The Astrophysical Journal, 968:16 (24pp), 2024 June 10 https://doi.org/10.3847/1538-4357/ad3e78

# Isolated Pulsar Population Synthesis with Simulation-based Inference

Vanessa Graber[1,2,3], Michele Ronchi[1,2], Celsa Pardo-Araujo[1,2], and Nanda Rea[1,2]

[1] Institute of Space Sciences (CSIC-ICE), Campus UAB, Carrer de Can Magrans s/n, 08193, Barcelona, Spain; v.graber@herts.ac.uk

[2] Institut d'Estudis Espacials de Catalunya (IEEC), Carrer Gran Capità 2–4, 08034 Barcelona, Spain

[3] Centre for Astrophysics Research, Department of Physics, Astronomy and Mathematics, University of Hertfordshire, College Lane, Hatfield AL10 9AB, UK

*Received 2023 December 21; revised 2024 April 11; accepted 2024 April 13; published 2024 June 5*

## Abstract

We combine pulsar population synthesis with simulation-based inference (SBI) to constrain the magnetorotational properties of isolated Galactic radio pulsars. We first develop a framework to model neutron star birth properties and their dynamical and magnetorotational evolution. We specifically sample initial magnetic field strengths, $B$, and spin periods, $P$, from lognormal distributions and capture the late-time magnetic field decay with a power law. Each lognormal is described by a mean, $\mu_{\log B}$, $\mu_{\log P}$, and standard deviation, $\sigma_{\log B}$, $\sigma_{\log P}$, while the power law is characterized by the index, $a_{\rm late}$. We subsequently model the stars' radio emission and observational biases to mimic detections with three radio surveys, and we produce a large database of synthetic $P$–$\dot{P}$ diagrams by varying our five magnetorotational input parameters. We then follow an SBI approach that focuses on neural posterior estimation and train deep neural networks to infer the parameters' posterior distributions. After successfully validating these individual neural density estimators on simulated data, we use an ensemble of networks to infer the posterior distributions for the observed pulsar population. We obtain $\mu_{\log B} = 13.10^{+0.08}_{-0.10}$, $\sigma_{\log B} = 0.45^{+0.05}_{-0.05}$ and $\mu_{\log P} = -1.00^{+0.26}_{-0.21}$, $\sigma_{\log P} = 0.38^{+0.33}_{-0.18}$ for the lognormal distributions and $a_{\rm late} = -1.80^{+0.65}_{-0.61}$ for the power law at the 95% credible interval. We contrast our results with previous studies and highlight uncertainties of the inferred $a_{\rm late}$ value. Our approach represents a crucial step toward robust statistical inference for complex population synthesis frameworks and forms the basis for future multiwavelength analyses of Galactic pulsars.

*Unified Astronomy Thesaurus concepts:* Neutron stars (1108); Radio pulsars (1353); Pulsars (1306)

## 1. Introduction

As one of the end points of stellar evolution of massive stars, neutron stars are influenced by many extremes of physics, including strong gravity, large densities, fast rotation, and extreme magnetic fields. Consequently, these compact objects have been connected with several of the most energetic transient phenomena in our Universe, such as fast radio bursts, superluminous supernovae, ultraluminous X-ray sources, long- and short-duration gamma-ray bursts, and gravitational-wave emission (e.g., Bachetti et al. 2014; Berger 2014; Metzger et al. 2014; Abbott et al. 2017; Margalit et al. 2018; Petroff et al. 2022). Accurately modeling these processes requires a detailed understanding of neutron star properties, which also set constraints on massive stellar evolution. Inferring the birth properties of neutron stars and the physics that govern their subsequent evolution is, thus, crucial for other fields of astrophysics.

Detecting and accurately characterizing individual objects within the entire neutron star population is, hence, critical. As a result, the number of known pulsars (those neutron stars that emit regular electromagnetic pulses) has steadily increased since the first detection in 1967 (Hewish et al. 1968), and we currently know of around 3500 of these objects (Manchester et al. 2005).[4] These are visible across the full electromagnetic spectrum, and their emission is predominantly driven by their enormous rotational energy reservoirs. Roughly 400 of these sources are confirmed to be in binaries, of which the majority were strongly influenced by accretion from their companions and spun up to short spin periods earlier in their lives. The remaining ∼3100 sources are primarily isolated neutron stars. Due to observational limitations and diverse emission properties, we cannot detect these with a single telescope, but instead have to focus on certain subpopulations. With around 1100 members, a subset of isolated radio pulsars constitutes the largest fraction of neutron stars detected in a single survey (Posselt et al. 2023). However, these numbers only cover a tiny portion of the overall neutron star population. We can provide a rough estimate of the neutron stars in the Milky Way by multiplying their birth rate (a core-collapse supernova rate of ∼2 per century; see Keane & Kramer 2008; Rozwadowska et al. 2021) by the age of the Milky Way (∼13 billion years; see, e.g., Conroy et al. 2022; Xiang & Rix 2022) to arrive at a total of 260 million Galactic neutron stars.

