Testing strengths, limitations, and biases of current pulsar timing arrays' detection analyses on realistic data

Serena Valtolina [1], Golam Shaifullah [2, 3, 4], Anuradha Samajdar[5, 6], Alberto Sesana[2, 3, 7]

1 Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Leibniz Universität Hannover, Callinstrasse 38, D-30167, Hannover, Germany
e-mail: serena.valtolina@aei.mpg.de
2 Dipartimento di Fisica "G. Occhialini", Universitá degli Studi di Milano-Bicocca, Piazza della Scienza 3, I-20126 Milano, Italy
3 INFN, Sezione di Milano-Bicocca, Piazza della Scienza 3, I-20126 Milano, Italy
4 INAF - Osservatorio Astronomico di Cagliari, via della Scienza 5, 09047 Selargius (CA), Italy
5 Institute for Gravitational and Subatomic Physics (GRASP), Utrecht University, Princetonplein 1, 3584 CC Utrecht, The Netherlands
6 Institut für Physik und Astronomie, Universität Potsdam, Haus 28, Karl-Liebknecht-Str. 24/25, 14476, Potsdam, Germany
7 INAF - Osservatorio Astronomico di Brera, via Brera 20, I-20121 Milano, Italy

March 25, 2024

ABSTRACT

State-of-the-art searches for gravitational waves (GWs) in pulsar timing array (PTA) datasets model the signal as an isotropic, Gaussian, and stationary process described by a power law. In practice, none of these properties are expected to hold for an incoherent superposition of GWs generated by a cosmic ensemble of supermassive black hole binaries (SMBHBs). This stochastic signal is usually referred to as the GW background (GWB) and is expected to be the primary signal in the PTA band. We performed a systematic investigation of the performance of current search algorithms, using a simple power-law model to characterise GW signals in realistic datasets. We used, as the baseline dataset, synthetic realisations of timing residuals mimicking the European PTA (EPTA) second data release (DR2). Thus, we included in the dataset uneven time stamps, achromatic and chromatic red noise, and multi-frequency observations. We then injected timing residuals from an ideal isotropic, Gaussian, single power-law stochastic process and from a realistic population of SMBHBs, performing a methodical investigation of the recovered signal. We found that current search models are efficient at recovering the GW signal, but several biases can be identified due to the signal-template mismatch, which we identified via probability-probability (P–P) plots and quantified using Kolmogorov-Smirnov (KS) statistics. We discuss our findings in light of the signal observed in the EPTA DR2 and corroborate its consistency with a SMBHB origin.

Key words. gravitational waves – black holes: physics – methods: data analysis – pulsars: general

1. Introduction

Recently, the pulsar timing array (PTA) community has announced the detection of a signal compatible with a gravitational wave (GW) origin. The latest datasets from the European PTA collaboration (EPTA, Antoniadis et al. 2023b), the Chinese PTA collaboration (CPTA, Xu et al. 2023), the North American Nanohertz Observatory for Gravitational Waves collaboration (NANOGrav, Agazie et al. 2023), and the Parkes PTA collaboration (PPTA, Reardon et al. 2023) all show evidence of a red noise process that is common among pulsars and shows correlation properties expected for the long sought after stochastic nano-Hertz (nHz) GW background (GWB, Hellings & Downs 1983).

Pulsar timing arrays are sensitive across the $10^{-9} - 10^{-7}$ Hz frequency range, where the dominant signal is expected to be an incoherent superposition of sinusoidal GWs emitted by a cosmic population of supermassive black hole binaries (SMBHBs, Rajagopal & Romani 1995; Jaffe & Backer 2003; Wyithe & Loeb 2003; Sesana et al. 2008; Rosado et al. 2015). Nonetheless, a correlated low-frequency signal in PTA data can also arise from GWs generated by early Universe phenomena, such as the non-standard inflationary scenario breaking the slow-roll consistency relations (e.g. Bartolo et al. 2007; Boyle & Buonanno 2008; Sorbo

2011), cosmic string networks (e.g. Damour & Vilenkin 2000), primordial curvature perturbations (e.g. Tomita 1967; Matarrese et al. 1993), turbulence arising in the aftermath of quantum chromodynamics (QCD) phase transitions (e.g. Kosowsky et al. 1992; Hindmarsh et al. 2014), and cosmic domain walls (e.g. Hiramatsu et al. 2014), or they might even originate from oscillations of the Galactic potential in the presence of ultra-light dark-matter (ULDM, Khmelnitsky & Rubakov 2014). Those scenarios inspired numerous studies aiming to test new physics in the early Universe (e.g. Vagnozzi 2023; Madge et al. 2023; Guo et al. 2023; Kitajima et al. 2023; Ellis et al. 2023; Cai et al. 2023; Figueroa et al. 2023; Franciolini et al. 2023) and have also been scrutinised by the EPTA+InPTA and NANOGrav collaborations in two comprehensive interpretation articles (Antoniadis et al. 2023a; Afzal et al. 2023). early Universe interpretations sometimes stress that the detected signal has a higher amplitude and flatter slope than what is expected from an astrophysical population of SMBHBs. This is an observation that has been challenged by Antoniadis et al. (2023a), who showed that a signal with the detected properties can naturally arise from a realistic ensemble of SMBHBs.

Under the hypothesis that the signal has an astrophysical, SMBHB origin, it is important to assess the performance of the

state-of-the-art PTA GWB search and parameter estimation algorithms as implemented in the PTA analysis suite `enterprise` (Ellis et al. 2019) on realistic data. In fact, current pipelines are searching for a stochastic GWB described as a Gaussian, isotropic, and stationary process characterised by a power-law Fourier spectrum, imprinting a correlated red signal in PTA residuals. The peculiar feature that allows us to disentangle those GW-induced delays from other noises is the inter-pulsar spatial correlation, first derived by Hellings & Downs (1983). Although pipelines searching for individual continuous GWs (CGWs), anisotropy, and broken power-law spectra exist (Babak & Sesana 2012; Ellis et al. 2012; Sampson et al. 2015; Taylor et al. 2020), the evidence for a GW-signal claimed by the different PTA collaborations is based on the aforementioned default assumptions. Conversely, none of these statistical properties are expected to hold for a signal produced by a cosmic ensemble of SMBHBs. Environmental effects and small number statistics are expected to produce spectra that significantly deviate from a smooth single power law (Sesana et al. 2008; Ravi et al. 2014). Furthermore, strong individual CGWs might produce extra power at specific frequencies, also resulting in highly anisotropic and non-Gaussian signals (Sesana et al. 2009; Ravi et al. 2012; Kelley et al. 2018). Finally, the eccentricity of SMBHBs might break the assumption of stationarity (Sesana 2013).

In a nutshell, there is a clear signal-template mismatch situation, where single power-law, Gaussian, isotropic, stationary templates are used to search for a signal that is unlikely to satisfy any of these hypotheses. Assessing the influence of this basic fact on the reliability of the analysis outcome is of paramount importance, especially in light of the recent PTA claims and of the numerous subsequent interpretation papers that rely on these results.

These issues were first investigated in Cornish & Sampson (2016), who demonstrated, using realistic GW models from (Rosado et al. 2015), that the isotropy hypothesis has only a mild effect on the detectability of the GW signal. Here we take a decisive step forward by conducting a systematic investigation of signal recovery on mock data designed to capture all the complexity of real PTA datasets. We generate mock versions of the EPTA `DR2new` dataset, described in Antoniadis et al. (2023e), including uneven time-stamps, white noise, chromatic and achromatic red noise, and multi-frequency observations (Antoniadis et al. 2023d). We then inject in our mock dataset two types of signals: (i) a Gaussian, isotropic, stationary GWB created with the `libstempo` (Vallisneri 2020)[1] function `createGWB`, and (ii) an incoherent superposition of the residuals induced by a population of circular GW-driven SMBHBs. In this latter case, residuals from both the pulsar and Earth terms of each binary are added one by one to the data. This results in a total GWB Fourier spectrum which on average is well fitted by a power-law function of frequency, but is generally much more structured. In particular, bright sources close to Earth can produce very pronounced peaks in power, breaking signal Gaussianity and isotropy at higher frequencies (Sesana et al. 2008). The rationale behind this choice is to test for differences in the performance of the GWB detection pipeline implemented in `enterprise` when (i) the injected signal matches the power-law template, and (ii) when there is a clear mismatch between the injected signal and power-law template.

The paper is structured as follows. In Section 2 we describe the simulated dataset, from the intrinsic pulsar noise to the realistic GWB modelling pipeline, and we present the signal recovery model implemented in the `enterprise` package. Section 3

presents the results of our simulations, which include two sets of `createGWB` injections (a strong and a weak signal), and a set of injections from a realistic SMBHB population. In Section 4, we discuss in detail selected realisations of the realistic injections, which help in understanding biases and issues that can arise from the signal-template mismatch. As we were completing this work, Bécsy et al. (2023) presented an independent, parallel investigation that touches on several points examined here. We subsequently discuss similarities and differences between the two works, together with further developments and future directions in Section 5.

2. Simulated data and analysis methods

The main goal of this work is to test our ability to recover a realistic GWB signal in a simulated data set with our existing model. To emulate the complexities of real data, we base all our simulations on the second data release from the EPTA collaboration (Antoniadis et al. 2023b,e,d). In particular, we use the 25 best EPTA pulsars, selected following (Speri et al. 2022). For these, we use the latest estimates of pulsar intrinsic red noise (RN) and dispersion measure (DM) variations to generate different simulated copies of the recently released `DR2new` dataset. `DR2new` is a reduced version of the entire second EPTA data release (`DR2full`), which includes only the last 10.3 years of observations, collected with the new generation wide-band backends.

In the following, we describe how we simulate PTA data using `libstempo` tools, a python wrapper around the `TEMPO2` pulsar timing software (Hobbs et al. 2006; Edwards et al. 2006), and how we obtain realistic GW-induced residuals from state of the art population of SMBHBs.

2.1. Timing model and pulsar noise properties

The timing models (TM) of the 25 pulsars are defined by the corresponding parameter files from the EPTA `DR2new`. We use `libstempo` to simulate times of arrival (ToAs) for a total observation time of 10.3 years, with a cadence defined by the observations, taken by the five European radio telescopes: the 100m Effelsberg radio telescope in Germany, the Lovell telescope at Jodrell Bank Observatory in the United Kingdom, the Nançay radio telescope operated by the Nançay Radio Observatory in France, the Sardinia Radio Telescope in Italy, and the Westerbork Synthesis Radio Telescope in the Netherlands. ToAs are simulated at two different frequency bands: 1400 and 2200 MHz. Having ToAs at different frequencies allows us to disentangle pulsar intrinsic RN from DM variations. We assume that an initial fit of the TM, obtained with `libstempo`, reduces it to a linear model where the coefficients are given by a design matrix. Following Van Haasteren et al. (2009), we analytically marginalise the likelihood over the TM parameter errors described by that linear model.

To make our simulations as close as possible to the real data, we include stochastic noise in our datasets. Stochastic noise in pulsar observations is customarily divided into three components: white noise, achromatic, and chromatic red noise. We include all three components in our simulations, following the analysis performed in Antoniadis et al. (2023d), based on the optimisation procedure outlined by Chalumeau et al. (2021).

