such as DSA-2000 (Hallinan et al. 2021), the Square Kilometre Array (SKA, Dewdney et al. 2009; Janssen et al. 2015), and the next-generation Very Large Array (ngVLA, Murphy et al. 2018) will lead to a torrent of new pulsars and observations. Future PTA data analysts will need metrics to judge which pulsars will most effectively characterize the GWB and resolve multiple individual SMBHBs out of this confusion background.

Therefore, exploring how to optimize the observing and analysis strategies of PTA experiments is crucial. In previous works, computational techniques to optimize the observational schedule (Lee et al. 2012; Lam 2018), and arrival-time precision as a function of radio frequency and bandwidth (Lam et al. 2018) have been investigated. In Roebber (2019), the author proposed a technique to optimize the disentangling between different spatial correlations and, therefore, to separate the signal due to GWs from that produced by clock or ephemeris errors. This paper also argued that such a method could be used to decide which pulsars should be included in PTAs. Beyond standard quality checks related to a pulsar's long-term timing stability, PTA searches aim to include as many pulsars as possible. However, a standard timing baseline cut of ~ 3 years is usually made in order to reduce the data volume while at the same time ensuring that all pulsars inform GW frequencies $\lesssim 10$ nHz where a GW background signal should be strongest.

In this work, we introduce for the first time a robust methodology for pulsar selection optimization in order to detect and characterize both the stochastic background and single continuous gravitational wave (CGW) sources. We develop ranking (or selection) methods to understand which pulsars contribute most to GW searches, where we target three key analyses: (*i*) detection of a GWB versus a Common Uncorrelated Red Noise (CURN) process, (*ii*) detection of a GWB versus Monopolar and Dipolar correlated signals, (*iii*) detection of CGW sources. These methods use statistical tools introduced in previous studies, making our methods easily implemented within established pipelines. Each method takes as input the intrinsic timing and noise properties of the whole pulsar array – which could be potentially provided by previous data releases– and outputs a ranked list of pulsars for a specified GW search.

This paper is organised as follows. We review the standard PTA statistical tools such as likelihood and frequentist and Bayesian hypothesis testing in Section 2.1 and 2.2. These tools are used to test the performance of the ranking methods introduced in Section 2.3 and 2.4. In particular, the ranking method based on signal-to-noise ratio (SNR) maximization is presented in 2.3.1, and the one aimed at disentangling different spatial correlations in 2.3.2. In Section 2.4 we develop a selection method that targets the search for continuous gravitational wave signals. The results are presented in Section 3 where the selection methods are tested using simulated datasets with increasing level of noise complexity. We conclude with our expectations for future investigations in Section 4.

2 METHODS

2.1 Pulsar Timing Array Likelihood

In this section we introduce the marginalized PTA likelihood which is ultimately the fundamental tool for the statistical analysis of PTA data (van Haasteren et al. 2009). We predominantly follow the "Gaussian process" treatment described in details in van Haasteren & Vallisneri (2014); Arzoumanian et al. (2016). The TOAs for each pulsar can be represented by a vector $\vec{t}$ of length $N_{\rm TOA}$. $\vec{t}$ can be written as a sum of a deterministic and a stochastic component: $\vec{t} = \vec{t}_{\rm det} + \vec{t}_{\rm sto}$.

The deterministic part comprises the so-called timing model which depends on a set of timing parameters $\vec{\beta}$. The timing model describes the intrinsic spin evolution of a source, propagation effects as well as time delays associated with the relative motion of a source and the Earth and kinematic and light propagation effects in the binary system (see e.g. Lorimer & Kramer 2012). The initial estimate of the m timing model parameters $\vec{\beta}_0$ is obtained using the minimization of the sum of the squares of the residuals $\delta\vec{t} = \vec{t} - \vec{t}_{\rm det}(\vec{\beta})$. This least-square linear fit to the timing model, which is performed using the `TEMPO2` software (Hobbs et al. 2006; Edwards et al. 2006), is equivalent to likelihood maximization when assuming Gaussian white noise errors. In reality the stochastic noise component is dominated by colored noises. Assuming that the initial estimate of the timing parameters $\vec{\beta}_0$ obtained from `TEMPO2` does not differ significantly from the final estimate $\vec{\beta}_f$ obtained from a full analysis that includes more sophisticated stochastic noise modelling, the timing model can be approximated to impact the timing residuals linearly via the term $\mathbf{M}\vec{\epsilon}$, where $\vec{\epsilon} = \vec{\beta}_f - \vec{\beta}_0$ and $\mathbf{M}$ is an $N_{\rm TOA} \times m$ *design matrix* (van Haasteren et al. 2009).

The correlated components of the stochastic piece $\vec{t}_{\rm sto}$ are modelled in terms of a Fourier decomposition (Lentati et al. 2013). In practice, the analysis focuses on the noise with dominant power at lower frequencies, so that only a finite number of Fourier components N_f

are used. In this case the signal can be written in a matrix form of the type $\mathbf{F}\vec{a}$, where the vector $\vec{a}$ of length $2N_{\rm freqs}$ contains the Fourier coefficients, whereas the $N_{\rm TOA}\times 2N_{\rm freqs}$ matrix $\mathbf{F}$ is constructed with alternating columns of sines and cosines evaluated at the TOAs of each pulsar. The base sampling frequency is given by the inverse of the observation timespan of the entire pulsar timing array, $1/T$.

The influence of white-noise on the timing residuals is described by the $N_{\rm TOA}\times N_{\rm TOA}$ white noise covariance matrix $\mathbf{N}$. Finally, the noise-mitigated timing residuals $\vec{r}$, which is our best approximation to the white noise $\vec{n}$ for each pulsar can be written in a compact form as a function of the input residuals $\delta\vec{t}$:

$$\vec{r} = \delta\vec{t} - \mathbf{T}\vec{b} \qquad \mathbf{T} = \begin{bmatrix}\mathbf{M} & \mathbf{F}\end{bmatrix} \qquad \vec{b} = \begin{bmatrix}\vec{\epsilon}\\ \vec{a}\end{bmatrix}, \tag{1}$$

and the likelihood is given by:

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

The prior covariance and corresponding Gaussian prior on the coefficients $\vec{b}$ are written as:

$$\mathbf{B} = \begin{bmatrix}\infty & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\phi}\end{bmatrix} \qquad p(\vec{b}|\vec{\phi}) = \frac{\exp\left\{-\frac{1}{2}\vec{b}^{\rm T}\,\mathbf{B}^{-1}\,\vec{b}\right\}}{\sqrt{2\pi\det\{\mathbf{B}\}}}\,, \tag{3}$$

so that the timing model piece of $\vec{b}$ is a uniform unconstrained prior on the timing model parameters $\vec{\epsilon}$, and the spectrum of all low-frequency processes enters in the variance $\boldsymbol{\phi}$ as:

$$\phi_{(ai),(bj)} = \Gamma_{ab}S_i\delta_{ij} + P_{ai}\delta_{ab}\delta_{ij}\,. \tag{4}$$

where the intrinsic low-frequency ("spin-noise") spectrum of pulsar a at the $i^{\rm th}$ sampling frequency is represented by P_{ai}, and the GWB spectrum, which is common to all pulsars, is given by S_i. Both of these processes can be modelled with a power-law functional form:

$$P_{ai} = \frac{A_a^2}{12\,\pi^2 T}\left(\frac{f_i}{\rm yr^{-1}}\right)^{-\gamma_a}{\rm yr}^2\,. \tag{5}$$

The reduction in correlated power due to the spatial separation of the pulsars is described by the overlap reduction function (ORF) Γ_{ab} between pulsars a and b. For an isotropic and stochastic GWB, the ORF is described by the HD curve (Hellings & Downs 1983), which depends only on the angular pulsar separation. If we group all the red noise and GWB spectral hyper-parameters into the vector $\vec{\eta}$ we can obtain the likelihood of the full PTA array (van Haasteren & Vallisneri 2014), marginalized over $\vec{b}$:

$$\begin{aligned}\mathcal{L}(\vec{\eta}) = p(\{\delta\vec{t}\}|\vec{\eta}) &= \int\prod_{a=1}^{N} p(\delta\vec{t}_a|\vec{b}_a)\times p(\{\vec{b}\}|\vec{\eta})d^N\vec{b}\,,\\ \ln\mathcal{L} &= -\frac{1}{2}\left[\delta\vec{t}^{\rm T}\,\mathbf{C}^{-1}\,\delta\vec{t} + {\rm Tr}\ln 2\pi\mathbf{C}\right],\end{aligned} \tag{6}$$

where $\mathbf{C} = \mathbf{N} + \mathbf{TBT}^{\rm T}$, and N is the total number of pulsars. A deterministic signal $\vec{s}(\vec{\theta})$ can be incorporated in the modelling by performing the following replacement $\delta\vec{t} \to \delta\vec{t} - \vec{s}(\vec{\theta})$. More details on likelihood construction and handling correlated noise processes in pulsar timing analysis can be found in e.g., van Haasteren & Levin (2013); Arzoumanian et al. (2015, 2016); Taylor (2021).

Having constructed the PTA marginalized likelihood, we can estimate the parameters $\vec{\eta}$. In frequentist inference, the true model parameters are considered to be fixed $\eta_{\rm True}$, and are estimated by maximizing the likelihood to obtain the maximum likelihood estimator (MLE), $\vec{\eta}_{\rm MLE}$. In Bayesian inference, model parameters are no longer regarded as fixed, but are themselves random variables. The probability distribution of the parameter values before the data acquisition (the prior distribution $p(\vec{\eta})$) is updated to a probability distribution after the data incorporation (the posterior distribution $p(\vec{\eta}|\delta\vec{t})$) through the likelihood of the observed data $\mathcal{L}(\delta\vec{t}|\vec{\eta})$. With several intrinsic noise parameters per pulsar, in addition to several global parameters describing the GW signal, the posterior distribution can be as high as $\mathcal{O}(100)$-dimensional. Thus, it is typically explored and sampled numerically using Markov chain Monte Carlo (MCMC) techniques.

2.2 Hypothesis testing

The essential step of the PTA analysis is testing whether the observed data are consistent with our expectations, e.g. the presence of a GW signal or its absence. Therefore, we use hypothesis testing to investigate if the data provides sufficient evidence for one hypothesis $\mathcal{H}_1$ with respect to another one $\mathcal{H}_2$. The tools developed in this section will be used in Sec. 3 as a proxy to test our selection methods.

If we adopt a frequentist approach, we can maximize the likelihood under each hypothesis to find the MLE for the parameters, i.e., $\vec{\eta}_{\rm MLE\,1} = \max_{\vec{\eta}}\ln\mathcal{L}(\vec{\eta}|\mathcal{H}_1)$ and analogously for $\mathcal{H}_2$. Then, the log-likelihood ratio defined as:

$$\ln\Lambda = \ln\mathcal{L}(\vec{\eta}_{\rm MLE\,1}|\mathcal{H}_1) - \ln\mathcal{L}(\vec{\eta}_{\rm MLE\,2}|\mathcal{H}_2) \tag{7}$$

can be used to test whether our data supports hypothesis $\mathcal{H}_1$ with respect to $\mathcal{H}_2$. Roughly speaking, a large value of $\ln\Lambda$ indicates a stronger support for $\mathcal{H}_1$ with respect to $\mathcal{H}_2$. Therefore, we can use $\ln\Lambda$ to assess if an optimally selected subset of pulsars supports our expectations as much as the full dataset.

To statistically quantify the significance of a measured log-likelihood value it is necessary to create multiple realizations of the data under the reference hypothesis $\mathcal{H}_2$. For each realization, we must then evaluate the log-likelihood ratio to obtain a distribution of $\ln\Lambda$ under the reference hypothesis. This distribution can be used to calculate the p−value of the measured log-likelihood. This approach is only viable if our ranking methods are tested on mock dataset realizations.

In reality, we cannot generate multiple realizations of the data because we do not have access to the true parameters and data generation process. We have access only to the most likely values of such parameters from previous data releases. Therefore, we can use those for the data generation of the reference hypothesis. By evaluating the p-value for the real dataset, we estimate the significance of such an experiment and check the consistency of our assumptions on the data generation process. Similar tests are extensively used in PTA analysis (see sky scrambles, phase shifts, and optimal statistic analysis, e.g., Chamberlin et al. 2015; Cornish & Sampson 2016; Taylor et al. 2017). We evaluate this procedure as a consistency check for hypothesis testing of a realistic PTA analysis in Sec. 3.2.

In Bayesian statistics, the Bayes Factor (BF)

$${\rm BF} = \frac{\int {\rm d}\vec{\eta}\mathcal{L}(\delta\vec{t}\,|\vec{\eta},\mathcal{H}_1)p(\vec{\eta},\mathcal{H}_1)}{\int {\rm d}\vec{\eta}\mathcal{L}(\delta\vec{t}\,|\vec{\eta},\mathcal{H}_2)p(\vec{\eta},\mathcal{H}_2)} \tag{8}$$

is used to assess which model is favored by the observations, assuming that the two models are equally probable a priori. A "rule of thumb" for interpreting Bayes' factors is presented in Kass & Raftery (1995), where BF > 20 is considered strong evidence for $\mathcal{H}_1$.[1]

[1] Alternatively, the distribution of the Bayes factor can be computed under the null hypothesis and used, in a frequentist way, to produce a mapping be-

If the posterior volumes of the two hypotheses are approximately the same, then the log-likelihood ratio at the MLE is approximately equal to the log-Bayes factor, i.e. $\ln \mathrm{BF} \approx \ln \Lambda$ (Romano & Cornish 2017; Pol et al. 2021).