To bridge the gap between expected and observed neutron stars, we take advantage of population synthesis. This approach relies on producing a large catalog of synthetic pulsar populations that are passed through a set of filters to mimic observational constraints. The resulting populations are then contrasted with the true observed sample to find those parameter regions that best explain the data. Although different versions of this methodology have been applied to pulsar data for several decades (e.g., Narayan & Ostriker 1990; Lorimer 2004; Faucher-Giguère & Kaspi 2006; Gonthier et al. 2007; Bates et al. 2014; Gullón et al. 2014, 2015; Cieślar et al. 2020), the complexity of models that capture the properties of observed Galactic neutron stars significantly complicates the comparison between the simulated populations and the observed one. This is especially true if we are interested in quantifying uncertainties for our neutron star parameters, because Bayesian Markov Chain Monte Carlo (MCMC) or nested

[4] https://www.atnf.csiro.au/research/pulsar/psrcat/; v2.1.0

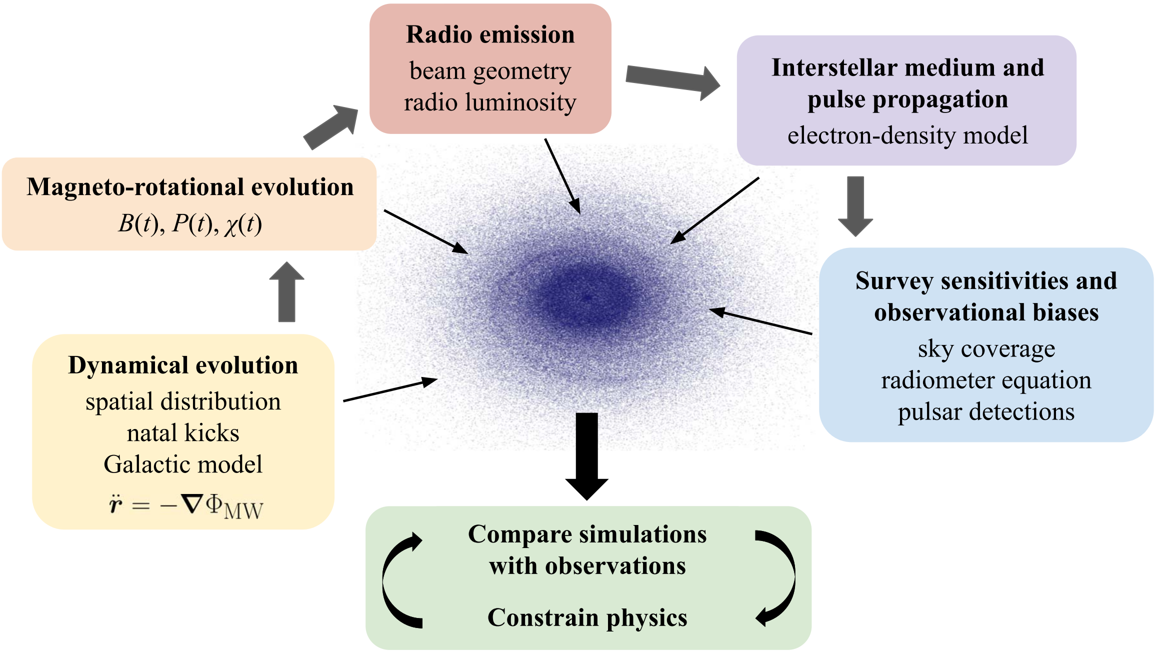
Radio emission
beam geometry
radio luminosity
Interstellar medium and pulse propagation
electron-density model
Magneto-rotational evolution
$B(t)$, $P(t)$, $\chi(t)$
Survey sensitivities and observational biases
sky coverage
radiometer equation
pulsar detections
Dynamical evolution
spatial distribution
natal kicks
Galactic model
$\ddot{\boldsymbol{r}} = -\boldsymbol{\nabla}\Phi_{\rm MW}$
Compare simulations with observations
Constrain physics

**Figure 1.** The key ingredients for pulsar population synthesis. Starting from the bottom left, this approach relies on modeling the neutron stars' dynamical evolution, as well as their magnetorotational properties. For a given beaming geometry and luminosity model, we then determine the pulsars' radio emission and its propagation across the Galaxy toward Earth. For the neutron stars pointing toward us, we subsequently invoke survey limitations and sensitivity thresholds to determine those objects that are detectable. The resulting synthetic populations are compared to the observed ones to constrain input physics.

sampling methods (the standard tools for this kind of question; see, e.g., Feroz et al. 2009; Foreman-Mackey et al. 2013; Sharma 2017; Ashton et al. 2019; Speagle 2020) become infeasible for pulsar population synthesis unless significant simplifications for simulation models and the likelihood function are made (Cieślar et al. 2020). The main reason for this is that we can no longer write down an explicit likelihood for realistic neutron star simulation frameworks. In this paper, we thus focus on simulation-based inference (SBI; also known as likelihood-free inference; for a recent review see Cranmer et al. 2020) in the context of pulsar population synthesis for the first time.

In the past few years, SBI has successfully challenged traditional approaches such as approximate Bayesian computation (e.g., Rubin 1984; Beaumont et al. 2002; Dean et al. 2011; Frazier et al. 2017) in those areas of science that rely on complex simulators, which lead to intractable likelihoods. The existence of such a simulator, essentially acting as a forward model, is the only requirement for SBI. As such, the approach is ideal for astrophysics and has been recently applied to parameter estimation in, e.g., cosmology (Alsing et al. 2019; Hahn et al. 2023; Lemos et al. 2023; Lin et al. 2023), high-energy astrophysics (Huppenkothen & Bachetti 2022; Mishra-Sharma & Cranmer 2022), gravitational-wave astronomy (Dax et al. 2021; Cheung et al. 2022; Bhardwaj et al. 2023), and exoplanet research (Vasist et al. 2023). SBI is particularly powerful in combination with neural networks, whose benefits for pulsar population synthesis studies were outlined in Ronchi et al. (2021) by inferring point estimates for the dynamical properties of radio pulsars in the Milky Way.