To model white noise, we distribute ToAs around the values predicted by the TM with a root-mean-square (rms) uncertainty given by:

$$\sigma = \sqrt{\text{EFAC}^2 \sigma_{\text{ToA}}^2 + \text{EQUAD}^2} \, , \tag{1}$$

---

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

where σ_{ToA} is the uncertainty value due to template-fitting errors obtained in EPTA DR2, the EFAC factor takes into account the ToA measurement errors. The EQUAD, added in quadrature, accounts for any other white noise, such as stochastic profile variations, and possible systematic errors. These parameters are specific for each observing backend. In our simulations, we defined EFAC and EQUAD parameters to be constant, setting EFAC = 1.0 and EQUAD = 10^{-6}, and identical for all pulsars. The ToA uncertainties, σ_{ToA}, used in our simulated data sets (see Table 1) are the maximum likelihood estimates obtained in EPTA DR2 (Antoniadis et al. 2023c).

The single-pulsar stochastic achromatic and chromatic red noises are time-correlated signals that can be modelled as a stationary Gaussian process with a power-law spectrum:

$$S_{\mathrm{RN/DM}}(f; A_{\mathrm{RN/DM}}, \gamma_{\mathrm{RN/DM}}) = \frac{A_{\mathrm{RN/DM}}^2}{12\pi^2} \left(\frac{f}{yr^{-1}}\right)^{\gamma_{\mathrm{RN/DM}}} yr^3, \quad (2)$$

where RN and DM refer to achromatic and chromatic red noise respectively. The achromatic red noise does not depend on the observing radio frequency and is commonly used in single-pulsar noise models to characterise the long-term variability of the pulsar spin. Conversely, chromatic red noise depends on the observing radio frequency and is due to dispersion measure (DM) variations. In fact, during its propagation, the pulsar radio emission interacts with the ionised interstellar medium (IISM), the Solar System interplanetary medium and the Earth's ionosphere. These interactions lead to frequency-dependent delays in the observed signal:

$$\Delta_{\mathrm{DM}} \propto \nu^{-2} \mathrm{DM}, \quad (3)$$

where ν is the radio observing frequency and DM is the path integral of the free-electron density along the line of sight to the pulsar. We take this effect into account in our timing model, which considers the DM value at a reference epoch together with its first and second derivatives. However, the inhomogeneous and turbulent nature of the IISM also induces stochastic variations in the DM value, which are modelled as chromatic red noise. DM variations become more and more important on the decade-long timescales of PTA data (see e.g. Keith et al. 2012).

We used the `libstempo` package to inject in each pulsar RN and DM with spectra given by Eq. (2). In particular, for each pulsar, we defined the values of $A_{\mathrm{RN/DM}}$ and $\gamma_{\mathrm{RN/DM}}$ to be equal to the recent maximum likelihood estimates from the EPTA DR2 (Antoniadis et al. 2023b). Those values were obtained from joint analysis runs on the `DR2new` data set (Antoniadis et al. 2023c) when an additional common red process was included alongside individual pulsar noise terms. We report a complete summary of the individual pulsar's noise parameter values in Table 1.

As mentioned above, we simulate stochastic noise following the results of the customised noise model analysis carried out in Antoniadis et al. (2023d). Thus, for three pulsars J0030+0451, J1455-3330 and J2322+2057, we simulate RN only. Five pulsars J0900-3144, J1012+5307, J1022+1001, J1713+0747, and J1909-3744 have both RN and DM variations, and for the remaining seventeen we include DM variations only.

2.2. GWB induced residuals: createGWB vs realistic SMBHB populations

Having addressed the system and pulsar-related properties of real data, the only thing necessary to complete our data set is the GW

signal. Following Rosado et al. (2015), we generate the stochastic GWB-induced signal from the incoherent superposition of individual sinusoidal GW signals emitted by inspiralling SMBHBs.

2.2.1. Ideal signal with createGWB

As described in Phinney (2001), the characteristic strain of the GW signal produced by a population of circular, GW-driven SMBHBs is the integral of the energy emitted by each system over the differential number density of sources per unit redshift z, and chirp masses $\mathcal{M}$[2]:

$$h_c^2(f) = \frac{4}{\pi f^2} \int \int \int dz \, dM_1 \, dq \, \frac{d^2 n}{dz d\mathcal{M}} \frac{1}{1+z} \frac{dE_{gw}(\mathcal{M})}{d\ln f_r}, \quad (4)$$

where $dE_{gw}(\mathcal{M})/d\ln f_r$ is the energy spectrum emitted by each source (binary). In the circular GW-driven approximation, we can rewrite the emitted energy spectrum as a function of the binary chirp mass and GW rest-frame frequency f_r:

$$\frac{dE_{gw}(\mathcal{M})}{d\ln f_r} = \frac{\pi^{2/3}}{3} \mathcal{M}^{5/3} f_r^{2/3}. \quad (5)$$

Here, f_r is defined as $f_r = (1+z)f$ and is twice the binary Keplerian rest-frame frequency. By inserting Eq. (5) into Eq. (4), it is straightforward to show that

$$h_c(f) = A_{\mathrm{GWB}} \left(\frac{f}{1yr^{-1}}\right)^{\alpha_{\mathrm{GWB}}}, \quad (6)$$

where A_{GWB} is a model-dependent amplitude of the signal at the reference frequency $f = 1\mathrm{yr}^{-1}$ and $\alpha_{\mathrm{GWB}} = -2/3$. The corresponding spectral density S_{GWB} takes the form:

$$S_{\mathrm{GWB}}(f) = \frac{h_c^2(f)}{12\pi^2 f^3} = \frac{A_{\mathrm{GWB}}^2}{12\pi^2} \left(\frac{f}{1yr^{-1}}\right)^{-\gamma_{\mathrm{GWB}}} yr^3, \quad (7)$$

with $\gamma_{\mathrm{GWB}} = 3 - 2\alpha_{\mathrm{GWB}} = 13/3$. We note that the form of Eq. (7) is very similar to that of the intrinsic RN spectra (Eq. (2)). In fact, in a PTA dataset, a stochastic GWB appears as a red noise that is common to all pulsars and induces a specific angular correlation among pulsar pairs. This correlation is expected to follow, on average, the Hellings & Downs (1983) overlap reduction function. Other examples of red signals correlated over all pulsars are clock and ephemeris errors, which produce, respectively, a monopole and a dipole correlation in the pulsar residuals (Tiburzi et al. 2016).

Using `libstempo`, it is possible to inject a GWB signal into a PTA dataset through the function *createGWB*. By default, this function simulates the GW-induced delays in a PTA dataset as a common RN, HD-correlated over pulsars and with a smooth power-law shaped spectrum with index $\alpha = 13/3$ (Eq. (7)). The only input parameter supplied is the amplitude of the GWB signal. See discussion in Chamberlin et al. (2015) for more details.

2.2.2. Realistic signal from an SMBHB population

Eq. (4) models the characteristic strain of the GW signal as a smooth continuous function of the frequency, as given by Eq. (5). In reality, the expected astrophysical signal is the incoherent

---

[2] The chirp mass is defined as $\mathcal{M} = (M_1 M_2)^{3/5}/(M_1 + M_2)^{1/5}$, where $M_2 < M_1$ are the masses of the two black holes forming the binary. In the circular GW-driven approximation, at the quadrupolar order, the signal depends only on this combination of the two masses.

Pulsar	d [kpc]	cadence [days]	σ_{ToA} [μs]	noise model	$\log_{10}A_{\mathrm{RN}}$	γ_{RN}	$\log_{10}A_{\mathrm{DM}}$	γ_{DM}
J0030+0451	1.1	1.1240	4.7321	RN	-16.2802	0.3401	-	-
J0613-0200	1.2	2.1498	2.3190	DM	-	-	-12.1099	2.2955
J0751+1807	0.9	1.5237	2.5138	DM	-	-	-11.5769	2.1921
J0900-3144	1.3	0.6809	4.6006	RN+DM	-12.7817	0.9483	-11.8937	4.5107
J1012+5307	2.1	0.8983	3.1012	RN+DM	-12.9283	1.4274	-12.6575	3.9428
J1022+1001	0.3	2.0854	3.3792	RN+DM	-17.3298	6.2477	-11.4570	0.2273
J1024-0719	0.2	1.7813	4.5679	DM	-	-	-11.9266	2.6198
J1455-3330	1.1	1.5463	11.6450	RN	-13.1206	1.4706	-	-
J1600-3053	0.7	1.4481	0.8195	DM	-	-	-12.8008	5.7666
J1640+2224	0.9	2.4334	5.7051	DM	-	-	-11.4996	0.3000
J1713+0747	0.8	0.9403	0.9645	RN+DM	-15.0142	3.0124	-12.1243	1.6056
J1730-2304	1.3	3.2348	3.4269	DM	-	-	-11.6328	1.5792
J1738+0333	1.5	5.0228	6.7555	DM	-	-	-11.2123	1.5681
J1744-1134	2.2	2.4413	1.9787	DM	-	-	-11.7996	0.7458
J1751-2857	1.5	12.3347	8.4155	DM	-	-	-11.0296	1.0261
J1801-1417	0.6	9.7971	6.0490	DM	-	-	-11.0496	2.2886
J1804-2717	1.1	5.8057	8.6381	DM	-	-	-11.2871	0.0786
J1843-1113	1.3	5.1115	2.3503	DM	-	-	-11.0364	2.3694
J1857+0943	0.6	3.2544	2.6596	DM	-	-	-12.4013	4.6670
J1909-3744	1.3	1.6435	0.5931	RN+DM	-16.8142	1.8981	-11.9277	1.6375
J1910+1256	2.3	8.1784	4.1336	DM	-	-	-11.9094	3.4703
J1911+1347	1.8	4.6388	2.0747	DM	-	-	-12.1507	3.1562
J1918-0642	1.0	3.3059	2.8809	DM	-	-	-12.3077	4.1191
J2124-3358	0.5	2.3498	6.6360	DM	-	-	-11.4152	0.6247
J2322+2057	0.6	5.5817	13.8094	RN	-15.1459	0.4594	-	-

Table 1: Values for the distance, timing, and noise parameters of the 25 best EPTA pulsars. The final four columns list the maximum likelihood values obtained from the data set `DR2new`, using customised noise models when a common red noise is also included in the recovery model (Antoniadis et al. 2023b). Thus, the missing RN and DM parameters refer to the fact that, according to the customised noise model, there is no relevant evidence for the given process in the EPTA `DR2new` data set.

superposition of independent sinusoidal waves produced by an ensemble of SMBHB systems. For this case, Eq. (4) takes the form (Sesana et al. 2008)

$$h_c^2(f) = \int_0^\infty \mathrm{d}z \int_0^\infty \mathrm{d}\mathcal{M} \frac{\mathrm{d}^3 N}{\mathrm{d}z\mathrm{d}\mathcal{M}\mathrm{dln}f_r} h^2(f_r) , \qquad (8)$$

where now $\mathrm{d}^3 N/(\mathrm{d}z\mathrm{d}\mathcal{M}\mathrm{dln}f_r)$ is the number of emitting systems per unit redshift, mass and logarithmic frequency interval, and the strain $h(f_r)$ is given by:

$$h(f_r) = \sqrt{2(a^2 + b^2)} \frac{(G\mathcal{M})^{5/3}(\pi f_r)^{2/3}}{c^4 r} . \qquad (9)$$

Here, r is the co-moving distance to the source and the functions $a = 1 + \cos^2\iota$ and $b = -2\cos\iota$ define the relative strength of the two strain polarisations as a function of the binary inclination angle ι (see Rosado et al. 2015, for details).