In practice, BFs are widely used to perform robust statistical analysis, including hypothesis testing, when processing real PTA datasets. In this work, full Bayesian inference is only used for computationally feasible analysis of simplified datasets. For the realistic mock datasets which require more sophisticated noise modelling, we utilise the log-likelihood ratio test as it requires fewer computational resources.

2.3 Ranking pulsars for stochastic signal searches

One of the primary goals of the current PTA experiments is to detect the stochastic GWB from a population of SMBHBs. An isotropic GWB manifests itself as a long timescale, low-frequency (or red) common signal across the pulsars in a PTA. This common signal is characterized by the common spectrum and the inter-pulsar spatial correlations. The distinctive signature of the gravitational nature lies in this correlation which depends only on the pulsar's angular separation and has an expectation value given by the HD curve (Hellings & Downs 1983). Current experiments found strong evidence for the presence of a common red noise signal. While such a signal could potentially represent the expected GWB from SMBHBs, there is not yet strong evidence for either HD or other alternative angular correlations.

Motivated by these latest results, in Sec. 2.3.1 we design a method to identify the optimal subset of pulsars for increasing the confidence in the detection of an HD correlation, whereas in Sec. 2.3.2 we use the decoupling formalism to find the best subset of pulsars for distinguishing this correlation from alternative hypotheses. Recent work has cautioned that GWB upper limits can be biased and even lie below the true value when small ($\lesssim 20$) combinations of pulsars are analyzed (Johnson et al. 2022). Our work here is likely immune from such unwanted effects for several reasons: (*i*) the field of PTAs has moved beyond the regime of setting upper limits, to now estimating the statistical parameters of a common process and performing model selection on spatial correlations; and (*ii*) our metrics here are based on the detectability and discrimination of stochastic processes, rather than upper limits.

2.3.1 *Spatially correlated signal-to-noise ratio maximization*

As previously mentioned, the target signal is described by a correlated red noise process $S(f)$ with spatial correlations Γ_{ab}. An optimal subset of pulsars can be constructed based on an optimal statistic that maximizes the detection probability at a fixed false alarm probability for this specific case. As a proxy for this, it is convenient to consider statistics that maximize the signal-to-noise ratio (SNR), which is the ratio of the expected value of a statistic in the presence of a signal, μ_1, to its standard deviation. The standard deviation can either be computed in the absence of a signal, σ_0, or in the presence of a signal, σ_1. In Rosado et al. (2015), the authors introduce two statistics: the A-statistic constructed by maximizing μ_1/σ_0 and the B-statistic constructed by maximizing μ_1/σ_1. This procedure leads to the respective SNR definitions:

$$\mathrm{SNR}_{\mathrm{A}}^2 = 2 \sum_{a>b} \int \frac{\Gamma_{ab}^2 \, S^2(f) \, T_{ab}}{P_a(f) P_b(f)} \mathrm{d}f \,, \tag{9}$$

$$\mathrm{SNR}_{\mathrm{B}}^2 = 2 \sum_{a>b} \int \frac{\Gamma_{ab}^2 \, S^2(f) \, T_{ab}}{[P_a(f) + S(f)][P_b(f) + S(f)] + S^2(f)\Gamma_{ab}^2} \mathrm{d}f \,. \tag{10}$$

We use these quantities as a proxy to identify the best subset of pulsars from the full array. $\mathrm{SNR}_{\mathrm{A}}$ and $\mathrm{SNR}_{\mathrm{B}}$ are obtained under the expectation value of the true hypothesis and do not depend on the timing residuals but only on the general properties of the pulsars' red and white noises. In Eq. (9-10), the sum is over the pulsar pair a, b, with $a > b$ and T_{ab} is the overlapping time of observation of the a, b arrays. The term $P_a(f)$ represents the sum of the intrinsic noise processes of pulsar a such as red noise, white noise, etc. :

$$\begin{aligned} P_a(f) &= P_{\mathrm{rn}} + P_{\mathrm{wn}} + \ldots \\ &= \frac{A_a^2}{12\pi^2} \left(\frac{f}{\mathrm{yr}^{-1}} \right)^{-\gamma_a} \mathrm{yr}^3 + 2\sigma^2 \Delta t + \ldots \end{aligned} \tag{11}$$

where σ is the root-mean-square (RMS) error and Δt is the cadence of the TOAs. We also assume that the correlated noise process $S(f)$ can be described by a power-law functional form.

As pointed out in Rosado et al. (2015), the $\mathrm{SNR}_{\mathrm{B}}$ is more robust in the strong-signal regime. In fact, as we can see from Eq. (9-10), one of the useful differences with respect to the other statistic is that $\mathrm{SNR}_{\mathrm{B}}$ does not diverge for $S \gg P_a$. The $\mathrm{SNR}_{\mathrm{B}}$ is very similar to the so-called optimal statistic SNR presented in (Siemens et al. 2013; Chamberlin et al. 2015), however the last term in the denominator of $\mathrm{SNR}_{\mathrm{B}}$ is missing in those studies.

One downside of using the $\mathrm{SNR}_{\mathrm{B}}$ of Eq. (10) is that it assumes the amplitude and slope of $S(f)$ to be known. Since we have constraints on such parameters from the current PTA experiments, we can assume these to be known and use them to calculate the SNR. We will later show that the selection procedure using this SNR is not strongly affected by the variations of these quantities when estimated over noise realizations. The $\mathrm{SNR}_{\mathrm{A}}$ definition has the advantage that the amplitude factors out and therefore its maximization is not affected by the choice of A_{GWB}.

In theory, we would need to compare the SNRs with all possible combinations of subsets of pulsars from the whole array. Since this is computationally intractable in practice, we start from a few fiducial pulsars and add pulsars one by one until we reach the desired level of SNR. We will see in Sec. 3.2 that this "one-by-one" implementation of SNR-maximization performs very well, reaching a high proportion of the full data set BF with only a small selection of pulsars. The small improvement that might be achieved from an exhaustive search of all possible pulsar subsets is unlikely to be worth the considerable increase in computational cost.

If we set the spatial correlation Γ_{ab} to be the HD correlation, we can use these SNRs to rank pulsars and increase the detection probability of a GWB. Therefore, the SNR-maximization selection method introduced here aims at providing the best pulsars for the hypothesis test of an HD correlation (hypothesis $\mathcal{H}_1$) versus a CURN (hypothesis $\mathcal{H}_2$).

2.3.2 *Maximization of the decoupling between spatial correlations*

An unambiguous detection of a GWB relies on the characterization of the angular correlation between pulsars. In order to claim a detection, PTA experiments must provide strong evidence that an HD

tween p-values and Bayes factors. However, this approach is computationally expensive.

correlation is clearly identified in the data. However, the detection of a GWB is complicated by the presence of other types of correlated signals. Specifically, errors in clocks used to calibrate timing residuals, and poorly determined solar system ephemeris induce large-scale correlations between pulsars and can mimic the effects of a GWB. The irregularities in terrestrial time standards produce signals with monopolar spatial correlation (Hobbs et al. 2012, 2020), while emphemeris errors can result in dipolar signals (Champion et al. 2010; Tiburzi et al. 2016). In order to provide an optimal separation of the quadrupole GWB signal from those produced by clock or ephemeris errors, Roebber (2019) proposed a method to minimize the leakage between spatially correlated noises. We briefly review this formalism here.

The degree to which power from one spatial harmonic can leak into another one can be quantified by the coupling matrix (Peebles 1973; Gorski et al. 1994; Wandelt et al. 2001; Hivon et al. 2002; Mortlock et al. 2002; Efstathiou 2004):

$$K_{(lm),(lm)'} = \int Y_{lm}(\Omega) W(\Omega) Y_{(lm)'}(\Omega) \mathrm{d}\Omega\,, \tag{12}$$

where Y_{lm} is the spherical harmonic of degree l and order m, $W(\Omega)$ is the window function, and the integral is performed over all sky directions, Ω. The Coupling Matrix formalism can be directly applied to the pulsar selection problem. Within the PTA framework, a GWB has maximum power at $l = 2$, while clock noise and ephemeris noise appear at $l = 0$ and $l = 1$, respectively. Therefore, the coupling matrix elements with l from 0 to 2 are of interest for the problem of mode disentangling. While forming an orthonormal basis in the case of continuous coverage ($W(\Omega) = 1$ everywhere on the sky), the coupling matrix loses its orthogonality when the sampling of the sky becomes discrete, resulting in non-zero off-diagonal elements in $K_{(lm),(lm)'}$.

In the context of PTA analysis, the window function is given by the Kronecker-delta modulated by the individual weights w of pulsars placed at sky positions $\hat{p}_a$:

$$W(\Omega) = \sum_a w^a \delta(\Omega - \hat{p}_a)\,. \tag{13}$$

In the case of all-equal pulsars, the choice of the weighting function is straightforward: $w^a = 1$ for all pulsars. However, the problem becomes less trivial when each pulsar has different properties (in terms of RMS residuals, observation time, intrinsic red noise, etc.). Roebber (2019) suggests to use the inverse of the RMS of a source, $1/\sigma_a^2$, as weights, to account for the relative sensitivity of different pulsars in an array. In order to additionally account for the coloured noise in an array, we will use $\mathrm{SNR_A} \sim 1/\sigma_a^2$ as weights in the coupling matrix formula, where $\mathrm{SNR_A}$ is defined using the self-term ($a = b$) of Eq (9). Although this is a natural choice, it is worth noting that the optimal choice of the weighting function for the coupling matrix construction does not have a unique solution and in some cases requires a heuristic approach (Efstathiou 2004). As shown in Appendix A, for the two realistic mock datasets described in Sec. 3, an $\mathrm{SNR_A^4}$-weighting on average performs better than the other types of weighting function considered. However, in order to provide a definitive solution to the problem of weight selection, extensive testing on more diversified samples of mock datasets is required, which we leave for future work.

The level at which one mode leaks to another is estimated via the ratio of minimum and maximum eigenvalues $\lambda_{\min}/\lambda_{\max}$ of $K_{(lm),(lm)'}$, which is 1 when the coupling matrix is diagonal and drops to 0 when the coupling matrix is ill-defined. Since we are mainly interested in decoupling the spherical harmonics with different l, we can average Eq. (12) over m. Thus, the final expression for the coupling matrix is Efstathiou (2004):

$$M_{l,l'} = \frac{1}{(2l+1)(2l'+1)} \sum_{m,m'} K_{(l,m)(l',m')}\,. \tag{14}$$

We construct the pulsar ranking list by selecting those that lead to the largest eigenvalue ratio $\delta_\lambda = \lambda_{\min}/\lambda_{\max}$ of the $M_{l,l'}$ matrix. The Coupling Matrix selection method introduced here aims at providing the best pulsars for the hypothesis test of an HD correlation (hypothesis $\mathcal{H}_1$) versus the presence of all three signals in the data, namely common uncorrelated, monopolar and dipolar spatially correlated red noise processes (hypothesis $\mathcal{H}_2$). As pointed out in Roebber (2019), the minimum number of pulsars required to disentangle up to $l_{\max}$ is $\sum_{l=0}^{l_{\max}}(2l+1) = (l_{\max}+1)^2$, which is 9 for $l = 2$. After averaging over m, the coupling matrix $M_{l,l'}$ is well-defined when the number of pulsars is ≥ 3, meaning that at least three pulsars are required to resolve the spatial modes up to the quadrupole. Therefore, when the Coupling Matrix formalism is applied to realistic datasets, in order to avoid ambiguity, the first three pulsars in the ranking are fixed to those with the highest self-SNR.

2.3.3 *Chimera method: combining SNR- and decoupling-maximization algorithms*

The Coupling Matrix selection method is aimed at disentangling different types of correlations, while the total SNR maximization is disregarded. Therefore, the Coupling Matrix can only be used as a complementary scheme for array optimization, especially, for an array of pulsars in mixed SNR regime[2]. Here we propose a new selection method that combines the merits of both the Coupling Matrix and SNR maximization: hereafter the "Chimera"[3] method. The basic idea is to add a new pulsar to a subset, so that the HD-SNR is maximized along with the decoupling power. One of the possible norms that satisfies the latter requirement is the multiplication of the relevant scores of both methods, i.e. SNR and eigenvalue ratio:

$$\mathrm{SC_{Chimera}} = \mathrm{SNR_B^2}\,\delta_\lambda\,. \tag{15}$$

Note that the ranking of pulsars within the Chimera approach is purely heuristic and the score that we offer in Eq. (15) is one of many possible choices. As in the case of the Coupling Matrix, the first three pulsars are selected according to the highest self-SNR, while the following ones are picked so that the score in Eq. (15) is maximized.