In this study, we take a Bayesian perspective to infer posteriors of neutron star parameters using SBI. For this purpose, we model the Galactic neutron star dynamics, the magnetorotational evolution, and the radio emission properties. We then run snapshots of the total pulsar population at the current time through a set of filters to mimic observational limitations. The resulting simulation output are synthetic $P$–$\dot{P}$ diagrams (where $P$ and $\dot{P}$ denote the pulsar spin period and its time derivative, respectively) of the observed pulsar population. We then construct an SBI pipeline, which we train, validate, and test on a large database of these synthetic $P$–$\dot{P}$ diagrams to infer posterior distributions of our input parameters. We specifically focus on five parameters related to the initial period distribution of pulsars and their magnetic field properties that crucially affect the positions of stars in the $P$–$\dot{P}$ plane. We then apply our optimized deep-learning framework, for the first time, to the radio pulsars detected in the Parkes Multibeam Pulsar Survey (PMPS; Manchester et al. 2001; Lorimer et al. 2006), the Swinburne Intermediate-latitude Pulsar Survey (SMPS; Edwards et al. 2001; Jacoby et al. 2009), and the low- and mid-latitude High Time Resolution Universe (HTRU) survey (Keith et al. 2010) (all recorded with Murriyang, the Parkes radio telescope).

The paper is structured as follows: Section 2 summarizes our population synthesis framework. We then provide a general overview of SBI and our choice of setup in Sections 3.1 and 3.2, respectively, whereas Section 3.3 summarizes the machine-learning experiments conducted for this study. We next address network training and inference results plus corresponding validation approaches in Section 4, specifically benchmarking our pipeline on test simulations before applying it to the observed pulsar population. Finally, we provide a detailed discussion of our approach and results, as well as an outlook into the future, in Section 5.

## 2. Pulsar Population Synthesis

### 2.1. Overview

The key ingredients for our pulsar population synthesis model are summarized in Figure 1. We first require a prescription for the star's dynamical properties to populate our synthetic Galaxy with neutron stars. To this end, we model their birth positions and velocities plus their subsequent dynamical evolution in the Milky Way. We further capture the stars' initial magnetic and rotational characteristics in addition to their evolution. For both these aspects, our framework broadly follows earlier works (see, e.g., Faucher-Giguère & Kaspi 2006; Gullón et al. 2014; Cieślar et al. 2020; Ronchi et al. 2021), and our simulator employs a Monte Carlo approach to sample relevant parameters at birth from corresponding probability density functions. We note that we save computation time by not evolving the dynamical properties for each single simulation. As the dynamical and magnetorotational properties are independent, we instead simulate a single dynamical database for a large number of current pulsar positions and velocities, and we subsequently sample from these distributions before determining the magnetorotational evolution. Next, we characterize the stars' radio emission by implementing a realistic beaming geometry. We then simulate detections by propagating the corresponding radio pulses across the Galaxy for a specific electron density model. The resulting emission for those pulsars pointing toward Earth is then contrasted to observational biases and sensitivity thresholds for a given radio survey to determine which synthetic pulsars would be detected. The resulting mock populations are then compared to the observed populations to constrain relevant model parameters. We explore SBI for this purpose as outlined in detail in Section 3.

### 2.2. Dynamical Evolution

To create our dynamical database from which we sample neutron star positions and velocities, we simulate $10^7$ neutron stars from birth to today. For each object, we randomly assign an age sampled from a uniform distribution up to a maximum age of $10^8$ yr, which ensures that our synthetic Milky Way is populated with a sufficient number of neutron stars within a reasonable computation time. As sources older than $10^8$ yr are no longer detectable as radio pulsars (see below), this approach provides a realistic description of the current positions and velocities of these objects.

We then define a cylindrical reference frame, $(r, \phi, z)$, whose origin is located at the Galactic center. Here $r$, $\phi$, and $z$ denote the distance from the origin in kpc, the azimuthal angle in radians, and the distance from the Galactic plane in kpc, respectively. In particular, we position our Sun at $r = 8.3$ kpc, $\phi = \pi/2$, and $z = 0.02$ kpc (see Pichardo et al. 2012, and references therein).

To determine the birth locations of individual neutron stars, we first focus on the distributions of their massive progenitors in the $(r, \phi)$-plane and along $z$ separately. Considering the distribution of free electrons as a tracer of star formation in the Milky Way that correlates with the massive OB stars that evolve into neutron stars, we sample the initial positions in $r$, $\phi$ according to the Galactic electron density distribution of Yao et al. (2017). This will also allow consistency when relating pulsar distances with their dispersion measures in Section 2.5. In addition, as the Galactic matter distribution is not static, we assume that the Milky Way rotates rigidly in a clockwise direction with an angular velocity $\Omega = 2\pi/T$, where $T \approx 250$ Myr (Vallée 2017; Skowron et al. 2019). For a given stellar age, we can thus retrace the angular coordinate, $\phi$, at birth.