Eq. (8) can be further manipulated by considering that the signal comes from a finite collection of discrete sources, and the spectrum is practically constructed in discrete frequency bins $\Delta f = 1/T$, where $T = 10.3\mathrm{yr}$ is the duration of the PTA experiment. The characteristic strain can be thus written as

$$h_c^2(f_i) = \sum_{j \in \Delta f_i} \frac{h_j^2(f_r)f_r}{\Delta f_i} , \qquad (10)$$

where f_i is the central frequency of the bin Δf_i, and the sum runs over all the systems for which $f_r/(1 + z) \in \Delta f_i$.

To practically inject the signal from a cosmic population of MBHBs in our PTA timing residual, we proceed as follows. The

list of emitting binaries is randomly sampled from the numerical distribution $\mathrm{d}^3 N/(\mathrm{d}z\mathrm{d}\mathcal{M}\mathrm{dln}f_r)$, which is obtained from the empirical, observation-based models described in Sesana (2013). The starting point is the galaxy merger rate, expressed as:

$$\frac{\mathrm{d}^3 n_g}{\mathrm{d}z\mathrm{d}M_g\mathrm{d}q_g} = \frac{\phi(M_g, z)}{M_g\ln 10} \frac{\mathcal{F}(z, M_g, q_g)}{\tau(z, M_g, q_g)} \frac{\mathrm{d}t_r}{\mathrm{d}z} , \qquad (11)$$

where the subscript 'g' stands for 'galaxy'. Here $\phi(M_g, z)$ and $\mathcal{F}(z, M_g, q_g)$ are the galaxy mass function and the galaxy differential pair fraction function at redshift z. Those quantities can be directly measured from observations, while the typical merger time scale $\tau(z, M_g, q_g)$ can be inferred by detailed simulations of galaxy mergers. The galaxy mass is then related to the SMBH mass via scaling relations of the form:

$$\log_{10}M_{\mathrm{BH}} = \alpha + \beta\log_{10}X , \qquad (12)$$

where X can be, depending on the model, the galaxy bulge mass, or its mid-infrared luminosity or velocity dispersion. We refer to Sesana (2013) for a list of those relations. Finally, SMBHs grow their mass through accretion in galaxy mergers. It is, however, unclear whether accretion mostly occurs before or after the SMBHB coalesces and, in the former case, whether accretion occurs preferentially on either of the two SMBHs.

For this work, we model $\phi(M_g, z)$ from Muzzin et al. (2013), $\mathcal{F}(z, M_g, q_g)$ from de Ravel et al. (2009) and $\tau(z, M_g, q_g)$ from Kitzbichler & White (2008). The SMBH mass is related to the galaxy mass via the $M - \sigma$ relation given by Kormendy & Ho (2013) and we assume that accretion occurs on the two black holes prior to the final merger, with preferential accretion on the
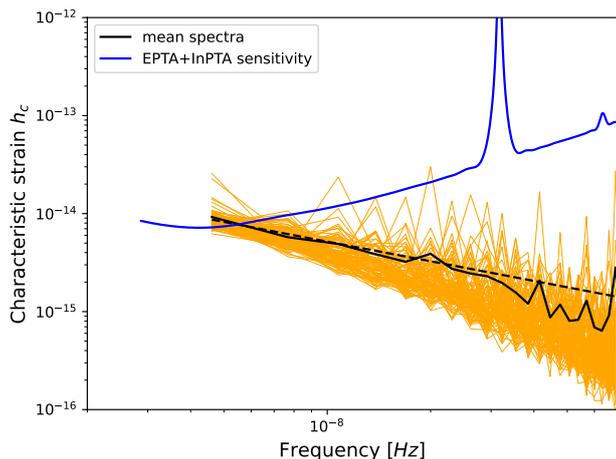
Fig. 1: Characteristic strain as a function of frequency for 100 different realisations of a GWB with nominal amplitude 2.4×10^{-15}. Each orange line corresponds to a different realisation (SMBHB population). The solid black line is the mean of those realisations, while the dashed black line highlights the nominal $f^{-2/3}$ spectrum. The sensitivity curve is derived in Collaboration et al. (2023).

secondary hole (Farris et al. 2014). These choices result in a GWB with nominal amplitude $A_{\mathrm{GWB}} = 2.4 \times 10^{-15}$ when computed with Eq. (4) and expressed in the power-law form given by Eq. (6); this value is consistent with the one inferred from the EPTA DR2new analysis (Antoniadis et al. 2023b).

We use this observation-driven astrophysical model to numerically construct the function $\mathrm{d}^3 N/(\mathrm{d}z\mathrm{d}\mathcal{M}\mathrm{d}\ln f_r)$. We then draw 100 Monte Carlo samples of this function in the appropriate mass, redshift and frequency region of the parameter space. Each draw results in $\approx$100K binaries, which we refer to as a *universe realisation*. For each binary, we specify its chirp mass, redshift, and GW-signal amplitude (Eq. (9)) in the observer frame, sky location coordinates, inclination and polarisation angles. All binaries are assumed to be circular.

Once the population of SMBHBs is defined, we construct the overall GW signal by directly injecting in the time domain the deterministic residulas imprinted by each individual system in the PTA dataset. To this end, we developed a custom *injection* pipeline, written partly in `python` (using `libstempo` functions) and partly in `fortran`. The script allows the user to take idealised pulsar timing models and produce ToAs for a given observing time span, add to each pulsar specific source noise as described in Section 2.1, and then add excess delays due to each of the SMBHB in the specified population. We inject both pulsar and Earth terms in the residuals, following the prescription of Babak et al. (2016). Similar injection pipelines have been applied to NANOGrav-like datasets in Bécsy et al. (2022) and Bécsy et al. (2023).

We can visually verify how this GWB signal definition method affects the spectra. The obtained characteristic strain amplitude for each of the 100 mock realisations of universe described above can be computed in the frequency domain from Eq. (10), and is shown in Fig. 1. As expected, the signal is much more structured than a plain $f^{-2/3}$ nominal power law computed through Eq. (4), with prominent spikes associated with rare, massive and (or) nearby sources. We note that the square-averaged signal sits on the theoretical curve, but there is considerable variance among different universe realisations.

2.3. Recovery model and analysis methods

To test the performance of current PTA GWB search and parameter estimation pipelines, we perform a set of inference analyses by using Bayes' theorem on the realistic datasets defined in Section 2. We search for individual and common noise parameters estimating model parameters characterised by their posterior probability distribution functions (PDFs), including those of RN, DM, and common noise. We then carry out Bayes-factor (BF) evaluations between HD-correlated and common, uncorrelated RN (CURN) models; and reconstruct the angular correlation from the data.

We conducted an in-depth analysis of each dataset using `enterprise` (*Enhanced Numerical Toolbox Enabling a Robust PulsaR Inference SuitE,* Ellis et al. 2019), a pulsar timing analysis package including functionalities for timing model evaluation, pulsar noise analysis, and GW searches. For each simulated array of pulsars we defined a noise model including: (i) EFAC and EQUAD parameters fixed respectively at 1.0 and 1e-6 for all pulsars, (ii) RN and DM variations according to the customised noise model used in EPTA `DR2new` (see Table 1), (iii) a common red noise process. We set the number of coefficients (i.e. the number of modes in the Fourier domain) for modelling the RN and DM processes in each pulsar to be, respectively, 30 and 100. We modelled DM variations as a Gaussian process.

Using this recovery model, we carried out Bayesian inference of the model parameter space using the Markov Chain Monte Carlo (MCMC) sampler included in `enterprise`: `PTMCMCSampler` (Ellis & van Haasteren 2017). Since we fixed the white noise parameters for each pulsar, the total number of parameters for the sampling is 62: 60 from pulsar intrinsic noise parameters and two parameters, log amplitude and slope, for the common red noise.

To decrease computational costs, we employed the *reweighting* method introduced in Hourihane et al. (2023). We first computed approximate posteriors defining the common process as an uncorrelated red noise. Thus, by temporarily ignoring the cross-correlation terms, the covariance matrix becomes block-diagonal, resulting in a much faster sampling. We then reweighted the obtained chains of samples to get the exact posteriors (corresponding to an HD-correlated common red process) via importance sampling. Besides being much faster than a direct search for an HD-correlated common signal, this method also provides an accurate estimate of the Bayes factor between the HD and the CURN models. In fact, the average of the weights computed for all samples is equal to the ratio of the marginal likelihood (or evidence) of the two considered models. While this method is mathematically exact, there are some limitations when the likelihood changes significantly between the two models. For example, as the GWB amplitude increases, the weights distribution becomes broader and the sampling efficiency of the method decreases. Hourihane et al. (2023) present a very detailed study of the limits of this method and conclude that the Bayes factor estimate stays robust up to $BF > 10^6$. See Hourihane et al. (2023) for a more detailed description.

To test the limits imposed by the sensitivity of the dataset and the consequences in the inference analysis, for some realisations we repeated the parameter estimation runs considering different numbers of Fourier components for the common process (see Section 4.2 for more details). The results presented in Sec. 3.1

Parameter	Prior type	Range
γ_{RN}, γ_{DM}, γ_{CURN}	Uniform	$[0, 7]$
A_{RN}	log-Uniform	$[10^{-18}, 10^{-11}]$
A_{DM}	log-Uniform	$[10^{-18}, 10^{-8}]$
A_{CURN}	log-Uniform	$[10^{-15.5}, 10^{-13.5}]$

Table 2: Prior distributions for the Bayesian inference analyses.

and 3.2 are obtained considering only the first nine frequency bins while searching for the common process.

We use uniform priors for the slope parameters and log-uniform priors for the amplitudes of the noise components. The prior ranges for the intrinsic noise and common process parameters are listed in Table 2.

For each simulated PTA dataset, we also computed the induced angular correlation in the timing residuals between pulsar pairs. We followed the method used for the frequentist analysis in EPTA `DR2new` (Antoniadis et al. 2023b). We used the *Optimal Statistic* (OS) framework developed by Anholm et al. (2009), Demorest et al. (2012), and Chamberlin et al. (2015), with the noise marginalisation described in Vigeland et al. (2018). We also computed the mean correlation and variance in the correlation recovery between different realisations of the same GWB signal, and compared the results with the theoretical predictions from Allen (2023). We followed the prescriptions in Allen & Romano (2023) when computing the average over pulsar pairs.