For reference, in Figure 1 we show how the three different selection methods for GWB searches pick equal-noise pulsars on the sky. The full array is composed of 200 pulsars uniformly distributed over the sky and the number of selected pulsars is 25. The first pulsar was randomly selected and the following ones were picked according to the different selection methods. The SNR depends on Γ_{ab}^2 and so the SNR-maximization method tends to add pulsars where the HD correlation is largest, i.e., with $\theta_{ab} = 0°$ and $180°$.

[2] This means that the vast majority of pulsars in an array are in the weak signal regime (Siemens et al. 2013) and only a few sources actually contain the detectable signal. In this case, the latter are expected to contribute a significant fraction of the whole array sensitivity, while the addition of the former sources is largely irrelevant.

[3] The name was inspired by the mythological creature composed of different animal parts. Homer describes it as follows in the Iliad: "she was of divine stock, not of men, in the fore part a lion, in the hinder a serpent, and in the midst a goat, breathing forth in terrible wise the might of blazing fire." Homer & Lattimore (2005)

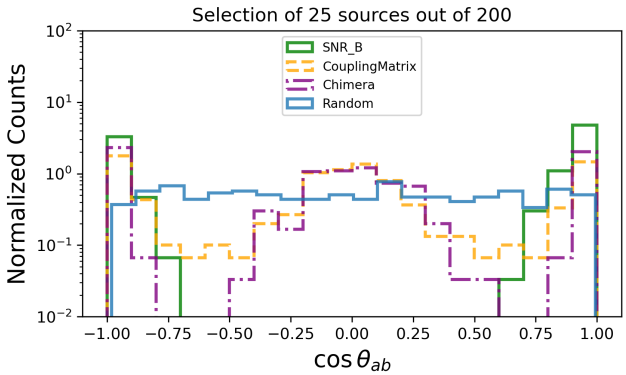

Figure 1. Distribution of angular separations of 25 pulsars selected with three selection methods, namely $\mathrm{SNR_B}$-maximization, Coupling Matrix and Chimera. These methods have been applied to a dataset consisting of 200 pulsars with uniform sky distribution and equal noise properties. For reference, we also show a random selection of 25 pulsars.

The region between -0.6 and 0.6 will be eventually filled as the number of selected pulsars increases[4]. The Coupling Matrix and Chimera methods also picked pulsars at $\theta_{ab} = 0°$ and $180°$, but the distribution of angular separations is broader and covers more values of θ_{ab}. We find that of the first 25 pulsars selected by the Chimera method, none of them are placed around $\cos\theta_{ab} \approx -0.7$ and $\cos\theta_{ab} \approx 0.7$. This might be due to some interaction between SNR-maximization and Coupling Matrix selection. Note that the pattern in Fig. 1 could change if we were starting with two or more pulsars with different sky locations.

2.4 Continuous gravitational wave SNR maximization

Continuous gravitational waves are deterministic signals and their analysis has been treated separately from the stochastic GWB. CGWs are included in the model as a periodic delay applied to the timing residuals $\vec{\delta t}$ while the effect of the GWB is included in the covariance matrix $\mathbf{C}$ of the likelihood. This fundamental difference between the two signals and their mathematical description calls for a different ranking method.

Here, we want to rank pulsars according to their response to a CGW signal. One way to proceed is to inject a large number of fake CGW signals with randomized parameters except for fixed frequency and amplitude (Babak et al. 2015). Then, for each pulsar, the CGW signal-to-noise ratio is computed for each injection and averaged numerically. In this way, we have the average response of each individual pulsar in the array at a given frequency of the CGW signal. This averaging can also be done analytically, as shown in the following paragraph. Note that we refer to the signal-to-noise ratio of CGWs using the acronym SNR. However, we use the symbol ρ to distinguish the SNR of CGWs from the previously defined SNRs.

In the likelihood of Eq. (6), the inclusion of a deterministic signal is performed by changing the timing residuals as $\vec{\delta t} \to \vec{\delta t} - \vec{s}(\vec{\theta})$ where $\vec{s}(\vec{\theta})$ is the signal template we aim to measure. In that case, the likelihood can be rewritten as:

$$\ln \mathcal{L} = -\frac{1}{2}\left[(\vec{\delta t}|\vec{\delta t}) + (\vec{s}|\vec{s}) - 2(\vec{\delta t}|\vec{s}) + \mathrm{Tr}\ln 2\pi\mathbf{C}\right], \tag{16}$$

where we have introduced the noise weighted inner product $(\vec{x}|\vec{y}) = x^T\mathbf{C}^{-1}y$.

We can now calculate this expression for the hypothesis of the presence of a CGW ($\mathcal{H}_1$) versus its absence ($\mathcal{H}_2$). The expectation value of the log-likelihood ratio becomes:

$$\langle \ln\Lambda\rangle_{\mathcal{H}_1} = \left\langle \ln\left(\frac{p(\vec{\delta t}|\vec{s})}{p(\vec{\delta t}|\vec{0})}\right)\right\rangle_{\mathcal{H}_1} = \langle(\vec{\delta t}|\vec{s}) - \frac{1}{2}(\vec{s}|\vec{s})\rangle_{\mathcal{H}_1} = \frac{1}{2}(\vec{s}|\vec{s}), \tag{17}$$

where $\rho_{\mathrm{Opt}} = \sqrt{(\vec{s}|\vec{s})}$ is the optimal SNR for the CGW source.

Since the source parameters are not known a priori, we average ρ^2_{Opt} over gravitational wave polarization ψ, initial phase ϕ_0, inclination ι, and sky location (θ, ϕ). To do so, we analytically compute the integral over the defined bounds of the CGW parameters:

$$\rho^2 = \int_0^{\pi}\frac{d\psi}{\pi}\int_0^{2\pi}\frac{d\phi_0}{2\pi}\int_1^{-1}\frac{d\cos\iota}{2}\int_1^{-1}\frac{d\cos\theta}{2}\int_0^{2\pi}\frac{d\phi}{2\pi}(\vec{s}|\vec{s}). \tag{18}$$

Using the formula for a CGW signal from a circular SMBHB, $\vec{s}(t, \Omega)$, as presented in Babak & Sesana (2012), the Earth-term SNR^2 averaged over CGW parameters takes this simple form:

$$\rho^2(h, f) = \frac{4}{15}\left(\frac{h}{2\pi f}\right)^2\left[(\cos 2\pi f t|\cos 2\pi f t) + (\sin 2\pi f t|\sin 2\pi f t)\right], \tag{19}$$

with

$$h = \frac{2\mathcal{M}^{5/3}(\pi f)^{2/3}}{d_L}, \tag{20}$$

where f and h are the gravitational wave frequency and amplitude, $\mathcal{M}$ is the chirp mass and d_L is the luminosity distance. For pulsar a, we evaluate ρ^2_a at the TOAs $\vec{t}_a$. We consider an Earth-term only SNR for simplicity as the inclusion of the pulsar term is unlikely to make a significant difference to the ranking. In the absence of a chirp, the contribution of the pulsar term to the SNR^2 is equal to that of the Earth term, therefore leaving the relative contribution of different pulsars unchanged. When the system is chirping this is no longer true as different pulsar terms contribute at different frequencies. However, it is slightly misleading to include these in the ranking on an equal footing with the Earth terms, since matching the pulsar terms in the data is much harder and requires good knowledge of the pulsar distance. In addition, the resulting ranking would be dependent on the nature of the source in the data, as this determines the frequencies of each of the pulsar terms, which would not be known until after the analysis using the reduced set of pulsars had been completed. The correlated noises (e.g. intrinsic and dispersion measure noises) are taken into account in the covariance matrix $\mathbf{C}$ of the noise-weighted inner product of the cosine and sine terms.

Common (correlated) processes were not included in our noise model, so the covariance matrix is block-diagonal. In this way, the likelihood can be factorized and SNR^2s can be computed independently for each pulsar. Common uncorrelated processes can be included without affecting the block diagonal form of the matrix, and this could be used as a proxy for the presence of a GWB background or other processes. In practice, we should incorporate these common processes in the noise model, but this adds another level of complexity that is irrelevant for the goal of the selection procedure[5]. The

[4] We included in the supplementary materials two animated figures that show how the SNR-maximization method squentially adds pulsars, see `animate_hist_HDvsNoise_loc_3d.gif` and `animate_hist_HDvsNoise.gif`.

[5] Furthermore, detectable CGW signals must be louder than the GWB. Since

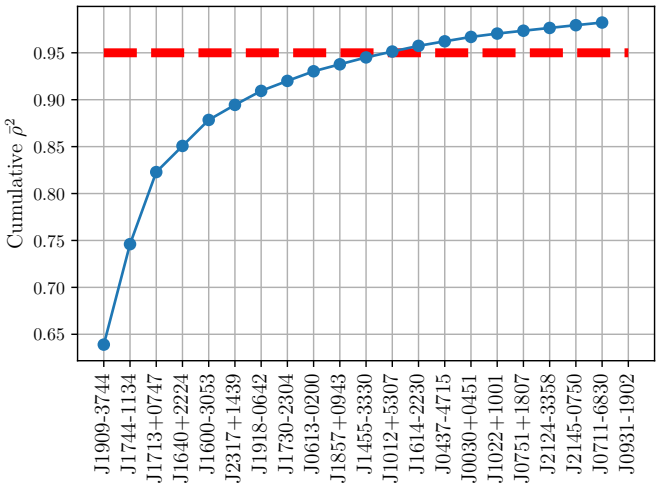

Figure 2. Cumulative $\bar{\rho}^2$ plot for the pulsars in the IPTA DR2 at CGW frequency of 5nHz. The pulsars above the red dashed line contribute less than 5% of the total SNR^2. This means only 12 pulsars out of 65 contribute on average to 95% of the total SNR^2 of the array at 5nHz. Note that, while only the best 22 pulsars are shown in the figure, the normalized total SNR has been evaluated using all 65 pulsars in the array.

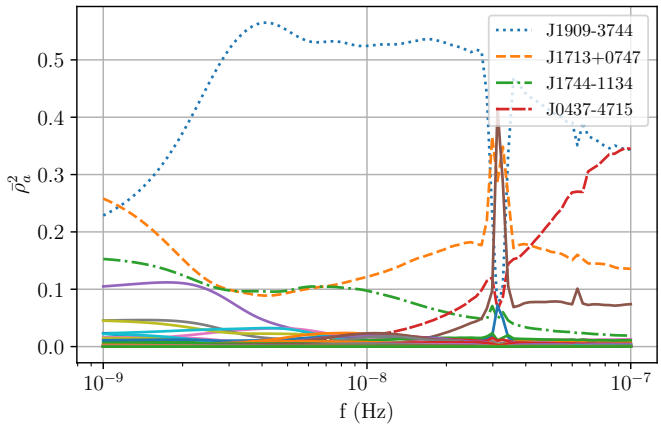

Figure 3. Normailzed $\bar{\rho}_a^2$ of the five best pulsars of the IPTA DR2, at different CGW frequencies. The glitches at the right of the plots are due to the one year and half-year peaks.

ultimate goal is identification of the best pulsars for CGW detection, and therefore, only the intrinsic properties of the pulsars were considered.

We estimate the relative contribution of one pulsar to the total SNR of the array using the normalized SNR^2:

$$\bar{\rho}_a^2(f) = \frac{\rho_a^2(h,f)}{\sum_b \rho_b^2(h,f)}, \tag{21}$$

Note that the amplitude h cancels out in this expression and the CGW frequency f remains the only parameter. Therefore we can fix h to any value without affecting the ranking.

We construct the cumulative sum of the normalized SNR^{2}s of the pulsars ranked from best to worst. We fix a threshold value for the SNR^2 cumulative sum above which pulsar contributions to the total SNR^2 are not considered significant. This value was chosen to be 0.95. The process is illustrated in Figure 2 and in the animated Figure (`cgw_ranking.gif` included in the supplementary materials) for pulsars from the IPTA second data release (DR2, Perera et al. 2019b).

Due to the strong dependence of $\bar{\rho}_a^2(f)$ on f, the resultant CGW pulsar ranking is also frequency dependent. This can be clearly seen from Figure 3. In our analysis, we use 100 log-spaced frequency bins between 10^{-9} and 10^{-7} Hz. Ranking lists were obtained separately for each frequency bin. In order to construct the final ranking catalog of best pulsars at a given frequency range, the lists at each frequency are merged together. This procedure ensures that we will gain at least, no matter the CGW frequency, 95% of the total SNR^2 of the array.

3 RESULTS

We create mock PTA datasets with increasing complexity in the noise models and test the performance of the selection methods. The PTA datasets are simulated using LIBSTEMPO[6] and analysed using ENTERPRISE (Ellis et al. 2020) giving the marginalized likelihood. Bayes factors are computed using DYNESTY (Speagle 2020).