Moreover, we assume that pulsar birth positions along the $z$-direction follow an exponential disk model (Wainscoat et al. 1992) and sample from a probability density function of the form

$$\mathcal{P}(z) = \frac{1}{h_c} \exp\left(-\frac{|z|}{h_c}\right). \tag{1}$$

We follow the pulsar population studies of Gullón et al. (2014) and Ronchi et al. (2021) and set the characteristic scale height, $h_c$, to a fiducial value of 0.18 kpc. Note that this is consistent with the distribution of young, massive stars in our Galaxy (Li et al. 2019). We then randomly assign each star's $z$-coordinate a positive or negative sign to distribute our population above and below the Galactic plane.

Next, we focus on the pulsars' birth velocities, which are a combination of the kick velocity, $\boldsymbol{v}_k$, imparted during the supernova owing to explosion asymmetries (see Coleman & Burrows 2022; Janka et al. 2022, and references therein) and the velocity, $\boldsymbol{v}_{\rm pr}$, inherited from the progenitors' orbital Galactic motion. Specifically, we sample the magnitude of the kick velocities, $v_k \equiv |\boldsymbol{v}_k|$, from a Maxwell distribution,

$$\mathcal{P}(v_k) = \sqrt{\frac{2}{\pi}} \frac{v_k^2}{\sigma_k^3} \exp\left(-\frac{v_k^2}{\sigma_k^2}\right), \tag{2}$$

and then assign a random direction to determine the kick along the $r$-, $\phi$-, and $z$-directions. For the dispersion parameter, $\sigma_k$, we take a fiducial value of $\sigma_k \approx 260\,{\rm km\,s^{-1}}$ (Hobbs et al. 2005), which is broadly consistent with observed proper motions of radio pulsars (Hobbs et al. 2005; Faucher-Giguère & Kaspi 2006). See, however, Verbunt et al. (2017) and Igoshev (2020), who find that a double Maxwellian characterizes the data better.

The second velocity component due to the progenitors' motion depends on the Galactic gravitational potential, $\Phi_{\rm MW}$, and points along the azimuthal direction:

$$\boldsymbol{v}_{\rm pr} = \sqrt{r \frac{\partial \Phi_{\rm MW}(r, z)}{\partial r}}\, \hat{\boldsymbol{\phi}}, \tag{3}$$

where $\hat{\boldsymbol{\phi}}$ is a unit vector in the $\phi$-direction. For this study, we consider a Galactic potential that is given as the sum of four components, i.e., the nucleus, $\Phi_n$, the bulge, $\Phi_b$, the disk, $\Phi_d$, and the halo, $\Phi_h$ (Marchetti et al. 2019). The nucleus and bulge contributions are described by a spherical Hernquist potential (Hernquist 1990):

$$\Phi_{n,b} = -\frac{GM_{n,b}}{R_{n,b} + R}, \tag{4}$$

where $R = \sqrt{r^2 + z^2}$ is the spherical radial coordinate and $G$ is the gravitational constant. The disk has a cylindrical Miyamoto–Nagai potential of the form (Miyamoto & Nagai 1975)

$$\Phi_d = -\frac{GM_d}{\sqrt{(a_d + \sqrt{z^2 + b_d^2})^2 + r^2}}, \tag{5}$$

where $a_d$ and $b_d$ represent the scale length and scale height of the disk, respectively. Finally, the halo is characterized by a spherical Navarro–Frenk–White potential (Navarro et al.

1996):

$$\Phi_h = -\frac{GM_h}{R}\ln\left(1+\frac{R}{R_h}\right). \tag{6}$$

The free parameters, $M_{n,b,d,h}$, $R_{n,b,h}$, $a_d$, and $b_d$, can be obtained through fits of the Milky Way's mass profile and are given in Table 2 of Ronchi et al. (2021) (see also Bovy 2015 and Table 1 of Marchetti et al. 2019).

After determining the initial positions and velocities for each of our $10^7$ neutron stars, we perform the dynamical evolution by solving the Newtonian equation of motion in cylindrical coordinates, $\ddot{\boldsymbol{r}} = -\boldsymbol{\nabla}\Phi_{\rm MW}$, according to the stars' respective ages. This way, we obtain a database of current pulsar positions and velocities in the Milky Way.

### 2.3. Magnetorotational Evolution

The primary diagnostic for the pulsar population is the $P$–$\dot{P}$ diagram. For our study, we focus on rotation-powered radio pulsars, which are the easiest to detect and constitute the largest class of neutron stars. Corresponding period and period derivative measurements for this population are enabled via radio timing. To first order, radio pulsars can be approximated as rotating magnetic dipoles, implying that their spin-down is driven by electromagnetic dipole radiation. The locations of individual neutron stars and the shape of the population's distribution in the $P$–$\dot{P}$ plane are, hence, determined by their dipolar magnetic fields and rotation periods at birth and the subsequent magnetorotational evolution. The latter couples the evolution of the pulsar period, $P$, the dipolar magnetic field strength, $B$, at the pole, and the inclination angle, $\chi$, between the magnetic and the rotation axis.