3. Results

Using the framework described in the previous section, we generate three sets of 100 mock EPTA `DR2new` datasets, for a total of 300 simulations. In each of these 300 simulations, the individual pulsar DM and RN are generated as a random realisation of a stochastic process described by the power-law spectra of Eq. (2), with amplitude and slope fixed to the ML value of the customised noise analysis performed in Antoniadis et al. (2023d) as reported in Table 1. The three sets of simulations differ for the injected GW signal:

- *LoudGWB_set*. We use *createGWB* to inject a loud, stochastic GWB with $A_{GWB} = 5 \times 10^{-15}$. This signal is easily detectable in the `DR2new` dataset and serves as a benchmark to test our simulations and analysis pipeline.
- *CreateGWB_set*. We use again *createGWB* to inject a stochastic GWB with $A_{GWB} = 2.4 \times 10^{-15}$, consistent with the signal observed in Antoniadis et al. (2023b). These simulations are meant to test the pipeline in the regime of a relatively weak signal that matches the template used in the likelihood evaluation.
- *SMBHB_set*. We inject individual residuals from an astrophysically motivated SMBHB population producing a GWB with a nominal average amplitude of $A_{GWB} = 2.4 \times 10^{-15}$. In this case, however, the signal is very different from the template used in the analysis pipeline, allowing us to investigate limitations and biases due to a mismatch between the signal present in the data and the model used in the analysis.

In Section 3.1 we discuss the performance of the pipeline applied to the *LoudGWB_set*; we then move to the comparison between the *createGWB_set* and the *SMBHB_set* in Section 3.2.
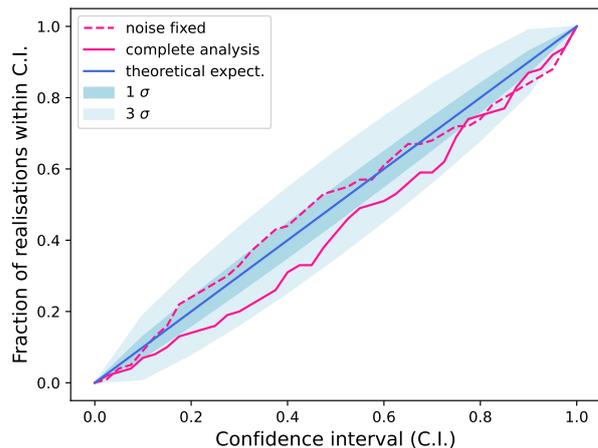
Fig. 2: P–P plot for the GWB-amplitude recovery for the *LoudGWB_set* of simulations. The dashed line refers to the inference runs where all the intrinsic noise parameters are fixed in a noise dictionary and the amplitude of the GWB is the only free parameter. The solid line is the result of MCMC runs over all 62 noise parameters. The theoretical expectation for an unbiased recovery and the predicted variance of one- and three- σ are represented in different shades of blue.
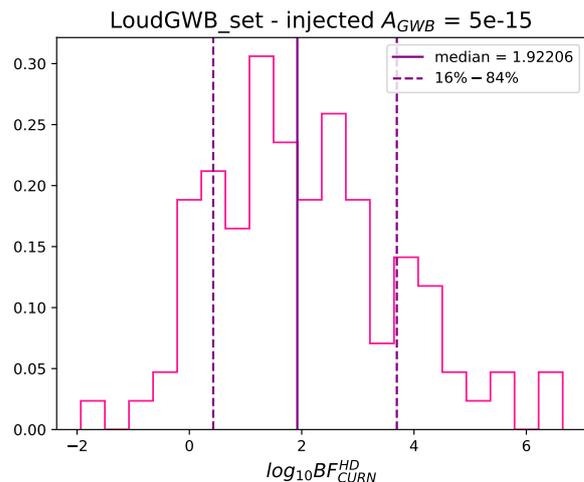
Fig. 3: $\log_{10}$BF distribution of HD vs CURN for the 100 realisations of the *LoudGWB_set*. The vertical lines show the median (solid) and the 16th and 84th percentile (dashed) of the distribution.

3.1. The LoudGWB_set of simulations: A benchmark for recovery convergence

As already mentioned, this set of simulations aims to define a benchmark for recovery convergence. In fact, the injected high GWB amplitude ($A_{GWB} = 5 \times 10^{-15}$) is expected to be relatively easy to recover and disentangle from pulsar noise.

Some useful tools to validate the performance of our model are the cumulative distributions of the number of times the injected value lies within a credible interval, the so-called probability-probability (P–P) plots (see Cook et al. 2006; Talts et al. 2020; Wilk & Gnanadesikan 1968, and others). These plots show on the *y* axis the fraction of times in which the nominal value
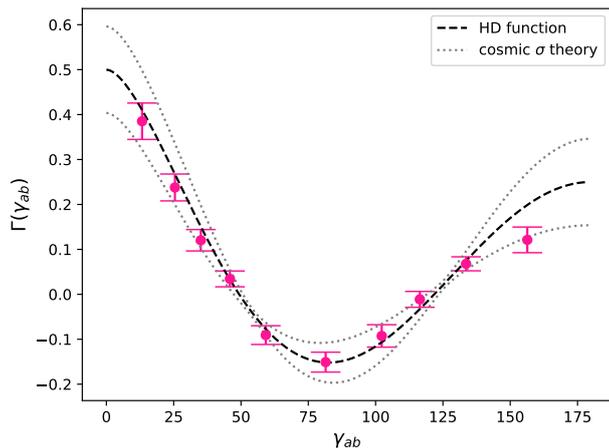
Fig. 4: HD correlation recovery for the *LoudGWB_set*. Each data point corresponds to the average of the optimal correlation estimators in that bin over all realisations. The expected cosmic variance is derived in Allen (2023).

(in our case, the injected A_{GWB}, normalised at the frequency of $1/1yr$) lies within the credible interval indicated on the x axis. In the case of unbiased inference, the data points follow the diagonal of the plot parameter space.

A P–P plot for the inferred amplitude of the recovered GWB is shown in Fig. 2. The theoretical expectation and the predicted variance (one-σ and three-σ levels) are shown in different shades of blue. The dashed line is obtained from inference runs where all the pulsar's intrinsic noise parameters and the slope of the GWB signal are fixed to the nominal injected values (Table 1, $\gamma_{\mathrm{GWB}} = 13/3$). Thus, the only free parameter is the amplitude of the GWB signal. In this case, the obtained distribution follows the diagonal within the one-σ variance and there is no evident bias in the recovery of A_{GWB}. This is expected since the GWB signal is injected via the *createGWB* function, which simulates a background with a spectrum very close to the nominal power law. The solid line, instead, is obtained from the full analysis sampling of all the 62 model parameters (the pulsars intrinsic RN and DM variation parameters, see Table 1, and the two GWB parameters). Although the recovered distribution is always within the three-σ expected interval, it systematically lies below the diagonal. This may indicate a slightly biased recovery towards lower amplitudes (and, consequently, higher γ_{GWB}) for the GWB spectrum. Although not particularly worrying, the origin of this potential bias is unknown and it might be due to leakage of power across different noise components.

Thanks to the reweighting method, we can also show the distribution of the obtained $\log_{10}$BF for the HD-correlated model versus CURN. The results are shown in Fig. 3. The median of the distribution (vertical solid line) is at $\log_{10}$BF = 1.91, which corresponds to a BF$\sim$ 81. The spread of the distribution highlights the impact of the stochastic nature of the pulsar noise in the signal recovery. The central 68% of the BF distribution spans more than three orders of magnitude, and depending on the specific realisation of the noise, the data can either provide decisive evidence of a GWB or an inconclusive result.

Finally, we compute the angular correlation induced in the residuals of each pulsar pair for all realisations. For an array of 25 pulsars, there are 300 independent pairs. For each realisation, we define ten angular separation bins of 30 pulsar pairs each, and
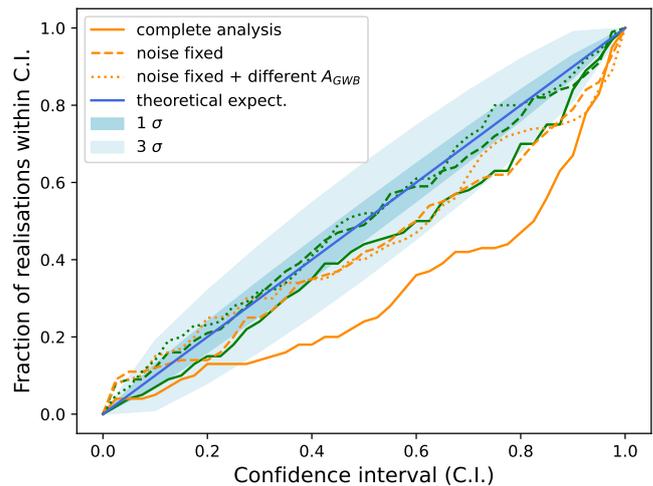
Fig. 5: P–P plot for the recovery of A_{GWB} from the *SMHBH_set* (orange) and *createGWB_set* (green). The solid lines are obtained from the $\log_{10}A_{\mathrm{GWB}}$ posteriors of the MCMC runs over all 62 noise parameters of the pulsars array. The dashed lines are obtained by fixing all noise parameters in the recovery model; thus, A_{GWB} is the only free parameter. The dotted lines are also from posteriors obtained sampling only over A_{GWB}, but in a dataset where the realisations do *not* all have the same spectral amplitude. Here, the injected background has an amplitude value extracted from the prior distribution used in the recovery.

compute the mean and variance of the correlation. The calculation follows the prescriptions in Allen & Romano (2023), taking into account the covariance between different pulsar pairs in the same angular separation bin. We then compute the mean over the whole *LoudGWB_set* as the average, in each angular separation bin, of such optimised correlation estimators. We present the result in Fig. 4, where each data point is computed over 3000 pulsar pairs: 30 for each realisation. For comparison, we also show the cosmic variance limit derived in Allen (2023) (see their Eq. 4.8 and G11). The pulsar variance contribution to the expected variance of the HD recovery is minimised by the weighted-average method described in Allen & Romano (2023); thus, in this case the cosmic contribution is the only significant one to compare our results with.

We note that the mean correlation estimated in each bin typically lies very close to the expected HD correlation. This is expected for such a loud GWB signal (5×10^{-15}). The only exception is the very last bin, which is difficult to constrain due to the limited number of pulsar pairs available at wide angular separation, forcing an averaging procedure over a wide bin.

3.2. Ideal vs real: Comparing the createGWB_set and the SMBHB_set

Having assessed the performance of the analysis pipeline on a loud, ideal signal, we now turn to the comparison of the signal recovery for the *createGWB_set* and the *SMBHB_set*.