3.1 Testing the selection methods for GWB searches

In this section we investigate the performance of the three ranking methods that target GWB searches (Sec. tion 2.3). We consider a simplified framework, in which the pulsar noise is white noise only, and there is an injected GWB with amplitude $A_{\rm GWB} = 3 \times 10^{-15}$ and slope $\gamma = 13/3$, consistent with findings from the EPTA analysis (Chen et al. 2021). We pick pulsars one by one using the SNR$_{\rm B}$-maximization, the Coupling Matrix method (with weights $w \sim$ SNR$_{\rm A}$), and the Chimera method, and we investigate the performance of these procedures by calculating the log-Bayes factor (lnBF natural logarithm) of the following hypothesis tests:

- HD vs CURN: Hellings & Downs correlation versus a common uncorrelated red noise process;
- HD vs CURN + MN + DN: Hellings & Downs correlation versus a combination of common uncorrelate red process, monopolar noise (MN) and dipolar noise (DN).

Since a detectable GWB signal is injected, we expect the log-Bayes factor to always increase in the limit of a high number of pulsars N. Of particular importance, however, are the dynamics of growth of the log-Bayes factor with respect to a random selection. A further comparison of these selection methods against a lowest-RMS selection procedure is presented in Appendix B

Note that the white noise parameters are kept fixed, and only the amplitudes and slopes of the common red noise processes are varied. In the next sections we present the evolution of the log-Bayes factor obtained with the N pulsars selected with the aforementioned methods. We anticipate that the performance of the selection methods strongly depends on the specifics of the dataset considered. Therefore, we tested our ranking methods with three different simulated datasets.

the GWB is stronger at lower frequencies, CGW signals are more likely to be found at high frequencies.

[6] https://github.com/vallis/libstempo

3.1.1 Galaxy-distributed dataset

We created an array of 200 pulsars with equal RMS of 100 ns with galaxy distribution on the sky. The sky coordinates were drawn randomly from the available values of known pulsars in the `psrcat` catalogue (Hobbs et al. 2004). The total timespan of the dataset is 10 years with a sampling rate of 28 days. A dataset consisting of all equal pulsars with a dense sky coverage serves to demonstrate how each selection method performs under idealised conditions. In Figure 4 we show the log-Bayes factor computed using the pulsars selected by the different ranking methods when applied to the Galaxy-distributed dataset for the hypothesis tests: HD vs CURN, and HD vs CURN + MN + DN. The very first pulsar in the array was selected at random 20 times, so that the log-Bayes factor shown in Figure 4 is an average over these realizations. This procedure was done in order to ensure that our results are independent of the initial pulsar choice. For reference, we also show the log-Bayes factor obtained with a random selection of pulsars.

The left panel of Figure 4 demonstrates that the Coupling Matrix method (dashed yellow line) performs similarly to the random selection (dotted blue line) for the HD vs CURN hypothesis test, with slightly better performance after $\sim$ 15 pulsars are included in the array. Both the SNR-maximization (solid green line) and Chimera method (purple dash-dotted line) outperform the other two types of selection. For the SNR maximisation method the log-Bayes factor increases with the number of pulsars in the array like $\sim 0.8N$, which results in almost double log-Bayes factor for $N = 25$ than the one obtained using random selection. These results are expected, since the SNR-maximization is designed to maximize the confidence of detecting the HD correlation versus a CURN process.

The hypothesis test HD vs CURN + MN + DN is proposed to demonstrate the benefits of the Coupling Matrix, as the method is designed to disentangle the HD correlation from other types of common correlated noises. The right panel of Figure 4 confirms these expectations. We see that, in this context, the Coupling Matrix and Chimera methods provide a log-Bayes factor for $N = 25$ pulsars which is 1.4 and 1.6 times larger than a random selection, respectively. The scaling of the log-Bayes factor for the Chimera selection is $\sim 0.8N$, while the SNR selection scales only as $\sim 0.2N$. The SNR-maximization is severely suboptimal for this test, as it tends to pick pulsars at locations where the HD overlap reduction function is the largest, i.e., at 180° and 0°, making it harder to discern HD from other types of correlation. A random selection of pulsars provides a more distributed sky coverage which improves the situation in this regard.

The slightly improved performance of the Chimera method in comparison to the Coupling Matrix formalism is due to the fact that it accounts for both the optimal sky coverage and total gain in SNR. These results confirms that both of these components are essential for PTA optimization and cannot be ignored. One can conclude that the inclusion of the SNR-maximiztion in the Chimera method is of special relevance in the case of non-equal pulsar arrays. The latter point is even more evident in one of the following subsection, where we consider a simplified EPTA dataset.

3.1.2 Mock MeerTime dataset

We now consider a PTA dataset which resembles the properties of the recently published 5-year MeerTime Large Survey (Spiewak et al. 2022). This survey is expected to significantly increase the sensitivity of current PTAs in the very near future. Using this as motivation, we created a mock MeerTime dataset consisting of 189 pulsars with sky positions taken from the survey. Observations were performed every 28 days on a baseline of 10 years. The white noise RMS is set to the median TOA uncertainties delivered by MeerTime, in which each observation epoch of each source consisted of 256 seconds of integration time with the MeerKat radio telescope. The dataset provides an insight on how the pulsar selection performs with a large dataset composed of non-equal pulsars with realistic sky positions.

We generate 20 noise realizations of this dataset and show the averaged log-Bayes factor in Figure 5. The first pulsar in the ranking is fixed to the one with the smallest RMS.

The left panel of Figure 5 shows the ranking for the HD vs CURN test, and it confirms that the Chimera method and the SNR-maximization are optimal in this case. Even though the pulsars selected with the Coupling Matrix method provide a log-Bayes factor smaller than the other methods, it still gives an evidence which is approximately 3 times larger in comparison to random selection for $N = 25$.

The evolution of the log-Bayes factor for the hypothesis test HD vs CURN + MN + DN is shown in the right panel of Figure 5. The Coupling Matrix and Chimera selections increase the log-Bayes factor up to $\log_{10}\mathrm{BF} \approx 12$. Differently from the "galaxy-distributed" dataset, the SNR-maximization performs slightly better than the random selection, although still worse than the Coupling Matrix and Chimera methods. Up to the first 18 pulsars, the Chimera method provides a stronger support for HD vs CURN + MN + DN than the Coupling Matrix, reaching similar levels for larger number of pulsars.

3.1.3 EPTA-simplified dataset

We construct an EPTA-simplified dataset, which consists of 40 pulsars with RMS and sky location of the latest EPTA dataset (Desvignes et al. 2016; Chen et al. 2021). The total timespan is fixed to 10 years with observations being performed every 28 days. In order to reduce required computational resources, only white noise was taken into account, ignoring the red intrinsic and interstellar medium noise contributions. Despite the significant simplification, this dataset serves to imitate a realistic PTA setup with a modest number of pulsars and representative pulsar sensitivities, which has been principally used for GW searches to date. We have simulated 20 statistically equivalent noise realizations. The averaged log-Bayes factor are shown in Figure 6. As in the case of the mock MeerTime dataset, the first initial pulsar is chosen to be the one with the smallest RMS.

It can be seen from both panels of Figure 6, that the restricted dataset of 25 pulsars chosen by the Chimera or SNR-maximization methods on average reaches higher log-Bayes factors than those selected randomly or using the Coupling Matrix formalism. Moreover, Figure 6 shows that by using only 25 of pulsars picked by one of the two former methods, we account for $\approx 90\%$ of the sensitivity of the whole array. The Coupling Matrix approach, on the other hand, falls behind, even for the HD vs CURN+MN+DN hypothesis test. These results clearly demonstrate that pulsar quality is as important as optimal sky location, when disentangling different types of correlations. The Coupling Matrix is not aimed at maximizing the SNR, therefore it can not be used as a selection method on its own, as some of the highly sensitive sources could be discarded. The best results are obtained when the optimal sky location and gain in SNR are finely balanced. Therefore, "good" pulsars must be picked at proper sky locations, which is the main idea behind the Chimera method. In other words, neither low-sensitivity sources selected at proper angular distances, nor high-SNR sources with poorly chosen coordinates, e.g. clustered at a specific location on the sky, can provide an adequate improvement in performance. The former case is the Coupling

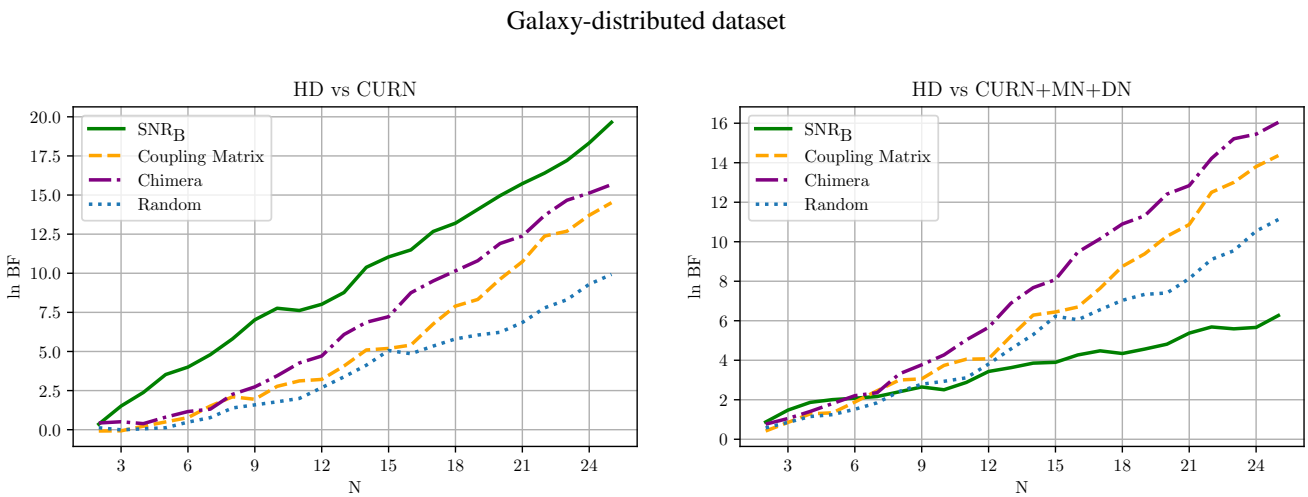

Figure 4. Log-Bayes factor as a function of the number of chosen pulsars by each of the selection methods (shown in different colors) for the Galaxy-distributed dataset and for different hypothesis tests: HD vs CURN (left), and HD vs CURN+MN+DN (right). The 200 simulated pulsars have the same noise properties and galaxy-distributed sky locations. The first pulsar is selected at random 20 times and the shown log-Bayes factors are the average over these 20 realizations. For 25 selected pulsars the mean and standard deviation values are: SNR$_B$: 20±6, Coupling Matrix: 15±7, Chimera: 16±5, Random: 10±4 (HD vs CURN hypothesis test (left)); SNR$_B$: 6±2, Coupling Matrix: 14±7, Chimera: 16±5, Random: 11±4 (HD vs CURN+MN+DN hypothesis test (right)). The log-Bayes factors of the whole array for one realization are 198 and 194 for HD vs CURN (left), and HD vs CURN+MN+DN (right), respectively.

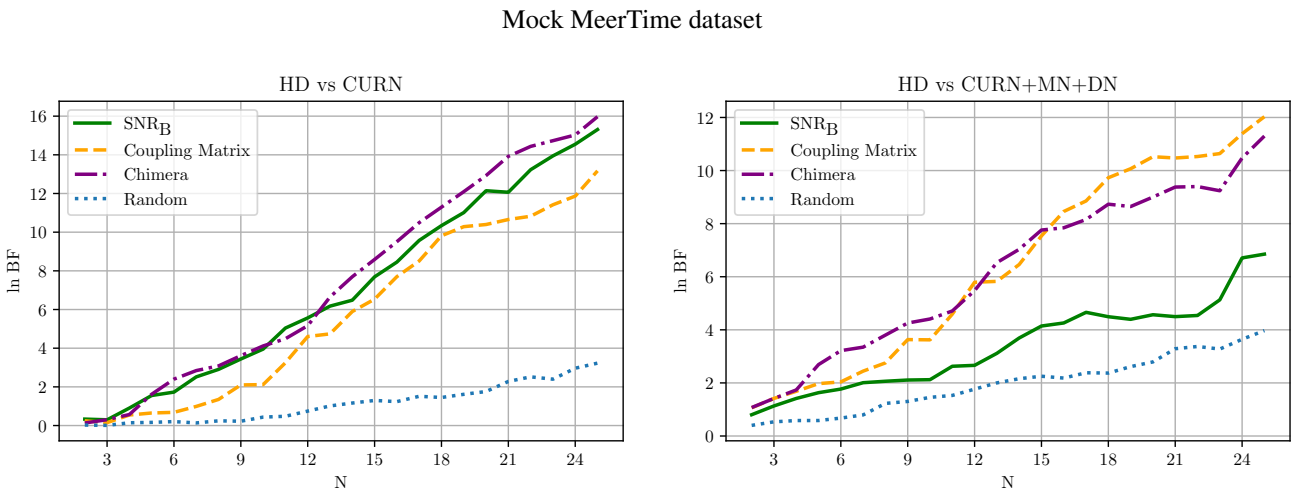

Figure 5. Log-Bayes factor as a function of the number of chosen pulsars by each of the selection methods (shown in different colors) for the mock MeerTime dataset and for different hypothesis tests: HD vs CURN (left), and HD vs CURN+MN+DN (right). The shown log-Bayes factors represent the average over 20 different noise realizations. For 25 selected pulsars the mean and standard deviation values are: SNR$_B$: 15±10, Coupling Matrix: 13±8, Chimera: 16±10, Random: 3±2 (HD vs CURN hypothesis test (left)); SNR$_B$: 7±3, Coupling Matrix: 12±6, Chimera: 11±6, Random: 4±2 (HD vs CURN+MN+DN hypothesis test (right)). The log-Bayes factors of the whole array are 57±21 and 47±16 for HD vs CURN (left), and HD vs CURN+MN+DN (right), respectively.