To capture these physics, we first sample the misalignment angle at birth, $\chi_0$, randomly in the range $[0, \pi/2]$ according to the probability density (Gullón et al. 2014)

$$\mathcal{P}(\chi_0) = \sin\chi_0. \tag{7}$$

We then sample the logarithm of the initial magnetic field, $B_0$ (measured in G), and the initial period, $P_0$ (measured in s), for each pulsar from normal distributions of the form (Popov et al. 2010; Gullón et al. 2014; Igoshev 2020; Igoshev et al. 2022; Xu et al. 2023)

$$\mathcal{P}(\log B_0) = \frac{1}{\sqrt{2\pi}\,\sigma_{\log B}}\exp\left(-\frac{\log B_0 - \mu_{\log B}}{2\sigma^2_{\log B}}\right), \tag{8}$$

$$\mathcal{P}(\log P_0) = \frac{1}{\sqrt{2\pi}\,\sigma_{\log P}}\exp\left(-\frac{\log P_0 - \mu_{\log P}}{2\sigma^2_{\log P}}\right). \tag{9}$$

The means, $\mu_{\log B}$, $\mu_{\log P}$, and the standard deviations, $\sigma_{\log B}$, $\sigma_{\log P}$, are free parameters of our model and four of those parameters, whose posteriors we set out to infer with our SBI approach in Section 3. We will specifically explore the ranges $\mu_{\log B} \in [12, 14]$, $\mu_{\log P} \in [-1.5, -0.3]$, $\sigma_{\log B} \in [0.1, 1.0]$, and $\sigma_{\log P} \in [0.1, 1.0]$ to encompass results of earlier analyses (e.g., Gullón et al. 2014).

Assuming that pulsars spin down owing to dipolar emission, we follow Spitkovsky (2006) and Philippov et al. (2014) and solve the following coupled differential equations:

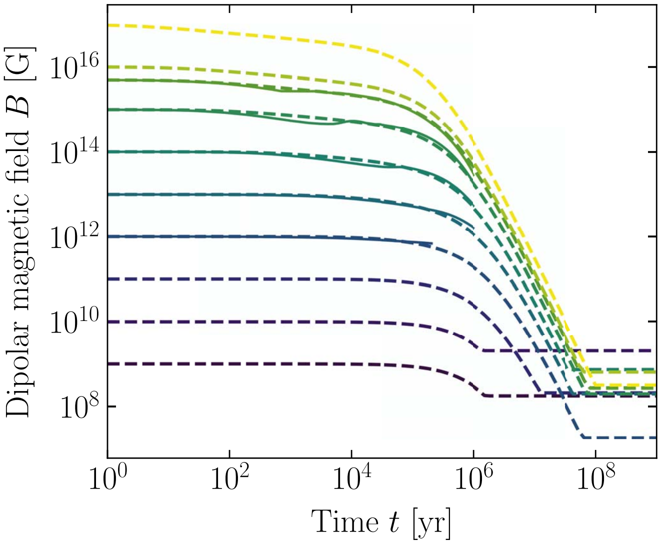

**Figure 2.** Illustration of the $B$-field parameterization used for this study. The five solid curves represent realistic two-dimensional simulations of magnetothermal evolution in the neutron star crust (Viganò et al. 2021). We fit these together with the late-time power-law evolution of the magnetic field with several broken power laws. The dashed curves shown here are determined for $a_{\rm late} = -3.0$. The colors represent the initial magnetic field strength, $B_0$. To avoid the field decaying to unrealistically small numbers at very late times, we sample the final fields from a Gaussian distribution. The procedure, which allows us to easily extract the dipolar field strength, $B$, at different times, $t$, to study the magnetorotational evolution of our synthetic pulsars, is described in detail in Appendix A.

$$\dot{P} = \frac{\pi^2}{c^3}\frac{B^2R^6_{\rm NS}}{I_{\rm NS}P}(\kappa_0 + \kappa_1\sin^2\chi), \tag{10}$$

$$\dot{\chi} = -\frac{\pi^2}{c^3}\frac{B^2R^6_{\rm NS}}{I_{\rm NS}P^2}\kappa_2\sin\chi\cos\chi, \tag{11}$$

where $c$ is the speed of light, $R_{\rm NS} \approx 11$ km is the neutron star radius, and $I_{\rm NS} \simeq 2M_{\rm NS}R^2_{\rm NS}/5 \approx 1.36 \times 10^{45}$ g cm$^2$ is the stellar moment of inertia (for a fiducial mass $M_{\rm NS} \approx 1.4 M_\odot$). For realistic pulsars surrounded by plasma-filled magnetospheres, we choose $\kappa_0 \simeq \kappa_1 \simeq \kappa_2 \simeq 1$, and we note that Equation (11) implies that $\chi$ decreases with time, i.e., our pulsars move toward alignment.