As in the previous section, we start by constructing the P–P plot for the A_{GWB} parameter, shown in Fig. 5. For both data sets, the nominal value of A_{GWB} for each realisation is 2.4×10^{-15}, at the reference frequency of $1/1yr$. In the *SMBHB_set* this value corresponds to computing the GWB through Eq. (4) and express-
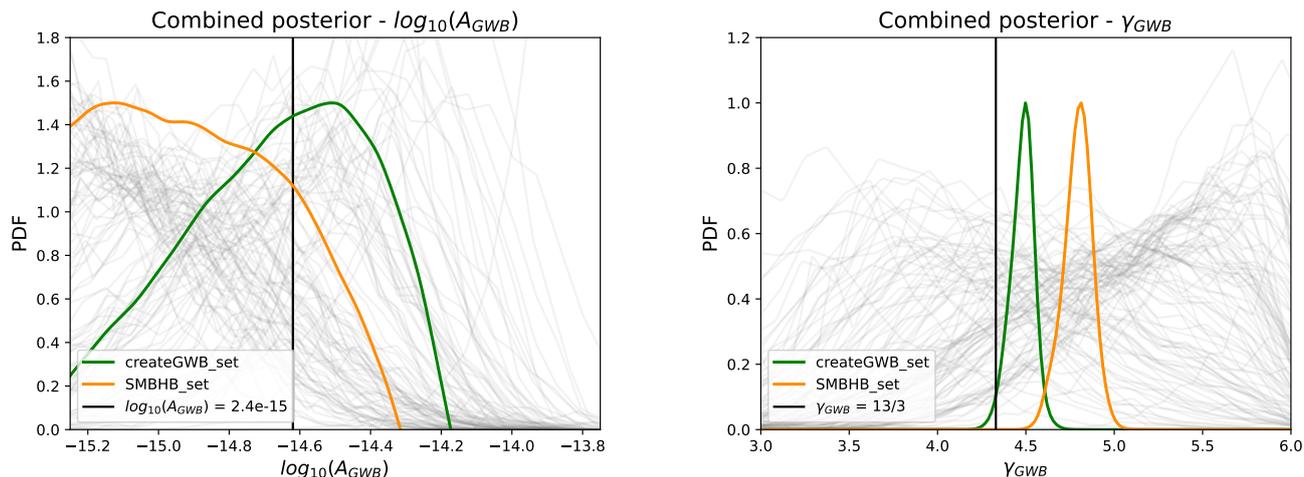
Fig. 6: Combined posterior for $\log_{10} A_{\text{GWB}}$ (left panel) and γ_{GWB} (right panel) parameter for the *createGWB_set* (green) and the *SMBHB_set* (orange). The injected values, marked by the black vertical lines, are $A_{\text{GWB}} = 2.4 \times 10^{-15}$ and $\gamma_{\text{GWB}} = 13/3$. The grey posteriors in the background are the ones obtained from the individual realisations of the *SMBHB_set*.

Model 1	Model 2	KS stat.	p-value
SMBHB_set noise fixed	diagonal	0.1707	0.5945
createGWB_set noise fixed	diagonal	0.0732	0.9999
SMBHB_set full analysis	diagonal	0.3590	0.0125
createGWB_set full analysis	diagonal	0.1538	0.7523
SMBHB_set noise fixed	*createGWB_set* noise fixed	0.1463	0.7789
SMBHB_set full analysis	*createGWB_set* full analysis	0.2821	0.0897

Table 3: KS statistic and correspondent p-values obtained from comparisons between different distributions shown in Fig. 5. With *diagonal* we refer to the theoretical prediction for unbiased recoveries, which corresponds to the diagonal of the plot. The other distributions analysed are the dashed and solid lines of that P–P plot. The p-value in the second raw is almost perfect (0.9999); this confirms the good fit between the A_{GWB} recovery for the *createGWB_set* when the other noise parameters are fixed.

Fig. 7: Histograms of the $\log_{10}\text{BF}^{\text{HD}}_{\text{CURN}}$ obtained from the reweighting analysis on the *createGWB* (green) and *SMBHB_set* (orange). The solid vertical lines show the medians of the distributions, while the dashed ones refer to the 16th and 84th percentiles. The black vertical line corresponds to the latest EPTA estimate for `DR2new`: $\text{BF}^{\text{HD}}_{\text{CURN}} \sim 60$.

ing it in the power-law form given by Eq. (6). The green lines are for the *createGWB_set*, while the orange ones are for the *SMBHB_set*. As in Fig. 2, the solid lines are computed using the posterior distributions of the MCMC runs sampling over all the 62 noise parameters of the array, while the dashed lines are obtained by searching over A_{GWB} keeping all other noise parameters fixed to the nominal value. When fixing all noise parameters, the *createGWB_set* closely follows the diagonal, indicating an unbiased recovery of the signal amplitude. Conversely, the *SMBHB_set* tends to be consistently below the diagonal, also crossing the three-σ confidence interval in the upper-right corner. When sampling over all the noise parameters, this bias is enhanced. As for the *LoudGWB_set*, the distribution for the *createGWB_set* (solid green) is still within the three-σ confidence interval, although systematically below the diagonal. The situa-

tion gets more extreme for the *SMBHB_set* (solid orange), which is dramatically biased towards the lower amplitude of the GWB spectra.

To quantify the distances between the different distributions shown in Fig. 5, we used the `scipy` package `scipy.stats.kstest` (Virtanen et al. 2020) to perform a series of non-parametric Kolmogorov-Smirnov tests (KS test). The test returns the maximum difference between two distributions and an estimate of the p-value under the null hypothesis that the two distributions are identical. Results are summarised in Table 3 and highlight the inconsistency of the signal recovery in the *SMBHB_set*.

We note that in these analyses there is a difference between the prior defined in the recovery model, $\log_{10} A_{\text{GWB}}$ uniform in
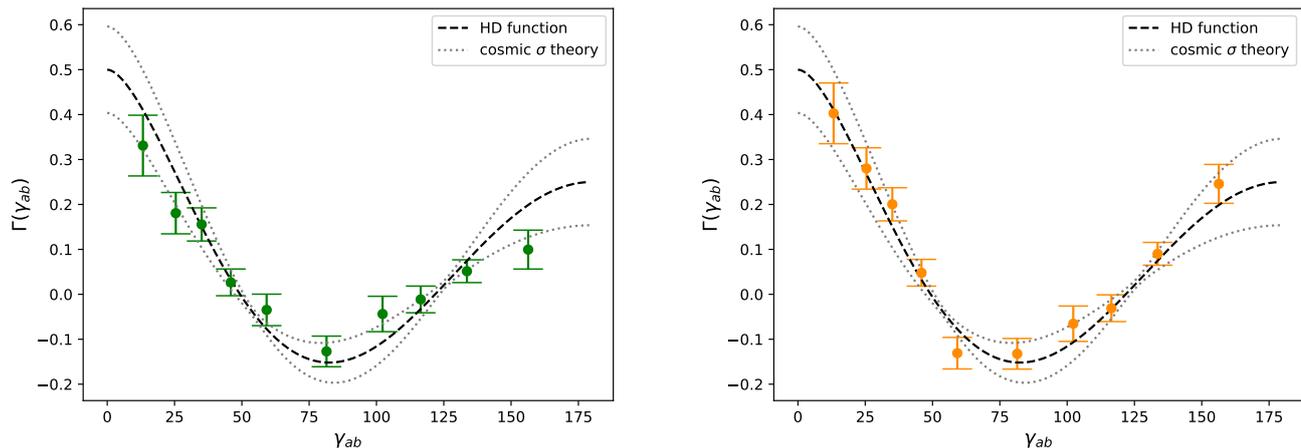
Fig. 8: Mean and variance of the angular correlation in the timing residuals of the *createGWB_set* (left panel) and for the *SMBHB_set* (right panel). Each data point corresponds to the average of the optimal correlation estimators in that bin over all realisations. We also show the expected cosmic variance derived in Allen (2023).

$[-15.5, -13.5]$, and the one from which the injected values are selected (basically a delta function centred in $A_{\mathrm{GWB}} = 2.4 \times 10^{-15}$). Since the latter prior is much narrower and completely included in the former, the statistical significance of the P–P plots are unaffected. To test this statement, we generated a new set of simulations where the pulsar's intrinsic noises are the same as for the other datasets and fixed, but now the GWB signal amplitudes have values which are randomly chosen from the uniform prior $\log_{10} A_{\mathrm{GWB}}$: $[-15.5, -13.5]$. The results are shown in Fig. 5 as dotted lines (green for the *createGWB_set*, orange for the *SMBHB_set*). As expected, there is no significant discrepancy between those lines and the dashed ones (same analysis on data where the GWB signal is the same in all realisations).

Since we have 100 simulations, each with a different realisation of population parameters but the same injected parameters of the GWB signal, we can utilise Bayes' theorem to obtain the combined and better constrained posterior PDFs for $\log_{10} A_{\mathrm{GWB}}$ and γ_{GWB}. In Fig. 6 we show the combined posteriors for $\log_{10} A_{\mathrm{GWB}}$ and γ_{GWB} for both the *createGWB_set* (green) and the *SMBHB_set* (orange). The posterior for the slope parameter is better constrained than that for the amplitude. While both parameters are compatible with the nominal injected value within $\sim$ one-σ for the *createGWB_set*, this is not the case for the *SMBHB_set*. The orange posteriors are clearly biased towards steeper spectra with lower amplitudes, consistent with what is shown by the P–P plots.

Systematic biases in the signal recovery for the *SMBHB_set* can be traced back to the h_c distribution of the individual realisations shown by the orange lines in Fig. 1. These realisations lie preferentially below the expected $f^{-2/3}$ line, featuring a steeper spectrum. Only sporadically, loud individual sources result in excess power at specific frequencies. When this happens, a power-law fit to the data can result in a flatter spectrum, above the $f^{-2/3}$ line. This is a general feature of realistic SMBHB populations characterised by a sparse, high-mass tail of loud GW sources. Although most signal realisations show a deficiency of power at high frequency, the few which feature loud sources ensure that the average signal amplitude sits on the expected $f^{-2/3}$ power law. It follows that the typical realisation of a nominal $f^{-2/3}$ power-law signal has in reality, a steeper spectrum. There is, therefore, a mismatch between the theoretically smooth GWB signal used in

the recovery model and the real signal produced by a discrete ensemble of SMBHBs, which can lead to systematic biases in the signal recovery and erroneous interpretation of the results.

We note that, during our whole analysis, we fixed the reference frequency for the recovery of the GWB amplitude to $1/1yr$. Changing this frequency to a lower one would result in a weaker dependence of A_{GWB} upon γ_{GWB}, but the one-dimensional posterior of the slope parameter remains unchanged. Reanalysing our data sets with the reference frequency set at $1/10yr$ shows that, as expected, the average recovery is still biased to higher γ_{GWB} values. In contrast, the A_{GWB} recovery is now less affected by the lack of power at higher frequencies, resulting in a combined posterior that agrees very well with the nominal amplitude value of the simulated GWB. The difference between the *SMBHB_set* and the *createGWB_set* also becomes not statistically significant. We refer to Sec. 2 of Antoniadis et al. (2023a) for further details on this point.

In this paper, we focus on testing the recovery of the GWB signal from PTA datasets. However, pulsar's intrinsic noise parameters are also free parameters in our inference runs and, thus, subject to possible biases in the recovery. We refer to Appendix A for a brief discussion on the recovery of pulsar's intrinsic RN parameters.

From the parameter estimation analysis carried out on the *createGWB_set* and on the *SMBHB_set*, we can build the distribution of the BFs of HD vs CURN obtained with the reweighting method in each realisation (see Sec. 2.3 for more details). Results are shown in Fig. 7. The median of the $\log_{10} BF$ distribution for the *createGWB_set* is ~ 0.5, while for the *SMBHB_set* is ~ 0.8. The estimated BF for EPTA `DR2new` is ~ 60 and is represented in the plot by the black vertical line. We note that this value is included in the 16th-84th percentile interval of the distribution for the *SMBHB_set*.

Finally, we also plot the recovered angular correlation in the timing residuals, following the procedure described in the previous section. In Fig. 8 we compare the mean and variance of the reconstructed correlation function to the HD curve and its cosmic variance (Allen 2023). In both cases, not only the different points are always compatible with the predicted HD correlation, but also the computed mean in each angular bin is typically included within the predicted cosmic variance.