Matrix selection for the EPTA-simplified dataset (yellow dashed line in the left panel of Figure 6), while the latter corresponds to SNR-maximization for the MeerTime dataset (solid green line in the right panel of Figure 5).

We want to remark that the Chimera implementation we offer in this paper is not the ultimate solution. Alternative ways to address this issue are proposed in Appendix A. Furthermore, as demonstrated in Appendix B, simpler ranking criteria might perform better than the Chimera method for some datasets. More thorough investigations are left for future works.

3.2 Optimizing the search for a GWB in a realistic EPTA dataset

To speed-up the assembly of the new dataset and to improve computational efficiency of the analysis, the EPTA collaboration decided to select a subsample of pulsars timed by its radio facilities. In this context, it is of paramount importance to wisely pick the pulsars to be included. Therefore, we create another simulated array to address this problem. We consider a dataset similar to the one of Sec. 3.1.3, i.e. 40 pulsars with RMS, timespan, and sky locations of the EPTA dataset, but more realistic in the sense that we include the intrinsic

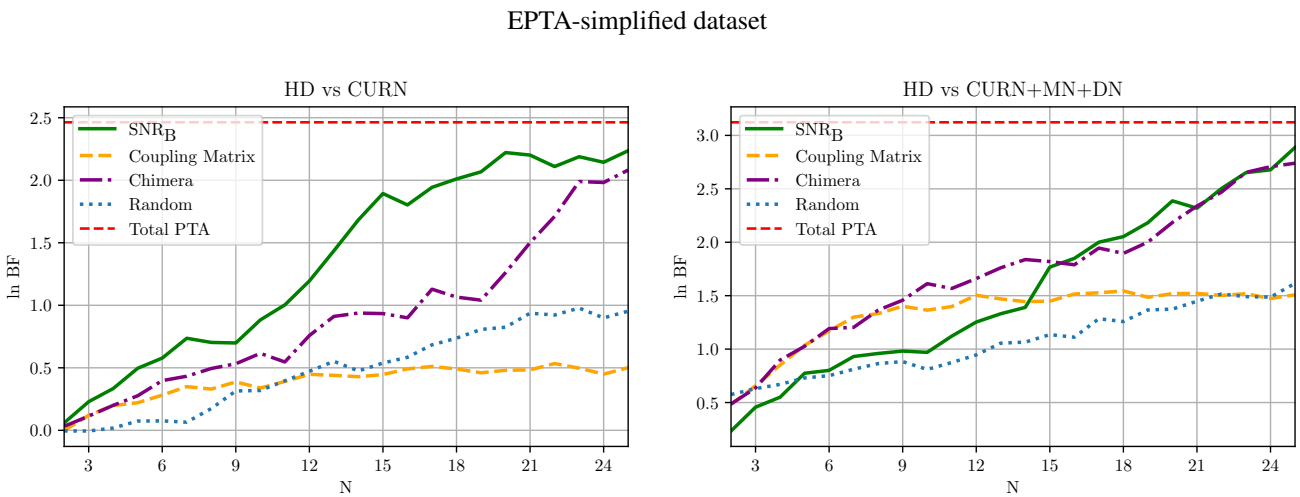

Figure 6. Log-Bayes factor as a function of the number of chosen pulsars by each of the selection methods (shown in different colors) for the EPTA-simplified dataset and for different hypothesis tests: HD vs CURN (left), and HD vs CURN+MN+DN (right). The shown log-Bayes factors represent the average over 20 different noise realizations. For 25 selected pulsars the mean and standard deviation values are: SNR$_B$: 2.2±1.9, Coupling Matrix: 0.5±1.1, Chimera: 2.1±1.9, Random: 1.0±1.2 (HD vs CURN hypothesis test (left)); SNR$_B$: 2.9±1.8, Coupling Matrix: 1.5±1.0, Chimera: 2.7±1.9, Random: 1.6±1.4 (HD vs CURN+MN+DN hypothesis test (right)). The red-dashed line shows the log-Bayes factor of the full dataset (N=40): 2.5±2.3 for HD vs CURN and 3.1±2.2 for HD vs CURN+MN+DN.

red-noise properties of the preliminary EPTA dataset[7] (Chen et al. 2021; Lentati et al. 2015).

For simplicity, we focus on ranking the best pulsars to distinguish an HD correlation (hypothesis $\mathcal{H}_1$) from a CURN process (hypothesis $\mathcal{H}_2$) and we study how this can be affected by possible noise realizations. As shown in the previous sections, SNR-maximization and the Chimera method should be a good selection proxy for this hypothesis test. Since the SNR-maximization method is constructed to target this hypothesis and it has been shown to perform as well as the Chimera method, we will only use this method for this study. The first six pulsars are fixed to those which constitute the preliminary combination of Chen et al. (2021): J1909-3744, J1713+0747, J1744-1134, J0613-0200, J1600-3053, J1012+5307.

Firstly, we estimate the number of sources that to be added to the preliminary combination in order to achieve a reasonable detection confidence. For this, we apply the SNR maximization selection using the injected GWB parameters, and iteratively add the pulsars which increase the SNR the most. Results are shown in Figure 7. SNR$_A$ tends to saturate more quickly than SNR$_B$. This is because the latter is suppressed by the term $S(f)$ in the denominator of Eq. (10). We find that with $N = 25$ pulsars we reach 94% of the total SNR$_B$. Therefore, adding 19 SNR-maximization selected pulsars to the starting six sources increases the SNR from 30% to 94% of the total SNR of the array.

Next, we want investigate whether the selection procedure is strongly affected by the choice of GWB parameters. To this end, we simulate the EPTA mock dataset 1000 times with the same injection parameters, and find the Maximum Likelihood Estimator using only the first six pulsars (preliminary dataset) and assuming an HD correlation only. The intrinsic red and white noise parameters were fixed to the true values. The results are shown in Figure 8. Different noise realizations lead the MLE values (blue dots) to be shifted from

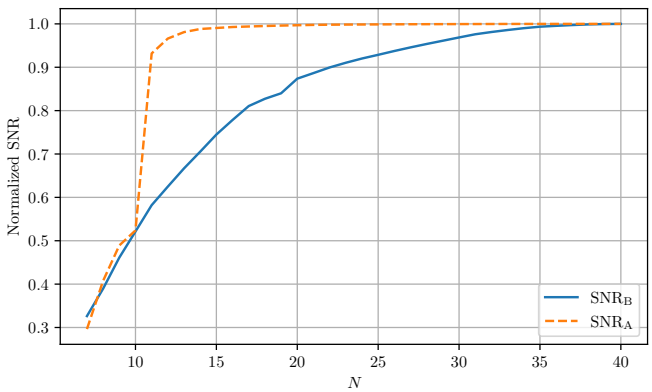

Figure 7. Normalized SNR evolution as a function of the number of selected pulsars N with the SNR maximization method of statistic B and A. The SNR is normalized to the total SNR of the dataset and the initial pulsar subset is composed of the 6 initial pulsars of the EPTA analysis (Chen et al. 2021).

the true parameters (orange dot). It can be clearly seen that the distribution of MLEs lies along the line over which the six initial pulsars are located (red triangles), and its median (dashed black lines) is consistent with the injected true parameters. For reference, we show the adopted intrinsic red noise parameters of the other pulsars in the simulated datasets as red crosses.

We now use each of the MLEs of Figure 8 as a new set of GWB parameters and run the SNR ranking procedure. The histogram of the best 25 selected pulsars is shown in Figure 9. Since the GWB parameters are different at every realization, the subset of selected pulsars slightly changes. As expected, the histogram for the SNR$_B$ selection has larger tails since different GWB parameters affect both the denominator and numerator of the Eq. 10. Instead, the SNR$_A$ is affected only by the variation in the GWB slope γ. Both SNR$_A$ and SNR$_B$ selections exclude 15 pulsars in each realization. This selection reduces the total number of TOAs to analyze from 18584 to 12191 (in median). Therefore, the SNR ranking procedure excludes

[7] For simplicity we adopt the best fit estimates as representative values from the EPTA constraints on the red noise parameters and set the time interval between observations to be 14 days.

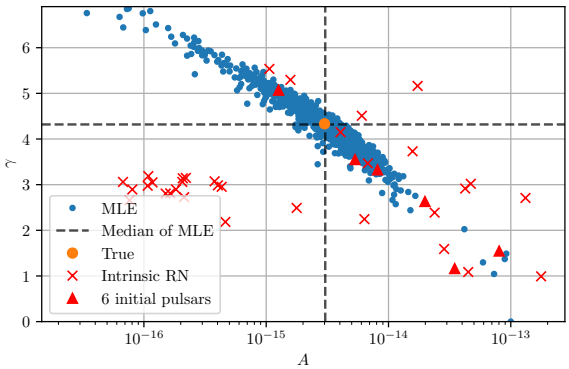

Figure 8. Maximum likelihood estimation of the amplitude A and slope γ of the stochastic gravitational-wave background using the first 6 pulsars of the EPTA mock dataset (the red triangles show the respective intrinsic red noise properties). The blue dots show the estimated values of A and γ per noise realization, and the dashed lines indicate the median distribution value. The orange dot shows the true injected value, whereas the red crosses show the values of the intrinsic red noises injected in the remaining pulsars.

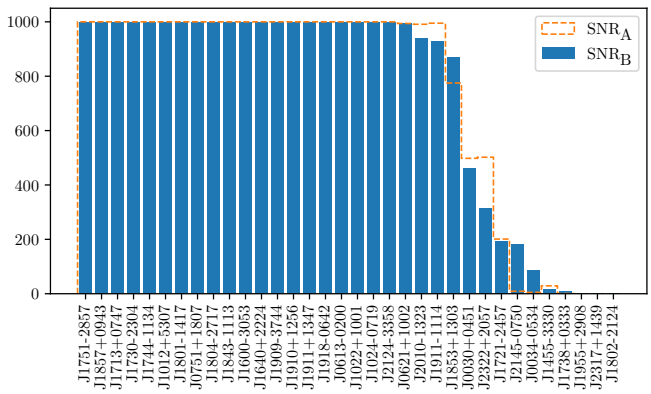

Figure 9. Histogram of the 25 pulsars selected with the $\mathrm{SNR_B}$ (blue) and $\mathrm{SNR_A}$ (orange) maximization over 1000 noise realizations.

6393/18584 ≈ 35% of the TOAs of the full dataset by excluding 15 out of 40 pulsars. As shown in Figure 9, both methods pick the same 20 pulsars in majority of the cases. In practice, we could find the best pulsars by performing the selection process with the GWB and intrinsic red noise parameters taken from posterior chains of the previous data release. However, such an analysis is beyond the scope of this work.

We now demonstrate that the SNR-maximization selection method performs better than a random selection, and it provides evidence comparable to the full dataset. For each of the 1000 noise realizations, we select 25 pulsars in three ways: using the SNR-maximization methods ($\mathrm{SNR_B}$ and $\mathrm{SNR_A}$) as done in Figure 9, and randomly. We compute the log-likelihood ratios obtained with the three different pulsar subsets and with the full dataset and we show the results in Figure 10. These distributions are evaluated at maximum-likelihood estimates of the parameters (amplitudes and slopes of the GWB). Based on the median values of the distributions, one finds that the optimally selected datasets provide a factor of 1.84 – 1.90 stronger evidence with respect to the random selection. Furthermore, we find that the log-likelihood ratio for the 25 optimally selected dataset is

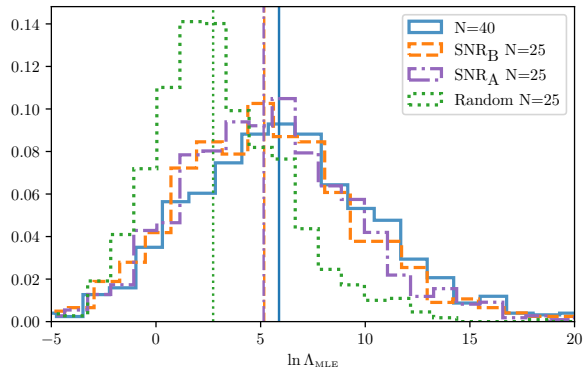

Figure 10. Distribution of log-likelihood ratios obtained with the full dataset $N = 40$ (solid blue) and with 25 pulsars selected with $\mathrm{SNR_B}$ (dashed orange) and $\mathrm{SNR_A}$-maximization (dash-dotted purple) for 1000 noise realizations. For each noise realization we also randomly select 25 pulsars and calculate the log-likelihood ratio of this distribution. The distribution of these log-likelihoods is also shown as a green dotted histogram. The medians of the distributions are shown as vertical lines and are 5.88 for $N = 40$, 5.17 for $\mathrm{SNR_B}$ $N = 25$, 5.14 for $\mathrm{SNR_A}$ $N = 25$, and 2.73 for Random $N = 25$. The log-likelihood ratios have been all evaluated at the maximum likelihood value.

in median ∼ 0.89 times the one obtained from the full array. The distributions of log-likelihood ratios evaluated at the true parameters do not significantly differ from those shown in Figure 10. Therefore, the search over the GWB parameters with the MLE is not affecting the distribution of log-likeliood ratios.