The final ingredient is a suitable prescription for the evolution of the dipolar magnetic field strength. While the $B$-field decay in the neutron star crust is typically assumed to be driven by the combined action of the Hall effect and ohmic dissipation (e.g., Aguilera et al. 2008), changes in the magnetic field are strongly coupled to the thermal properties of the neutron star interior (e.g., Pons & Viganò 2019). This is particularly important for strongly magnetized neutron stars with fields above $\sim 10^{13}$ G and, hence, relevant for a significant fraction of our simulated pulsar population. In the past decade, several theoretical and numerical efforts have begun to unveil the complex processes of magnetothermal evolution (e.g., Viganò et al. 2013, 2021; De Grandis et al. 2021; Igoshev et al. 2021; Dehman et al. 2023). As corresponding simulations are highly time-consuming, we instead develop a new approach, outlined in detail in Appendix A and summarized in Figure 2, that parameterizes a range of magnetothermal simulations for different magnetic field strengths (Viganò et al. 2021). This prescription allows us to extract magnetic fields up to pulsar ages of around $10^6$ yr. Above this value, current numerical simulations become unreliable because they rely on implementations of complex microphysics that are unsuitable for cold, old stars. Moreover, they do not capture the highly uncertain physics of neutron star cores, which become relevant at large ages. We instead incorporate the cores' field evolution at late times by means of a power law of the form

$$B(t) \propto \left(1 + \frac{t}{\tau_{\rm late}}\right)^{a_{\rm late}}, \tag{12}$$

where $\tau_{\rm late} \approx 2 \times 10^6$ yr, $t$ is the time, and the power-law index, $a_{\rm late}$, is the fifth free parameter of our model. We note that although the details of core field evolution are not known, Equation (12) is physically motivated because several known mechanisms exhibit similar power-law behavior (see Appendix A). Hence, we will explore the parameter range $a_{\rm late} \in [-3.0, -0.5]$. Finally, to prevent the dipolar magnetic field from decaying to arbitrarily small values (in disagreement with observations of old, recycled millisecond pulsars; see, e.g., Lorimer 2008), we assume that the field eventually reaches a constant value. Therefore, we sample the logarithm of the field, $B_{\rm final}$, from a normal distribution with a mean $\mu_{\log B,{\rm final}} = 8.5$ and a standard deviation $\sigma_{\log B,{\rm final}} = 0.5$, in line with observations of old pulsars.

Following this prescription allows us to determine the spin periods, dipolar field strengths, and misalignment angles for our simulated pulsars at the current time.

### 2.4. Emission Characteristics

We next implement a prescription for the radio emission geometry to determine those pulsars whose beams sweep over Earth and are, in principle, detectable. In the canonical model of radio pulsars, their emission is produced close to the stellar surface in the cone-shaped, open field-line region (Lorimer & Kramer 2012; Johnston et al. 2020). Assuming that this entire region is involved in the emission, geometric considerations allow us to estimate the half opening angle of the emission beam, $\rho_b$ (in rad), via (Gangadhara & Gupta 2001)

$$\rho_b \simeq \sqrt{\frac{9\pi r_{\rm em}}{2cP}}, \tag{13}$$

where $r_{\rm em}$ is the emission height. The latter is thought to be period independent, and we set it to 300 km following Johnston et al. (2020; see also references therein). Note that several studies of pulsars with stable emission profiles have recovered this $\rho_b \propto P^{-1/2}$ behavior (e.g., Kramer et al. 1994; Maciesiak & Gil 2011; Skrzypczak et al. 2018). Knowledge of $\rho_b$, then, allows us to obtain the solid angle, $\Omega_b$, covered by a pulsar's two radio beams. More specifically,

$$\Omega_b = 4\pi(1 - \cos\rho_b). \tag{14}$$

As we do not expect biases in how we observe this conal emission for any given pulsar, we draw a random line-of-sight angle, $\alpha$, with respect to the rotation axis in the range $[0, \pi/2]$ using the probability density $\sin\alpha$. Combined with the half opening angle, $\rho_b$, and the evolved inclination angle, $\chi$, we can then determine those pulsars whose radio beams are visible from Earth. We note that as a result of this purely geometric argument, between ∼60% and 95% of our generated pulsars (depending on the specific choice of magnetorotational parameters) are typically not detectable.

We proceed with determining the emission characteristics of those neutron stars that point toward Earth. In particular, we follow Maciesiak et al. (2011) and express the intrinsic pulse width (measured in s) of our simulated pulsars as follows:

$$w_{\rm int} = \frac{2}{\pi}\arcsin\sqrt{\frac{\sin^2\left(\frac{\rho_b}{2}\right) - \sin^2\left(\frac{\alpha-\chi}{2}\right)}{\sin(\alpha)\sin(\chi)}}\; P. \tag{15}$$

Finally, as the radio emission is ultimately driven by the stars' rotational energy reservoir, we assume that the intrinsic radio luminosity, $L_{\rm int}$ (in erg s$^{-1}$), for each star depends on the spin-down power, $|\dot{E}_{\rm rot}| = 4\pi^2 I_{\rm NS}\dot{P}/P^3$. In particular, we consider

$$L_{\rm int} = L_0\sqrt{\frac{\dot{P}}{P^3}}, \tag{16}$$

where $L_0$ is a normalization factor whose logarithm we sample from a normal distribution with mean $\mu_{\log L} = 35.5$ and standard deviation $\sigma_{\log L} = 0.8$ (see also Faucher-Giguère & Kaspi 2006; Gullón et al. 2014) to eventually recover observed luminosities.