4. A close look at astrophysical variance: Notable realisations of the *SMBHB_set*

So far, we focussed on the collective properties of the recovered signal, highlighting possible systematic biases due to the idealised model used in the analysis pipeline. It is also interesting, however, to have a close look at the 'zoology' of signals arising from a realistic SMBHB population, to get an idea of how specific features are reflected in the outcome of the analysis. We focus here on four selected realisations of the *SMBHB_set*, stressing that all of them are independent statistical realisations of the same underlying astrophysical model.

4.1. Case1 and Case2: Different spectra resulting in similar posteriors

The first two selected cases are characterised by the spectra shown in the left panel of Fig. 9. While *Case1* (brown line) has a flat, smooth bump in the lowest frequency bins, *Case2* (orange line) is characterised by a jagged behaviour due to few loud sources. In the right panel of Fig. 9 we show the GWB parameter posteriors obtained for those two realisations. The recovery model is the one described in Sec. 2.3. Although the two spectra are very different, the inference runs produce very similar posteriors for the GWB amplitude and slope. This can be qualitatively understood by comparing the spectra to the sensitivity curve associated to the considered PTA (blue curve in Fig. 9, left panel). In *Case1*, the array is mostly sensitive to the last couple of frequency bins, where the signal is flatter than the nominal $f^{-2/3}$ power law. This leads to a posterior with a slight preference for small γ and high A_{GWB}. Conversely, in *Case2*, besides detecting the signal at the lowest frequency bin, the inference pipeline picks some correlated power due to the marginally detectable loud source at $f \approx 1.5 \times 10^{-8}$ Hz. Since the built-in model is a single power law, this extra high-frequency power also results in a posterior with a slight preference for small γ and high A_{GWB}.

It is also interesting to notice that the recovered GWB parameters for these two selected realisations are quite consistent to the DR2new posteriors presented in Fig. 1 of Antoniadis et al. (2023b). This just exemplifies that there is no conflict between the signal observed in the latest PTA data and astrophysical expectations.

4.2. Case2 and Case3: Loud and louder sources.

To further demonstrate the impact of sparse, particularly massive and (or) nearby SMBHBs on the GWB signal recovery, we selected two realisations featuring some prominent loud sources. Those are *Case2*, already introduced in the previous subsection, and *Case3*. The characteristic strain frequency spectra of those two realisations are presented in the top panel of Fig. 10. The loud source at $f \approx 2 \times 10^{-8}$ present in *Case3* produces a peak in the power that is marginally above the nominal sensitivity of the simulated PTA, we can therefore expect a strong impact on the recovered signal.

This is shown in the central and bottom panels of Fig. 10. When modelling the signal with the nine lower Fourier components (frequency bins) in the GWB search, the presence of this peak biases the recovery towards low values of γ_{GWB} and high amplitudes (red posterior in the central panel). If, instead, we consider only the first four frequency bins, the bright source falls outside the frequency domain of the model and, as expected, the recovered posteriors are much closer to the nominal values for the injected SMBHB population (red posterior in the bottom panel).

A similar effect, although to a lesser extent, is seen for *Case2*. In this case, the loud source is less prominent, contributing only marginally to the overall detected power. Still, by comparing the orange posteriors in the central and bottom panels of Fig. 10, we can see a significant shift of the posterior to the lower right when restricting the model from nine to four frequency bins. The possible bias towards a flatter GWB spectrum due to a particularly bright GW source has been recently discussed also in Bécsy et al. (2023). Here we provided concrete examples of the effect on the recovery of the GWB spectra when bright sources are involved.

4.3. Case4: A familiar HD correlation recovery

Finally, we present a fourth interesting realisation. In *Case4* the characteristic strain as a function of frequency does not present any particularly pronounced peak (see the left panel in Fig. 11). The interesting feature observed while analysing this PTA dataset is the induced angular correlation in the residuals. In Fig. 11, right panel, we show the results obtained with the OS packages (as described in Sec. 2.3) for this specific realisation (light blue points), and compare them with the results obtained for the EPTA DR2new dataset (red points). Each data point of the plot represents the average correlation in a bin containing 30 pulsar pairs. The averages over pulsar pairs are computed following the prescriptions in Allen & Romano (2023).

The reconstructed angular correlation closely follows the HD prediction. What caught our attention is the close resemblance between the right panel of Fig. 11 and the results from the EPTA DR2new dataset (see Fig. 6 of Antoniadis et al. 2023b). In both plots, the considered pulsars are the same (EPTA best 25 pulsars) and the correlation is computed following the same procedure.

The similarity in the reconstructed HD is also reflected in the estimated BF for HD correlated common signal vs CURN. Using the reweighting technique in the inference run, we obtain BF≈ 62 for *Case4*, which is very consistent with the BF≈ 60 found in EPTA DR2new (Antoniadis et al. 2023b).

The BF (HD vs CURN) of those four notable cases are summarised in Table 4 for completeness, and they give a flavour of the role played by stochasticity in the estimate of signal significance and parameters. For example, the spectra of *Case1* and *Case4* look very similar, with perhaps *Case1* showing a bit more power in the lowest frequency bins. Still, in this case, the result of the analysis is inconclusive (BF= 1.179 for HD vs CURN), whereas in *Case4* there is strong evidence of an HD correlated process (BF= 62.5). This is because, in these early stages of detection, the output of the analysis is very sensitive to the specific realisation of the noise processes and to the specific sky locations of the loudest systems contributing to the GWB with respect to the best pulsars in the array. The key role played by the stochastic realisation of the noise processes involved is also demonstrated by the BF distributions shown in Fig. 7; even when we inject a nearly ideal signal with *createGWB*, the distribution of Bayes factors returned by the analysis spans several orders of magnitudes.

5. Discussion and conclusions

In this paper, we carried out an extensive investigation of the performance of current PTA GW analyses on simulated PTA datasets injected with different types of GW signals. Our simulations included realistic levels of white noise, red noise, and DM variations, that were gauged to create mock data equivalent to the recently published EPTA DR2new. We injected in those data either a stochastic, stationary, Gaussian GWB with a standard $f^{-2/3}$
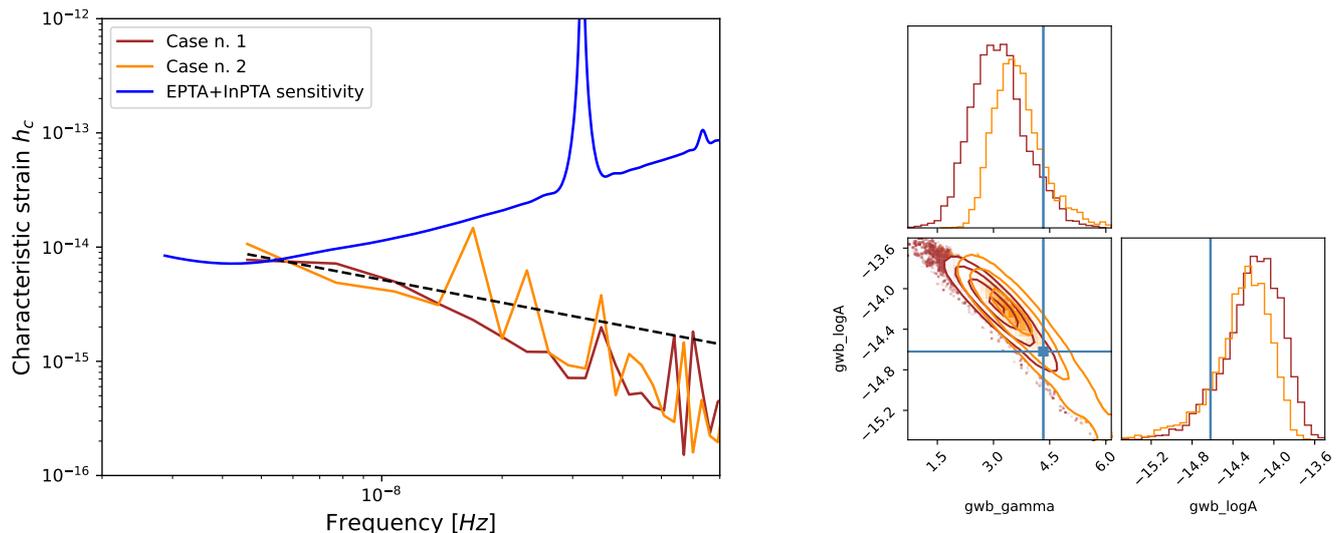
Fig. 9: Comparison between the two realisations denoted as *Case1* and *Case2*. In the left panel we compare the characteristic spectra with the the $f^{-2/3}$ trend, represented by the dashed black line. The sensitivity curve is the one presented in Collaboration et al. (2023). Right panel: corner plot of the GWB parameter posteriors for *Case1* (brown posterior) and *2* (orange posterior). The simulated GWB amplitude is 2.4×10^{-15}.

Case n.	$\text{BF}^{\text{HD}}_{\text{CURN}}$
1	1.179
2	12.358
3	70.133
4	62.503

Table 4: BF obtained from the reweighted nine-frequency bins inference runs (Hourihane et al. 2023) for the four notable cases analysed in Sec.4. The BF is computed for an HD-correlated signal over a common uncorrelated red noise.

power-law spectrum (the *createGWB_set* of simulations) or the incoherent superposition of sinusoidal signals from a cosmic population of SMBHBs (the *SMBHB_set* of simulations), paying particular attention to possible limitations and biases arising from the mismatch between the signal present in the data and the model used for the inference. The injected signals were calibrated on the results of Antoniadis et al. (2023b), where the amplitude of the signal was estimated to be $A_{\text{GWB}} \approx 2.4 \times 10^{-15}$ for a power-law spectrum with $\gamma_{\text{GWB}} = -13/3$.

We quantified the performance of the analysis model by constructing P–P plots for the amplitude of the recovered signal for each set of simulations. The *createGWB_set* demonstrates that, when the model matches with the injected signal, the outcome of the analysis is reliable, although some mild bias can arise due to the complex multi-dimensional nature of the parameter space that needs to be searched over. In fact, when fixing all the parameters but the GWB amplitude, the estimate of this parameter is unbiased; conversely, when performing a joint search on the GWB and noise parameters (including RN and DM) the signal amplitude is slightly biased towards lower amplitudes and steeper spectra. Such bias has also been seen in simulations of individual pulsar noise analysis (Antoniadis et al. 2023d), and although it is still between the one- and two-σ level (Fig 5, 6), it requires further investigation.

In the case of the more realistic *SMBHB_set*, where there is a mismatch between the injected signal and the simplified recovery model, the bias is much more prominent and the recovered spectra for the common noise tend to be steeper (higher γ_{GWB}) and with lower amplitude than the injected signal (Fig 5, 6). This is due to the nature of the SMBHB population, which features a large number of weak sources with a tail of sparse, loud systems, as discussed in Sec. 3.2. When one of these loud systems is present in the SMBHB population, the recovered GWB spectra can appear flatter (lower γ_{GWB}) and with a higher amplitude.

Using the reweighting method, we were also able to build a distribution of BF from the different realisations of both the *createGWB_set* and the *SMBHB_set* (see Fig. 7). We found no significant difference between the two distributions, although the *SMBHB_set* results on average in slightly higher BFs. We note that in both sets, the log*BF* distribution has a large scatter, with one-σ confidence interval spanning index covering the $[0, 2]$ interval. We also computed the induced angular correlation in the timing residuals of the different datasets and showed that they agree, within the predicted variance, with the HD correlation.