These results demonstrate that the SNR-maximization selection method is a good proxy for choosing pulsars and it is robust against noise realizations. Furthermore, we have demonstrated that the log-likelihood ratio obtained with a subset of 25 pulsars is comparable to the one from the full array.

Now, we establish the significance achieved by the optimally selected pulsars. To this purpose, we simulate two sets of realistic EPTA datasets: with an injected CURN process; and with an injected HD correlated process. The two injected common processes are characterized by the same amplitudes and slopes. We show in Figure 11 the log-likelihood ratios obtained using the full dataset (N=40) and the 25 $\mathrm{SNR_B}$ selected pulsars for the HD and CURN injection subsets. The median of the log-likelihood ratios of the best 25 pulsars for the HD injection (orange dashed-line histogram) corresponds to a p-value of $\approx 2\times10^{-3}$ with respect to the CURN log-likelihood ratio distribution (black dashed-line histogram). The log-likelihood ratio distributions for the full array (N = 40) are shown in Fig. 11 as solid-line histograms for the CURN (gray) and HD injection (blue), respectively. Since the median of the latter distribution (HD) is above all the log-likelihood ratios obtained with the CURN injection with N=40 pulsars, we estimate the respective p-value as smaller than one over the number of noise realizations/samples, i.e. $\lesssim 10^{-3}$. We caution the reader that the aforementioned p-values are only approximate. In fact, to resolve the tails of the CURN log-likelihood distribution, we would need to run our analysis for a larger number of noise realizations. Nevertheless, these results demonstrate that the selection of pulsars does not significantly affect the statistical significance of the hypothesis test.

We showed that the SNR-maximization selection method is a good proxy for ranking pulsars and it allows to reach detection confidence comparable to the full array. However, it is important to remark that these results are obviously dependent on the specific pulsars' sky

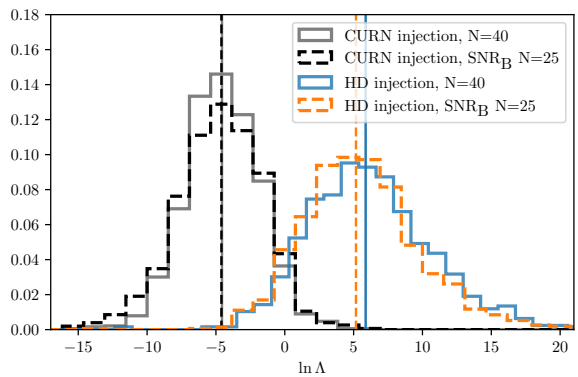

Figure 11. Distribution of log-likelihood ratios $\ln\Lambda$ for the hypothesis test of the HD correlation versus Common uncorrelated Red Noise process over many noise realizations and different injections. The dashed lines show the distribution when the log-likelihood is computed using the 25 pulsars selected with the SNR$_B$ maximization, whereas the solid lines when all 40 pulsars are used. The median values of the distributions for the CURN injection are -4.56 and -4.60 for $N = 40$ and SNR$_B$ $N = 25$, respectively, whereas for the HD injection these are 5.87 and 5.17 for $N = 40$ and SNR$_B$ $N = 25$, respectively. The log-likelihood ratios have all been evaluated at the true injected parameters.

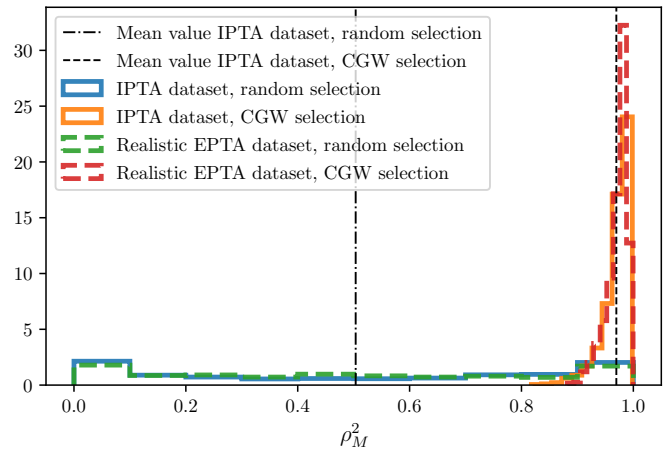

Figure 12. Distribution of the normalized SNR^2 coverage for 1000 different sets of CGW parameters. The distributions are obtained with the list of pulsars chosen according to the CGW selection method, in this case 22 for both the real IPTA dataset and the realistic EPTA dataset. For comparison, we also show the distribution of the normalized SNR^2 obtained with a random selection.

localizations and noise properties and on the tested hypothesis (here HD vs CURN). We expect this ranking method to be well suited also for other PTA datasets where the pulsars have very different noise properties.

We remark that similar results can be obtained also with a lowest RMS selection. However, such a method becomes sub-optimal once the observation cadence is not the same across all pulsars. For a more detailed investigation see Appendix B2

3.3 Optimizing IPTA and EPTA analysis of CGW signals

We now test the performance of the CGW ranking method using noise-parameter values previously extracted from individual pulsar noise analyses of the latest IPTA data release (Perera et al. 2019b) and the realistic EPTA dataset created in the previous Sec. 3.2.

Because the ranking method is based on an exact noise-averaged formula, it is unnecessary to simulate noise realizations to test its performance. However, we still want to prove that the selected pulsars recover most of the total SNR in the presence of a true (i.e. non-averaged) signal. We test this by comparing the fraction of total SNR^2 obtained using the CGW ranked pulsars to that obtained from a random pulsar selection. For an array of N pulsars, the fraction of total SNR^2, given a list of $M < N$ pulsars, is defined as:

$$\rho^2_M = \sum_{a=1}^{M} \bar{\rho}^2_a \,, \qquad \text{with } 0 < \rho^2_M < 1 \,, \tag{22}$$

where $\bar{\rho}^2_a$ is the normalized SNR^2 defined in Eq (21).

After extracting the list of best pulsars, we test the selection procedure as follows:

- We draw the CGW signal parameters $\vec{\theta}$ from a uniform distribution with bounds defined as in the integral of Eq. (18), and with frequency between 1 and 100 nHz. As pointed out in Sec. (2.4) the strain amplitude has no influence on the ranking and therefore we fix it to $h = 10^{-14}$.
- We compute the non-averaged optimal SNR $\rho_{\rm Opt} = \sqrt{(\vec{s}|\vec{s})}$ for each pulsar for a CGW signal $\vec{s}(t, \vec{\theta})$ and we use this quantity to calculate the normalized $\bar{\rho}^2_a$ defined in Eq. (21).
- We compute $\rho^2_{M-\rm CGW}$ for the list of best selected pulsars and $\rho^2_{M-\rm rand}$ for a random subset of pulsars of random size M.
- We repeat the previous steps one thousand times.

This gives us 1000 values of $\rho^2_{M-\rm CGW}$ and $\rho^2_{M-\rm rand}$ that we plot as histograms on Figure 12. For the IPTA dataset, the distribution of fractional $\rho^2_{M-\rm CGW}$ for the selected pulsars is narrowly peaked around a mean value 0.97. The random selection $\rho^2_{M-\rm rand}$ gives an almost uniform distribution with 0.50 mean value. The distribution is not uniform because ρ^2_a is not uniform and a few ρ^2_a values are much bigger while many others are very small. Similar results are obtained for the realistic EPTA dataset. We find that the number of pulsars which gives 95% of the SNR^2 is 22 for both datasets, and these pulsars represents respectively 61% of the total number of TOAs (= 18584) for the realistic EPTA dataset, and 76% of the total number of TOAs (= 210148) for the IPTA dataset.

Now we briefly discuss the comparison between the CGW and GWB selection methods. Focusing on the realistic EPTA dataset, we find an overlap between the identified best pulsars with the CGW method and GWB method as shown in Table 1. This time we run the Chimera and SNR$_B$-maximization ranking without fixing the six initial pulsars of the EPTA. We find that 17 pulsars are common to all three selection methods (highlighted in bold).

In summary, when true CGW signals are injected in the data, the CGW ranking method selects the pulsars which provides most of the SNR of the array, whereas a random selection is inefficient. This method extracts the few best pulsars to optimize the search for a CGW signal.

4 CONCLUSIONS AND FUTURE OUTLOOK

PTA data analysis requires both significant human and computational resources. As the computational burden of such analyses grows with the number of pulsars, the problem will be further exacerbated by the discovery of many new pulsars by next-generation radio facilities. In

CGW ranking	Chimera method	SNR_B maximization
J0030+0451	**J0030+0451**	**J0030+0451**
J0613−0200	J0034−0534	**J0613−0200**
J0751+1807	**J0613−0200**	J0621+1002
J1012+5307	J0621+1002	**J0751+1807**
J1022+1001	**J0751+1807**	J1022+1001
J1024−0719	J1012+5307	**J1024−0719**
J1600−3053	**J1024−0719**	**J1600−3053**
J1640+2224	J1455−3330	**J1640+2224**
J1713+0747	**J1600−3053**	**J1713+0747**
J1730−2304	**J1640+2224**	**J1730−2304**
J1744−1134	**J1713+0747**	**J1744−1134**
J1751−2857	**J1730−2304**	**J1751−2857**
J1804−2717	**J1744−1134**	J1801−1417
J1853+1303	**J1751−2857**	**J1804−2717**
J1857+0943	J1801−1417	J1843−1113
J1909−3744	**J1804−2717**	J1853+1303
J1910+1256	J1843−1113	**J1857+0943**
J1911+1347	**J1857+0943**	**J1909−3744**
J1918−0642	**J1909−3744**	**J1910+1256**
J2010−1323	**J1910+1256**	J1911+1347
J2124−3358	J1911−1114	J1911−1114
J2145−0750	**J1918−0642**	**J1918−0642**
	J2010−1323	**J2010−1323**
	J2124−3358	**J2124−3358**
	J2322+2057	J2322+2057

Table 1. List of the first 22 pulsars selected with the CGW ranking method and the 25 pulsars selected with the Chimera method and SNR_B-maximization in the realistic EPTA dataset. Bold font indicate the 17 pulsars that are selected by all three methods.

this work, we introduced the concept of pulsar selection optimization for specific analyses. We emphasize that the ranking procedure is not straightforward and depends on the properties of the sought signal, and the optimization requirements. Therefore, we considered optimal selection criteria for deterministic CGW and stochastic GWB searches separately.

For the GWB, we presented three different ranking methods that target different aspects of a GWB search: SNR-maximization, Coupling Matrix, and Chimera method. The performance of our methods was assessed using frequentist and Bayesian hypothesis testing on simulated datasets.

The SNR-maximization method aims to increase the detection confidence in favor of the HD correlation with respect to a CURN process. Pulsars selected with this method provide an evidence for the HD vs CURN hypothesis larger than a random selection for all the considered datasets. For instance, using the EPTA-simplified dataset we obtained a log-Bayes factor which is double the one obtained with the random selection. Additionally, it was demonstrated that with this dataset we can reach 88% of the total sensitivity after including $N = 25$ pulsars out of 40. The SNR-maximization method was further studied in Sec. 3.1.3 for the case of a realistic EPTA dataset with intrinsic red noise included. We found that the first ~ 20 pulsars are included regardless of the particular noise realization and respective GWB parameter estimations. It was shown that the method selects pulsars which provide $1.8 - 1.9$ times larger log-likelihood ratio than a random selection. Furthermore, 25 pulsars out of the 40 selected by the SNR-maximization method accounted for 89% of the log-likelihood ratio of the full dataset.

Inherently, the SNR-maximization method tends to pick pulsars that maximize the HD ORF, which results in clustering of the sources at angular separations of $0°$ and $180°$. This fact can be detrimental for disentangling the HD from other spatially correlated noise processes. The Coupling Matrix selection is aimed at resolving this issue by maximizing the decoupling between different correlations, so that the HD spatial mode disentangles from the monopolar and dipolar correlations. This method has been shown to be efficient at increasing the evidence in the hypothesis test HD vs CURN+MN+DN in two out of the three datasets. The main pitfall of this method is that it weakly depends on the relative sensitivity of selected sources. As a consequence, some of the high-SNR sources are left behind, which is the main reason for the loss of sensitivity to GWB.