### 2.5. Simulating Detections

Armed with the knowledge of intrinsic pulsar properties, we now turn to the possibility of detecting those objects whose emission beams cross our line of sight. First, the bolometric radio flux, $S$, that reaches us from any given simulated pulsar is equal to

$$S = \frac{L_{\rm int}}{\Omega_b d^2}, \tag{17}$$

where $d$ is the distance known from the dynamical evolution outlined in Section 2.2. To determine the corresponding radio flux density, $S_f$ (measured in Jy), at a specific observing frequency, $f$, we follow Lorimer & Kramer (2012) and assume that the radio emission spectrum follows a power law in $f$. In particular, we set the spectral index to $-1.6$ (Jankowski et al. 2018). We can, hence, approximate the total fluence of a pulse with width $w_{\rm int}$ as $S_f w_{\rm int}$. Assuming that this fluence stays constant as the radio signal propagates from the pulsar toward us, we estimate the flux density, $S_{f,{\rm obs}}$, that reaches Earth as

$$S_{f,{\rm obs}} \simeq S_f \frac{w_{\rm int}}{w_{\rm obs}}, \tag{18}$$

where $w_{\rm obs}$ is the observed pulse width.

Specifically, as a radio pulse propagates, it experiences dispersion and scattering caused by interactions with the free electrons and density fluctuations in the interstellar medium (ISM), respectively. Both mechanisms result in a broader pulse when compared with the intrinsic width, $w_{\rm int}$. Further broadening is caused by instrumental effects, which are dominated by the sampling time, $\tau_{\rm samp}$, of the hardware used to record radio observations. Accounting for these processes, we can write the observed pulse width as (Cordes & McLaughlin 2003)

$$w_{\rm obs} \simeq \sqrt{w_{\rm int}^2 + \tau_{\rm samp}^2 + \tau_{\rm DM}^2 + \tau_{\rm scat}^2}. \tag{19}$$

We follow Bates et al. (2014) to determine $\tau_{\rm DM}$, encoding the pulse smearing due to dispersion for a single frequency channel

**Table 1**
Survey Parameters for PMPS, SMPS, and the Low- and Mid-latitude HTRU Survey Taken from Manchester et al. (2001), Lorimer et al. (2006), Edwards et al. (2001), Jacoby et al. (2009), and Keith et al. (2010), Respectively

| Survey | PMPS | SMPS | HTRU mid | HTRU low |
|---|---|---|---|---|
| Sky region | $-100° < l < 50°$ | $-100° < l < 50°$ | $-120° < l < 30°$ | $-80° < l < 30°$ |
| | $\lvert b\rvert < 5°$ | $5° < \lvert b\rvert < 30°$ | $\lvert b\rvert < 15°$ | $\lvert b\rvert < 3\overset{\circ}{.}5$ |
| $f$ (GHz) | 1.374 | 1.374 | 1.352 | 1.352 |
| $\Delta f_{\rm ch}$ (kHz) | 3000 | 3000 | 390.625 | 390.625 |
| $\tau_{\rm samp}$ ($\mu$s) | 250 | 125 | 64 | 64 |
| $G$ (K  Jy$^{-1}$) | 0.735 | 0.735 | 0.735 | 0.735 |
| $n_{\rm pol}$ | 2 | 2 | 2 | 2 |
| $\Delta f_{\rm bw}$ (MHz) | 288 | 288 | 340 | 340 |
| $t_{\rm obs}$ (s) | 2100 | 265 | 540 | 4300 |
| $\beta$ | 1.5 | 1.5 | 1.5 | 1.5 |
| $T_{\rm sys}$ (K) | 21 | 21 | 23 | 23 |
| S/N threshold | 9 | 9 | 9 | 9 |

**Note.** We provide the survey region where completeness is above 90% in Galactic longitude ($l$) and latitude ($b$), the central observing frequency ($f$), the channel width ($\Delta f_{\rm ch}$), the sampling time ($\tau_{\rm samp}$), the telescope gain ($G$), the number of observed polarizations ($n_{\rm pol}$), the observing bandwidth ($\Delta f_{\rm bw}$), the integration time ($t_{\rm obs}$), the degradation factor ($\beta$), the system temperature ($T_{\rm sys}$), and the S/N threshold for each of the surveys. Corresponding units are given in parentheses in the first column.

of the telescope's receiver. Specifically,

$$\tau_{\rm DM} = \frac{e^2}{\pi m_e c}\frac{\Delta f_{\rm ch}}{f^3}\,{\rm DM}, \tag{20}$$

where $e$ is the electronic charge, $m_e$ is the corresponding mass, $\Delta f_{\rm ch}$ us the hardware-specific width of a frequency channel at observing frequency $f$, and DM is the dispersion measure. We further use the empirical fit relationship from Krishnakumar et al. (2015) for $\tau_{\rm scat}$, the pulse smearing due to scattering of radio waves by an inhomogeneous and turbulent ISM:

$$\tau_{\rm scat} = 3.6 \times 10^{-9}{\rm DM}^{2.2}(1 + 1.94 \times 10^{-3}{\rm DM}^2), \tag{21}$$

where $\tau_{\rm scat}$ is measured in s. We moreover account for a significant scatter in the underlying data (see Figure 3 in Krishnakumar et al. 2015) by drawing $\log \tau_{\rm scat}$ values from a Gaussian distribution around the fit in Equation (21) with a standard deviation of 0.5. We also incorporate the fact that Krishnakumar et al. (2015) analyzed observations at 327 MHz by rescaling to a given observing frequency $f$, assuming a Kolmogorov spectrum, i.e., $\tau_{\rm scat} \propto f^{-4.4}$ (see Lorimer & Kramer 2012, for details). As $\tau_{\rm DM}$ and $\tau_{\rm scat}$ both depend on the pulsars' respective dispersion measure, we again employ the Galactic electron density distribution of Yao et al. (2017) to convert our simulated neutron star positions from Section 2.2 into DM values.