These results allow us to make several interesting considerations. First, as already shown in Cornish & Sampson (2016) the mismatch between the model and the data does not seem to affect our ability to recover the GW signal. This is probably because the detection significance is based on the intra-pulsar correlation properties of the signal (i.e. the HD overlap reduction function) which is a feature that emerges for any collection of GW signal, regardless on its specific properties (e.g. stationarity, Gaussianity, isotropy, spectral shape). Conversely, the reconstruction and interpretation of the observed signal can be severely biased by the use of a simplified GWB model. For example, while loud individual sources can cause a flattening of the spectrum which might erroneously misinterpreted as environmental effects or high eccentricity, the lack of them might result in a steep inferred spectrum which can be (again erroneously) claimed to be inconsistent with an astrophysical origin. Finally, the stochastic nature of the noise has a major impact on the outcome of the analysis. Even
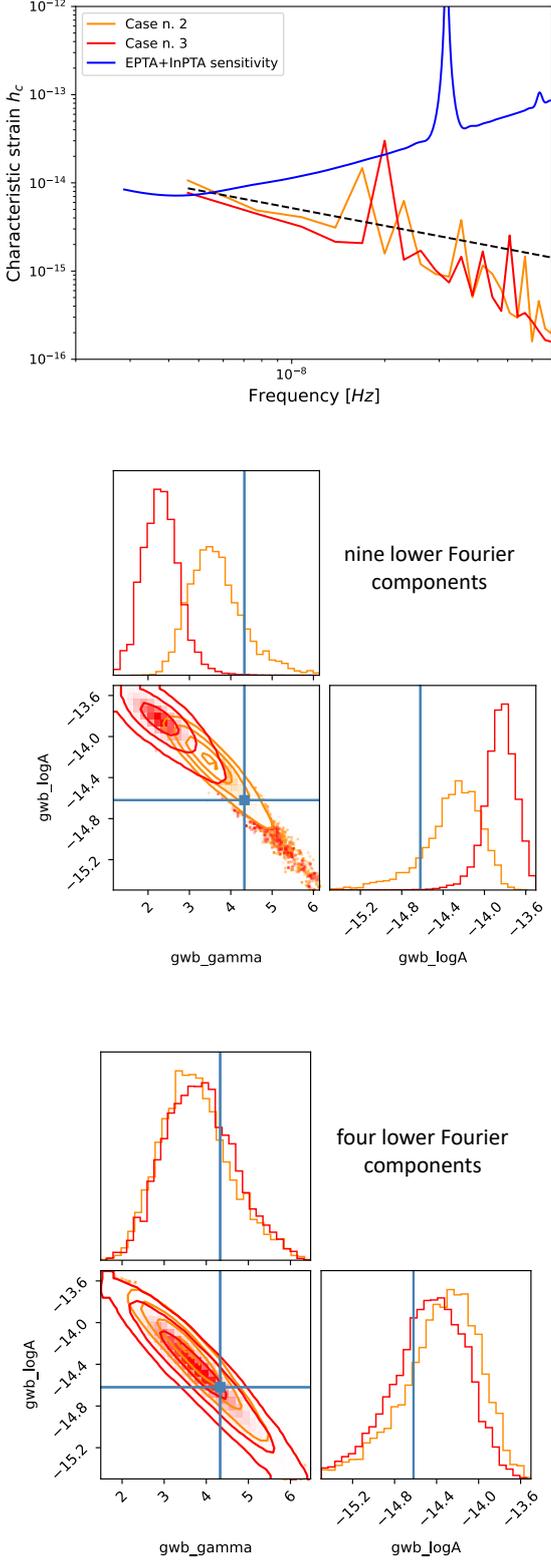
Fig. 10: Notable realisations *Case2* (orange) and *Case3* (red). Top panel: GWB spectra. The dashed black line represents the $f^{-2/3}$ trend. The sensitivity curve is the one presented in Collaboration et al. (2023). The central and bottom panels show corner plots of the GWB parameter posteriors inferred from the analysis using nine (central panel) and four (bottom panel) frequency bins for the GWB recovery.

for the *createGWB_set*, where we effectively always inject the same signal, the analysis can either return a detection supported by strong evidence or an inconclusive result, just depending on different realisation of the stochastic process describing the noise.

Within this diverse and complex phenomenology, we highlighted also some notable realisations that exemplifies some of the possible analysis outcomes. In particular, we highlighted that: (i) little astrophysical information can be drawn from an inference run using a simple power-law GWB model, by showing that very different GW signals can result in similar inferred GWB parameters, (ii) the presence of loud sources can introduce a bias in the recovery of the common process, (iii) some realisations of our simulations result in a recovered HD correlation and BF completely in line with what observed in EPTA DR2.

Finally, Bécsy et al. (2023) presented a similar set of simulations based on the NANOGrav 15yr dataset. They also use astrophysically motivated SMBHB populations to generate the GW signal and carry out a thorough analysis of a realistic dataset including unevenly sampled data and pulsar red noise. They perform Bayesian inference from the data, computing HD vs CURN Bayes factors and conclude that the simple GWB model implemented in the current analysis is able to recover a realistic GW signal, although they stress that loud sources might affect the inference. Compared to their work, our investigation adds several layers of sophistication. Our simulated data also include observations at two frequencies per epoch, allowing the inclusion of DM as a further source of noise, which allows as to test the analysis performance on a more complicated situation, closer to the real data. We cast our results in terms of P–P plots, and we compute combined posteriors of several realisations of the same dataset, which allowed us to identify some interesting systematic biases in the recovered signal. Finally, we carried out a systematic comparison on the analysis performed on realistic signal injections vs an ideal GWB generated by the *createGWB* function, which allowed us to identify potential difficulties due to the signal vs template mismatch.

Now that evidence of a GW signal is emerging independently from several PTA data, it is important to assessthe reliability of our analysis methods, in order to maximise the astrophysical potential of the PTA experiments. The present work, along with Bécsy et al. (2023), represents an important first step in this direction, which needs to be extended to include increasingly realistic situations. For example, in this article, we did not investigate complex signal spectra due to environmental effects or binary eccentricity. Likewise, we used the exact same model for noise injection and recovery. As shown in Antoniadis et al. (2023a), the excess or mis-modelled noise in the data can be absorbed within the common correlated signal in the analysis, leading to further biases and interpretation issues. Finally, including single sources along with a GWB in the recovery model might significantly improve the quality of the inference, especially when prominent peaks are present in the GW spectrum.

The main scripts used to analyse the data sets here described can be found at https://github.com/serevaltolina/PTAsim_fromSMBHBpop. The *SMBHB_set* is also available on zenodo under the following Digital Object Identifier: 10.5281/zenodo.10276364. In particular, for each universe realisation we uploaded: the SMBHBs specifications, the chain file obtained from the inference runs (while sampling for each pulsar intrinsic noise parameters and a common uncorrelated red noise), each pulsar .par and .tim files.
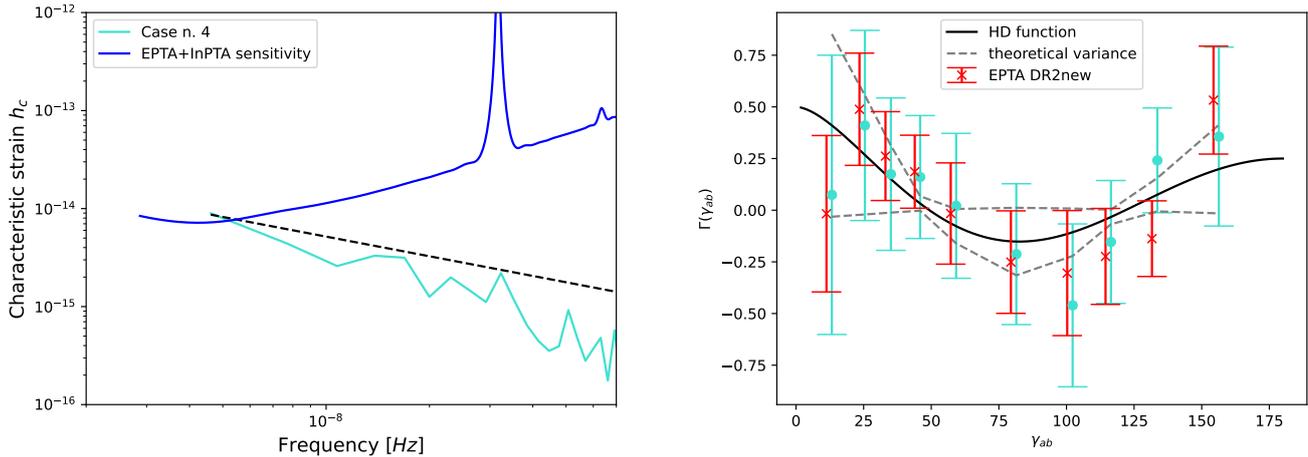
Fig. 11: *Case4* analysis. The left panel shows the characteristic strain spectrum compared to the $f^{-2/3}$ trend (dashed black line). The sensitivity curve is the one presented in Collaboration et al. (2023). Right panel: comparison between the recovered angular correlation in the residuals for this specific realisation (light blue points) and the results for the EPTA `DR2new` dataset (red). The expected variance of the optimal correlation estimator, for the EPTA 25 best pulsars array, is derived from prescriptions in Allen & Romano (2023).

Acknowledgements

References

Afzal, A., Agazie, G., Anumarlapudi, A., et al. 2023, The Astrophysical Journal Letters, 951, L11

Agazie, G., Anumarlapudi, A., Archibald, A. M., et al. 2023, The Astrophysical Journal Letters, 951, L8

Allen, B. 2023, Physical Review D, 107

Allen, B. & Romano, J. D. 2023, Physical Review D, 108

Anholm, M., Ballmer, S., Creighton, J. D. E., Price, L. R., & Siemens, X. 2009, Physical Review D, 79

Antoniadis, J., Arumugam, P., Arumugam, S., et al. 2023a, The second data release from the European Pulsar Timing Array: V. Implications for massive black holes, dark matter and the early Universe

Antoniadis, J., Arumugam, P., Arumugam, S., et al. 2023b, Astronomy & Astrophysics, 678, A50

Antoniadis, J., Arumugam, P., Arumugam, S., et al. 2023c, The second data release from the European Pulsar Timing Array III. Search for gravitational wave signals, data for EPTA DR2 paper 3.