The Chimera method combines the two approaches to optimize both the sky coverage and the gain in total SNR. Even though its formulation is heuristic, this selection method has been a good proxy for selecting the pulsars that increase confidence in a GWB detection comparable to Coupling Matrix and SNR maximization. Specifically, for the simplified-EPTA dataset the method is able to recover 90% of the sensitivity of the whole array with $N = 25$ pulsars. In future work this formalism is going to be further examined. In particular, it would be interesting to explore if the Information matrix formalism introduced recently in Ali-Haïmoud et al. (2021, 2020) could be used to develop a more rigorous Chimera method, or a selection method targeting anisotropic searches.

The CGW SNR maximization is constructed to find the best pulsars to detect a CGW from a SMBHB. In contrast to the GWB case, CGW ranking deals with purely deterministic signals and this allows us to treat every pulsar independently, within our formalism. The method is based on an averaged SNR formula, and was applied to continuous wave signal searches in the IPTA and realistic EPTA mock datasets. Because of the strong dependence of an individual pulsar's SNR response $\bar{\rho}_a(f)$ on the CGW frequency f, ranking was performed separately for different frequency bins. In order to find the best pulsars on some frequency range, we had to take the union of the best pulsars that were identified for several frequency bins. Using the 22 best-ranked pulsars we recovered more then 95% of the total SNR^2 for both the IPTA and realistic EPTA datasets. Furthermore, we found that 17 of these pulsars are also selected by the SNR-maximization and Chimera methods.

The main takeaway points of our study can be summarized as follows:

- Although the addition of new pulsars inevitably increases the sensitivity of a PTA towards CGW and GWB detection (see Siemens et al. 2013), there exists an optimal subset of pulsars which is responsible for a larger portion of the sensitivity of a PTA, especially if the pulsar have different noise properties. This behaviour is confirmed in Figure 2 for CGWs, and Figures 6 and 10 for a GWB. If pulsars have all equal noise properties, it is possible to include pulsars such that the increase in the evidence is steeper than a random selection. This can be seen in Figure 4.
- In contrast to intuitive expectations, covering the sky uniformly with pulsars is not the most optimal strategy of pulsar selection for the purpose of disentangling different spatial modes, even in the case that all pulsars are equally sensitive. Instead, as can be seen from Figure 1, the ultimate distribution of pulsars in $\cos\theta_{ab}$ has three distinctive peaks at angular separations of $0°, 90°$ and $180°$. We expect that this distribution will converge to a uniform distribution, if we aim to resolve all multipoles.
- We stress that although a high SNR provides a steeper increase in the log-Bayes factor when HD is compared to all other considered types of common processes, it does not guarantee an optimal decoupling of spatial modes. This is clearly illustrated with the Galaxy-distributed and mock MeerTime datasets.

- Good sky coverage alone does not guarantee the effective decoupling of spatial modes. The optimal pulsar selection criterion should balance between proper sky localization and high sensitivity. The Chimera method is an attempt to create such a criterion which accounts for both properties. However, as demonstrated in Appendix B, simpler selection methods might perform better than the Chimera method for some datasets. The optimal weighting between the position and the sensitivity of a pulsar will be the subject of future investigations.

The purpose of these ranking methods is not to discard the analysis of some pulsars but only to evaluate their contribution to the full PTA analysis. Even though these results depend on the noise properties of the PTA dataset considered, the selection of a subset of pulsars has been shown to be a good proxy for having an informative dataset and at the same time reducing the computational burden of the analysis. Therefore, if a collaboration decides to limit pulsar sources due to resource restrictions, these tools will be essential for understanding how to make such a selection. These methods will be crucial to extend the array of existing experiments and target specific analyses when the next generation of radio facilities discover a large number of new pulsars.

ACKNOWLEDGEMENTS

We thank Stanislav Babak, Golam Shaifullah, Anuradha Samajdar, David Champion, Aditya Parthasarathy for useful discussions. We are very thankful to the anonymous referee for improving this manuscript. SRT acknowledges support from NSF AST-2007993, the NANOGrav NSF Physics Frontier Center PHY-2020265, and an NSF CAREER Award PHY-2146016. AS acknowledges financial support provided under the European Union's H2020 ERC Consolidator Grant "Binary Massive Black Hole Astrophysics" (B Massive, Grant Agreement: 818691). We made use of numpy and scipy (Virtanen et al. 2020; Harris et al. 2020).

DATA AVAILABILITY

The timing data and codes used in this article shall be shared on reasonable request to the corresponding authors.

REFERENCES

Ali-Haïmoud Y., Smith T. L., Mingarelli C. M. F., 2020, Phys. Rev. D, 102, 122005
Ali-Haïmoud Y., Smith T. L., Mingarelli C. M. F., 2021, Phys. Rev. D, 103, 042009
Antoniadis J., et al., 2022, Monthly Notices of the Royal Astronomical Society, 510, 4873–4887
Arzoumanian Z., et al., 2015, ApJ, 810, 150
Arzoumanian Z., et al., 2016, ApJ, 821, 13
Arzoumanian Z., et al., 2020, ApJ, 905, L34
Babak S., Sesana A., 2012, Physical Review D, 85
Babak S., et al., 2015, Monthly Notices of the Royal Astronomical Society, 455, 1665–1679
Bailes M., et al., 2018, arXiv e-prints, p. arXiv:1803.07424
Caballero R. N., et al., 2018, MNRAS, 481, 5501
Camilo F., et al., 2018, Astrophys. J., 856, 180
Chamberlin S. J., Creighton J. D., Siemens X., Demorest P., Ellis J., Price L. R., Romano J. D., 2015, Physical Review D, 91
Champion D. J., et al., 2010, ApJ, 720, L201
Chen S., et al., 2021, Monthly Notices of the Royal Astronomical Society, 508, 4970
Cordes J. M., Shannon R. M., 2010, arXiv e-prints, p. arXiv:1010.3785
Cornish N. J., Sampson L., 2016, Phys. Rev. D, 93, 104047
Desvignes G., et al., 2016, MNRAS, 458, 3341
Detweiler S., 1979, The Astrophysical Journal, 234, 1100
Dewdney P. E., Hall P. J., Schilizzi R. T., Lazio T. J. L., 2009, Proceedings of the IEEE, 97, 1482
Edwards R. T., Hobbs G. B., Manchester R. N., 2006, Monthly Notices of the Royal Astronomical Society, 372, 1549
Efstathiou G., 2004, MNRAS, 349, 603
Ellis J. A., Vallisneri M., Taylor S. R., Baker P. T., 2020, ENTERPRISE: Enhanced Numerical Toolbox Enabling a Robust PulsaR Inference SuitE, Zenodo, doi:10.5281/zenodo.4059815, https://doi.org/10.5281/zenodo.4059815
Foster R. S., Backer D. C., 1990, ApJ, 361, 300
Goncharov B., et al., 2021, ApJ, 917, L19
Gorski K. M., Hinshaw G., Banday A. J., Bennett C. L., Wright E. L., Kogut A., Smoot G. F., Lubin P., 1994, ApJ, 430, L89
Guo Y. J., Li G. Y., Lee K. J., Caballero R. N., 2019, MNRAS, 489, 5573
Hallinan G., Ravi V., team D. S. A., 2021, Bulletin of the AAS, 53
Harris C. R., et al., 2020, Nature, 585, 357
Hellings R. W., Downs G. S., 1983, ApJ, 265, L39
Hivon E., Górski K. M., Netterfield C. B., Crill B. P., Prunet S., Hansen F., 2002, ApJ, 567, 2
Hobbs G., Manchester R., Teoh A., Hobbs M., 2004, The atnf pulsar catalogue (arXiv:astro-ph/0309219)
Hobbs G. B., Edwards R. T., Manchester R. N., 2006, Monthly Notices of the Royal Astronomical Society, 369, 655–672
Hobbs G., et al., 2012, MNRAS, 427, 2780
Hobbs G., et al., 2020, MNRAS, 491, 5951
Homer Lattimore R., 2005, The Iliad of Homer. Recording for the Blind & Dyslexic, Princeton, N.J., http://www.worldcat.org/search?qt=worldcat_org_all&q=9780226469409
Janssen G., et al., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14). p. 37 (arXiv:1501.00127)
Jiang P., et al., 2019, Science China Physics, Mechanics, and Astronomy, 62, 959502
Johnson A. D., Vigeland S. J., Siemens X., Taylor S. R., 2022, ApJ, 932, 105
Joshi B. C., et al., 2018, Journal of Astrophysics and Astronomy, 39, 51
Kass R. E., Raftery A. E., 1995, Journal of the American Statistical Association, 90, 773
Kramer M., Champion D. J., 2013, Classical and Quantum Gravity, 30, 224009
Lam M. T., 2018, ApJ, 868, 33
Lam M. T., McLaughlin M. A., Cordes J. M., Chatterjee S., Lazio T. J. W., 2018, The Astrophysical Journal, 861, 12
Lee K. J., 2016, Prospects of Gravitational Wave Detection Using Pulsar Timing Array for Chinese Future Telescopes
Lee K. J., Bassa C. G., Janssen G. H., Karuppusamy R., Kramer M., Smits R., Stappers B. W., 2012, MNRAS, 423, 2642
Lentati L., Alexander P., Hobson M. P., Taylor S., Gair J., Balan S. T., van Haasteren R., 2013, Phys. Rev. D, 87, 104021
Lentati L., et al., 2015, Monthly Notices of the Royal Astronomical Society, 453, 2577–2599
Lorimer D. R., Kramer M., 2012, Handbook of Pulsar Astronomy
Manchester R. N., et al., 2013, Publ. Astron. Soc. Australia, 30, e017
McLaughlin M. A., 2013, Classical and Quantum Gravity, 30, 224008
Middleton H., Sesana A., Chen S., Vecchio A., Del Pozzo W., Rosado P. A., 2021, MNRAS, 502, L99
Mortlock D. J., Challinor A. D., Hobson M. P., 2002, MNRAS, 330, 405
Murphy E. J., et al., 2018, arXiv preprint arXiv:1810.07524
Ng C., 2018, in Weltevrede P., Perera B. B. P., Preston L. L., Sanidas S., eds, IAU Symposium Vol. 337, Pulsar Astrophysics the Next Fifty Years. pp 179–182 (arXiv:1711.02104), doi:10.1017/S1743921317010638
Peebles P. J. E., 1973, ApJ, 185, 413
Perera B. B. P., et al., 2019a, Mon. Not. Roy. Astron. Soc., 490, 4666
Perera B. B. P., et al., 2019b, MNRAS, 490, 4666

Pol N. S., et al., 2021, Astrophys. J. Lett., 911, L34
Roebber E., 2019, The Astrophysical Journal, 876, 55
Romano J. D., Cornish N. J., 2017, Living Reviews in Relativity, 20, 2
Romano J. D., Hazboun J. S., Siemens X., Archibald A. M., 2021, Phys. Rev. D, 103, 063027
Rosado P. A., Sesana A., Gair J., 2015, Monthly Notices of the Royal Astronomical Society, 451, 2417–2433
Sazhin M., 1978, Sov. Astron, 22, 36
Shannon R. M., Cordes J. M., 2010, ApJ, 725, 1607
Siemens X., Ellis J., Jenet F., Romano J. D., 2013, Classical and Quantum Gravity, 30, 224015
Speagle J. S., 2020, MNRAS, 493, 3132
Spiewak R., et al., 2022, arXiv e-prints, p. arXiv:2204.04115
Swarup G., 1990, Indian Journal of Radio and Space Physics, 19, 493
Taylor S. R., 2021, arXiv e-prints, p. arXiv:2105.13270
Taylor S. R., Lentati L., Babak S., Brem P., Gair J. R., Sesana A., Vecchio A., 2017, Phys. Rev. D, 95, 042002
Tiburzi C., et al., 2016, MNRAS, 455, 4339
Vallisneri M., et al., 2020, ApJ, 893, 112
Verbiest J. P. W., et al., 2009, MNRAS, 400, 951
Verbiest J. P. W., et al., 2016, MNRAS, 458, 1267
Virtanen P., et al., 2020, Nature Methods, 17, 261
Wandelt B. D., Hivon E., Górski K. M., 2001, Phys. Rev. D, 64, 083003
van Haasteren R., Levin Y., 2013, MNRAS, 428, 1147
van Haasteren R., Vallisneri M., 2014, Phys. Rev. D, 90, 104012
van Haasteren R., Levin Y., McDonald P., Lu T., 2009, MNRAS, 395, 1005

APPENDIX A: IMPLEMENTATION OF DIFFERENT WEIGHTS FOR COUPLING MATRIX FORMALISM OPTIMIZATION

In this paragraph we provide further clarifications on the choice of the weighting function w_α from Eq.(13). As mentioned in the main text, the weights for the construction of the coupling matrix should have a direct correspondence to the relative sensitivity of a source in an array. Here, we tested the performance of the Coupling Matrix formalism using as the weighting function $\mathrm{SNR_A}$ raised to the power of 2, 4 and 6. The optimal performance is obtained using $\mathrm{SNR_A^4}$-weights. Coupling matrix selection with weights of lower power of $\mathrm{SNR_A}$ tends to pick pulsars with a triple-peak distribution on the sky (see Figure 1), while the individual sensitivity of a source is relegated to the background. The degradation of the efficiency of $\mathrm{SNR_A^6}$-weighting for the mock MeerTime dataset is due to a saturation of the coupling matrix by the high SNR pulsars, so that it becomes essentially insensitive to adding further sources of lower sensitivity, or in some cases even ill-defined. In order to evade the problem of saturation, we have proposed to use the eigenvalue-ratio δ_λ ($w^a = 1$) and the individual SNRs of the pulsars combined in a Chimera-like manner: $\delta_\lambda \prod_{a=1}^{N_{\rm psr}} \mathrm{SNR}_\mathrm{A}^a$. The performance of the latter method is comparable to the one of the Coupling Matrix formalism with $\mathrm{SNR_A^4}$-weights. The efficacy of the Coupling Matrix selection and its modifications is going to be investigated more thoroughly in future work on a broader range of datasets.