Antoniadis, J., Arumugam, P., Arumugam, S., et al. 2023d, Astronomy & Astrophysics, 678, A49

Antoniadis, J., Babak, S., Bak Nielsen, A.-S., et al. 2023e, Astronomy & Astrophysics, 678, A48

Babak, S., Petiteau, A., Sesana, A., et al. 2016, MNRAS, 455, 1665

Babak, S. & Sesana, A. 2012, Phys. Rev. D, 85, 044034

Bartolo, N., Matarrese, S., Riotto, A., & Väihkönen, A. 2007, Phys. Rev. D, 76, 061302

Bécsy, B., Cornish, N. J., & Kelley, L. Z. 2022, The Astrophysical Journal, 941, 119

Boyle, L. A. & Buonanno, A. 2008, Phys. Rev. D, 78, 043531

Bécsy, B., Cornish, N. J., Meyers, P. M., et al. 2023, The Astrophysical Journal, 959, 9

Cai, Y.-F., He, X.-C., Ma, X.-H., Yan, S.-F., & Yuan, G.-W. 2023, Science Bulletin, 68, 2929

Chalumeau, A., Babak, S., Petiteau, A., et al. 2021, Monthly Notices of the Royal Astronomical Society, 509, 5538

Chamberlin, S. J., Creighton, J. D. E., Siemens, X., et al. 2015, Phys. Rev. D, 91, 044048

Collaboration, T. I. P. T. A., Agazie, G., Antoniadis, J., et al. 2023, Comparing recent PTA results on the nanohertz stochastic gravitational wave background

Cook, S. R., Gelman, A., & Rubin, D. B. 2006, Journal of Computational and Graphical Statistics, 15, 675

Cornish, N. J. & Sampson, L. 2016, Physical Review D, 93

Damour, T. & Vilenkin, A. 2000, Phys. Rev. Lett., 85, 3761

de Ravel, L., Le Fèvre, O., Tresse, L., et al. 2009, A&A, 498, 379

Demorest, P. B., Ferdman, R. D., Gonzalez, M. E., et al. 2012, The Astrophysical Journal, 762, 94

Edwards, R. T., Hobbs, G. B., & Manchester, R. N. 2006, Monthly Notices of the Royal Astronomical Society, 372, 1549

Ellis, J., Lewicki, M., Lin, C., & Vaskonen, V. 2023, Phys. Rev. D, 108, 103511

Ellis, J. & van Haasteren, R. 2017, jellis18/PTMCMCSampler: Official Release

Ellis, J. A., Siemens, X., & Creighton, J. D. E. 2012, ApJ, 756, 175

Ellis, J. A., Vallisneri, M., Taylor, S. R., & Baker, P. T. 2019, ENTERPRISE: Enhanced Numerical Toolbox Enabling a Robust PulsaR Inference SuitE, Astrophysics Source Code Library, record ascl:1912.015

Farris, B. D., Duffell, P., MacFadyen, A. I., & Haiman, Z. 2014, ApJ, 783, 134

Figueroa, D. G., Pieroni, M., Ricciardone, A., & Simakachorn, P. 2023, arXiv e-prints, arXiv:2307.02399

Foreman-Mackey, D. 2016, The Journal of Open Source Software, 1, 24

Franciolini, G., Iovino, A. J., Vaskonen, V., & Veermäe, H. 2023, Physical Review Letters, 131

Guo, S.-Y., Khlopov, M., Liu, X., et al. 2023, arXiv e-prints, arXiv:2306.17022

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357

Hellings, R. W. & Downs, G. S. 1983, ApJ, 265, L39

Hindmarsh, M., Huber, S. J., Rummukainen, K., & Weir, D. J. 2014, Phys. Rev. Lett., 112, 041301

Hiramatsu, T., Kawasaki, M., & Saikawa, K. 2014, J. Cosmology Astropart. Phys., 2014, 031

Hobbs, G. B., Edwards, R. T., & Manchester, R. N. 2006, Monthly Notices of the Royal Astronomical Society, 369, 655

Hourihane, S., Meyers, P., Johnson, A., Chatziioannou, K., & Vallisneri, M. 2023, Physical Review D, 107

Hunter, J. D. 2007, Computing in Science & Engineering, 9, 90

Jaffe, A. H. & Backer, D. C. 2003, ApJ, 583, 616

Keith, M. J., Coles, W., Shannon, R. M., et al. 2012, Monthly Notices of the Royal Astronomical Society, 429, 2161

Kelley, L. Z., Blecha, L., Hernquist, L., Sesana, A., & Taylor, S. R. 2018, MNRAS, 477, 964

Khmelnitsky, A. & Rubakov, V. 2014, Journal of Cosmology and Astroparticle Physics, 2014, 019

Kitajima, N., Lee, J., Murai, K., Takahashi, F., & Yin, W. 2023, arXiv e-prints, arXiv:2306.17146

Kitzbichler, M. G. & White, S. D. M. 2008, MNRAS, 391, 1489

Kormendy, J. & Ho, L. C. 2013, ARA&A, 51, 511

Kosowsky, A., Turner, M. S., & Watkins, R. 1992, Phys. Rev. D, 45, 4514

Madge, E., Morgante, E., Puchades-Ibáñez, C., et al. 2023, Journal of High Energy Physics, 2023

Matarrese, S., Pantano, O., & Saez, D. 1993, Phys. Rev. D, 47, 1311

Muzzin, A., Marchesini, D., Stefanon, M., et al. 2013, ApJ, 777, 18

Phinney, E. S. 2001, A Practical Theorem on Gravitational Wave Backgrounds

Rajagopal, M. & Romani, R. W. 1995, ApJ, 446, 543

Ravi, V., Wyithe, J. S. B., Hobbs, G., et al. 2012, ApJ, 761, 84

Ravi, V., Wyithe, J. S. B., Shannon, R. M., Hobbs, G., & Manchester, R. N. 2014, MNRAS, 442, 56

Reardon, D. J., Zic, A., Shannon, R. M., et al. 2023, The Astrophysical Journal Letters, 951, L6

Rosado, P. A., Sesana, A., & Gair, J. 2015, Monthly Notices of the Royal Astronomical Society, 451, 2417

Sampson, L., Cornish, N. J., & McWilliams, S. T. 2015, Phys. Rev. D, 91, 084055

Sesana, A. 2013, Classical and Quantum Gravity, 30, 224014

Sesana, A. 2013, Monthly Notices of the Royal Astronomical Society: Letters, 433, L1

Sesana, A., Vecchio, A., & Colacino, C. N. 2008, Monthly Notices of the Royal Astronomical Society, 390, 192

Sesana, A., Vecchio, A., & Volonteri, M. 2009, MNRAS, 394, 2255

Sorbo, L. 2011, J. Cosmology Astropart. Phys., 2011, 003

Speri, L., Porayko, N. K., Falxa, M., et al. 2022, Monthly Notices of the Royal Astronomical Society, 518, 1802

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. 2020, Validating Bayesian Inference Algorithms with Simulation-Based Calibration

Taylor, S. R., van Haasteren, R., & Sesana, A. 2020, Phys. Rev. D, 102, 084039

Tiburzi, C., Hobbs, G., Kerr, M., et al. 2016, MNRAS, 455, 4339

Tomita, K. 1967, Progress of Theoretical Physics, 37, 831

Vagnozzi, S. 2023, Journal of High Energy Astrophysics, 39, 81

Vallisneri, M. 2020, libstempo: Python wrapper for Tempo2, Astrophysics Source Code Library, record ascl:2002.017

Van Haasteren, R., Levin, Y., McDonald, P., & Lu, T. 2009, Monthly Notices of the Royal Astronomical Society, 395, 1005

Vigeland, S. J., Islo, K., Taylor, S. R., & Ellis, J. A. 2018, Physical Review D, 98

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261

Wilk, M. B. & Gnanadesikan, R. 1968, Biometrika, 55, 1

Wyithe, J. S. B. & Loeb, A. 2003, ApJ, 590, 691

Xu, H., Chen, S., Guo, Y., et al. 2023, Research in Astronomy and Astrophysics, 23, 075024
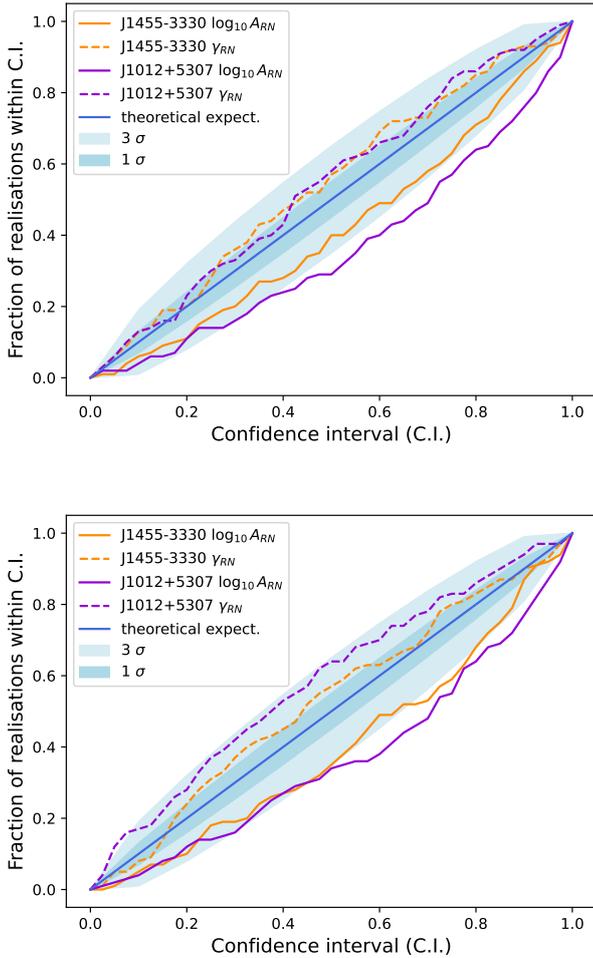
Fig. A.1: P–P plots for the recovery of pulsars J1455-3330 and J1012+5307 RN parameters from the MCMC inference runs performed on the *createGWB* (top panel) and *SMBHB_set* (bottom). Dashed lines are for the power-law slope, while the solid lines refer to the RN amplitude.

Antoniadis et al. (2023d) carried out a similar test on the RN recovery. In their Fig. 6, they show a P–P plot test on their noise parameter estimation from simulated data containing pulsar intrinsic noise only. Interestingly, they observe a bias in the *opposite direction* with respect to our results. Their conclusion is that, when the spectral slope of RN is greater than ~ 4, the recovery is, on average, biased towards flatter spectra with higher amplitudes. In our case, the spectral slope of the tested pulsars was much lower than that threshold (about 1.4 for both pulsars, see Table 1). This may be the at the root of the different results. We defer a deeper investigation of those mild biases to future studies.

Appendix A: Intrinsic noise parameters recovery

While we focussed on testing the viability and robustness of the recovery of the GWB signal, it is also interesting to briefly look at the recovered values for the pulsar's intrinsic RN when a common signal is included in the model. In particular, we choose the pulsars J1455-3330 and J1012+5307 (see Table 1 for the injected noise). We used the posteriors obtained from the MCMC runs over the *createGWB* and the *SMBHB_set* (see Sec. 3.2) to build the P–P plots for the inferred RN amplitude and slope for those two pulsars. The results are shown in Fig. A.1.

In this case there is no obvious distinction between the results from the *createGWB* and the *SMBHB_set*. In fact, the intrinsic pulsars RN is modelled in the same way in both of them and the different GWB injected does not seem to affect the RN inference. However, in both cases, the recovered distributions are slightly biased. Even if they all lie in a range comparable with the three-σ confidence interval, the distributions for $\log A_{RN}$ are constrained below the diagonal, while the ones for γ_{RN} are always above. This means that, for both pulsars, on average the recovered RN has spectra biased towards lower amplitudes and higher slopes; the same type of bias that we obtained for the GWB recovery.