APPENDIX B: SIMPLE ALTERNATIVE SELECTION METHODS

Throughout the paper we compared our selection methods to a random pulsar selection, because only a random selection can be considered independent of the specifics of the datasets. However, such a selection method would not be adopted in a realistic setting. Therefore, we explore how the selection methods compare to more realistic, still simple, ranking criteria: selecting pulsars based on their lowest RMS noise and longest timespan.

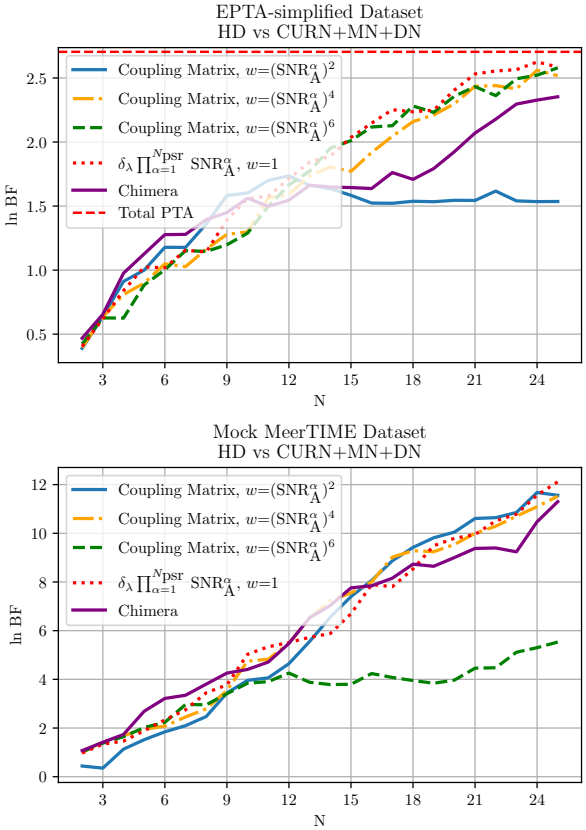

Figure A1. Log-Bayes factor of the hypothesis test HD vs CURN+MN+DN as a function of the number of pulsars selected by various modifications of the Coupling Matrix formalism (shown in different colors). The corresponding result for the Chimera method (purple color) are also shown for comparison. The upper panel shows the result for the simplified EPTA dataset averaged over 45 noise realizations, and the log-Bayes factor of the full array is indicated with a horizontal red dashed line. The bottom panel demonstrates the results for the mock MeerTime dataset.

For the case of the Galaxy-distributed dataset (Sec. 3.1.1) where all the pulsars have the same RMS and timespan, it is already clear that our ranking methods outperform a lowest-RMS selection or a longest-timespan selection, which are equivalent to the random selection. For the EPTA-simplified dataset (Sec. 3.1.3) and the Mock MeerTime dataset (Sec. 3.1.2) we perform only the RMS selection because all the pulsars' timespans are equal.

For the Mock MeerTime dataset (Fig. B1), the RMS selection method provides Bayes factors comparable to those of the Coupling Matrix and worse than the $\mathrm{SNR_B}$ and Chimera method, for the hypotheis test HD vs CURN. However, for the hypothesis test HD vs CURN+MN+DN, the RMS selection method performs better than all the others.

For the EPTA-simplified dataset (Sec. 3.1.3) the results are shown in Fig. B2. The RMS selection method provides Bayes factors comparable to the ones of the Chimera method for 25 pulsars and slightly smaller than the $\mathrm{SNR_B}$ method, for the hypothesis test HD vs CURN. For the hypothesis test HD vs CURN+MN+DN, the RMS selection method yields a Bayes factor comparable to the one of the $\mathrm{SNR_B}$ selection. The reason why for the hypothesis test HD vs CURN+MN+DN in the EPTA-simplified and mock MeerTime datasets the RMS selection performs better than other selection meth-

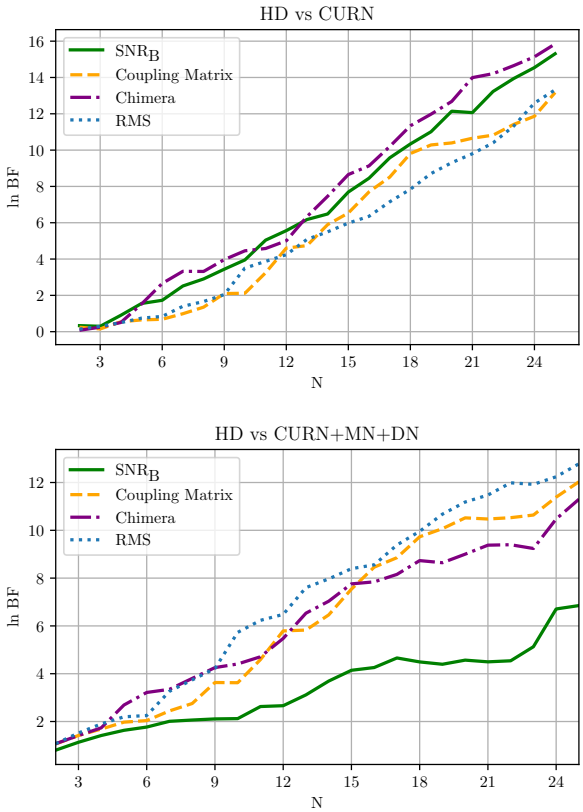

Figure B1. Log-Bayes factor as a function of the number of chosen pulsars for each of the selection methods (shown in different colors) for the Mock MeerTime dataset and for different hypothesis tests: HD vs CURN (top), and HD vs CURN+MN+DN (Bottom). The shown log-Bayes factors represent the average over 20 different noise realizations.

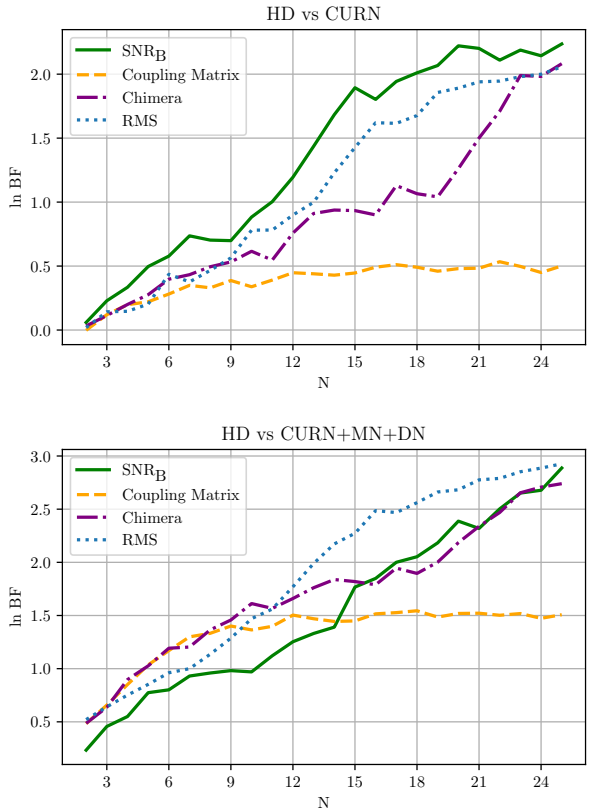

Figure B2. Log-Bayes factor as a function of the number of chosen pulsars by each of the selection methods (shown in different colors) for the EPTA-simplified dataset and for different hypothesis tests: HD vs CURN (top), and HD vs CURN+MN+DN (Bottom). The shown log-Bayes factors represent the average over 20 different noise realizations.

ods is that the lowest-RMS pulsars are almost uniformly distributed on the sky, so that the most sensitive pulsars of the array are picked in sufficiently optimal parts of the sky. For the arrays in which low-RMS pulsars are clustered in a specific region of the sky, this will not be the case. For the hypothesis test HD vs CURN, the RMS method does not differ significantly from the SNR-maximization, because the SNR formula already takes into account the RMS values and the aforementioned datasets are affected only by white noise.

For the realistic EPTA datasets (Sec. 3.2), we performed the lowest-RMS and longest-timespan selections, and we show the results in the top panel of Fig. B3. The lowest-RMS selection does not seem to differ from the SNR-maximization selection and it yields in median approximately the same log-likelihood ratio, which is ~ 0.87 times the total one. The longest-timespan selection performs slightly worse than the SNR-maximization and lowest-RMS selections, and it provides a log-likelihood ratio 0.71 times the one from the full dataset.

To highlight the difference between the lowest-RMS selection and the SNR-maximization selection we created a new dataset which is identical to the realistic EPTA dataset of Sec. 3.2, apart from the number of TOAs of each pulsar. The pulsars simulated for the realistic EPTA dataset have the same timespan as the real EPTA dataset, but with TOAs observed every 14 days. Now, the new dataset has the same number of TOAs as the real EPTA dataset and their TOA cadence range between one per day up to one every 18 days. The results of the same analysis of Sec. 3.2 are shown in the bottom panel of Fig. B3. Contrary to the previous results, the lowest-RMS selection method is now sub-optimal compared to the SNR-maximization method. The contribution to the total noise power due to white and red noise has changed as the TOA cadence is different. This has an impact on the selection methods. In fact, the SNR ranking recovers 88% of the total log-likelihood, whereas the lowest-RMS selection reaches only 79%.

Even if the SNR-ranking method does not perform as well as the RMS selection in some scenarios, it is more flexible and its relatively cheap computational cost makes it worth using it instead of RMS or longest timespan selection, when testing the HD vs CURN hypothesis.

This paper has been typeset from a TEX/LATEX file prepared by the author.

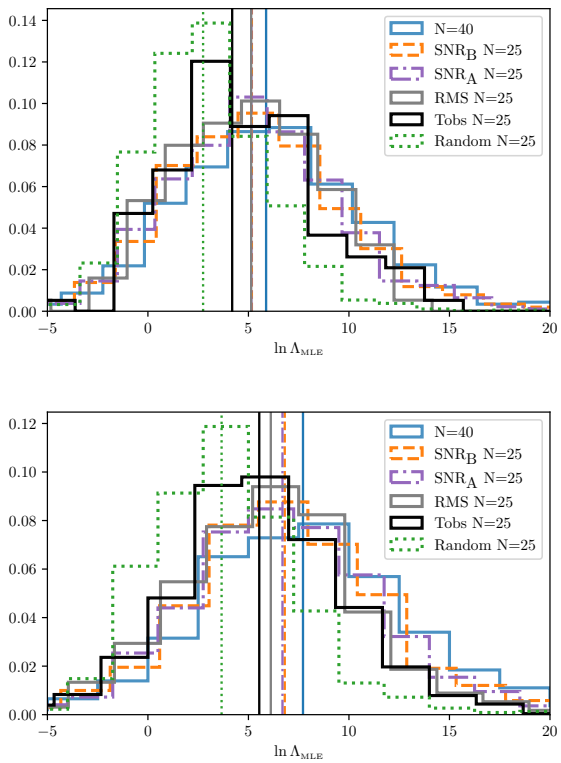

Figure B3. (*Top*): Distribution of log-likelihood ratios obtained as in Fig. 10 but with the addition of the distributions of log-likelihood ratios obtained with the lowest-RMS (RMS) and the longest-timespan (Tobs) selections. The median values for the shown distributions are: 5.88 (N=40), 5.17 (SNR$_B$), 5.14 (SNR$_A$), 5.13 (RMS), 4.19 (Tobs), 2.73 (Random). (*Bottom*): Same analysis as above but for the simulated realistic EPTA dataset with a number of TOAs as in the real EPTA dataset and not every 14 days as in the (simulated) realistic EPTA dataset. The median values for the shown distributions are: 7.71 (N=40), 6.80 (SNR$_B$), 6.69 (SNR$_A$), 6.11 (RMS), 5.53 (Tobs), 3.66 (Random).