At this stage, we require information for the radio surveys we want to emulate. We specifically focus on three surveys recorded with Murriyang, the Parkes radio telescope: PMPS (Manchester et al. 2001; Lorimer et al. 2006), SMPS (Edwards et al. 2001; Jacoby et al. 2009), and the low- and mid-latitude HTRU survey (Keith et al. 2010). All relevant survey parameters (including the sampling time, $\tau_{\rm samp}$, the observing frequency, $f$, and the channel width, $\Delta f_{\rm ch}$, needed to calculate $w_{\rm obs}$) are summarized in Table 1.

To assess whether those simulated stars that cross our line of sight are detectable with a given survey, we first determine whether they are located in the surveys' fields of view. While PMPS and HTRU have a similar sky coverage, we highlight that SMPS detected pulsars at higher Galactic latitude (see left panel of Figure 3). This survey is thus sensitive to older neutron stars that have had sufficient time to move away from their birth positions closer to the Galactic plane, providing complementary information on the pulsar population. For those objects that fall within our survey coverage, we subsequently establish whether they are sufficiently bright to be detected. To do so, we calculate the pulsars' signal-to-noise ratio (S/N) using the radiometer equation (Lorimer & Kramer 2012):

$${\rm S/N} = \frac{S_{\rm mean}\, G \sqrt{n_{\rm pol}\Delta f_{\rm bw}\, t_{\rm obs}}}{\beta[T_{\rm sys} + T_{\rm sky}(l,\, b)]}\sqrt{\frac{P - w_{\rm obs}}{w_{\rm obs}}}. \tag{22}$$

Here $S_{\rm mean} \simeq S_{f,{\rm obs}} w_{\rm obs}/P$ denotes the mean flux density averaged over a single rotation period $P$, $G$ is the receiver gain (see Lorimer et al. 1993; Bates et al. 2014, for details), $n_{\rm pol}$ is the number of detected polarizations, $\Delta f_{\rm bw}$ is the observing bandwidth, $t_{\rm obs}$ is the integration time, and $\beta > 1$ is a degradation factor that accounts for imperfections during the digitization of the signal. Moreover, $T_{\rm sys}$ denotes the system temperature, and $T_{\rm sky}(l,\, b)$ is the sky background temperature dominated by synchrotron emission of Galactic electrons, which varies strongly with latitude, $l$, and longitude, $b$. To model the latter, we use results from Remazeilles et al. (2015), who provided a refined version of the temperature map of Haslam et al. (1981, 1982). As the underlying data were obtained at 408 MHz, we rescale to the relevant observing frequencies by assuming a power-law dependence of the form $T_{\rm sky} \propto f^{-2.6}$ (Lawson et al. 1987; Johnston et al. 1992).

A synthetic pulsar counts as detected if the value obtained from Equation (22) exceeds the surveys' sensitivity thresholds. We aim to recover the numbers of detected isolated Galactic radio pulsars for each survey, i.e.,

$$\begin{aligned} &\text{PMPS: 1009 observed pulsars,}\\ &\text{SMPS: 218 observed pulsars,}\\ &\text{HTRU: 1023 observed pulsars.} \end{aligned} \tag{23}$$

To obtain these values, we used the data from the ATNF Pulsar Catalogue (Manchester et al. 2005)[5] and removed extragalactic sources and those in globular clusters. We further applied a

[5] https://www.atnf.csiro.au/research/pulsar/psrcat/; v1.69

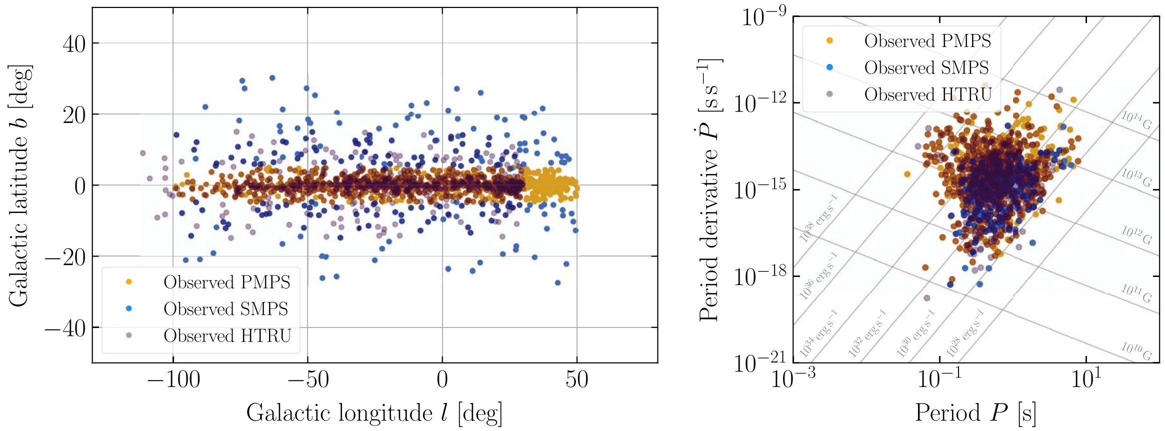

**Figure 3.** Observed populations of isolated Galactic radio pulsars detected with PMPS, SMPS, and the low- and mid-latitude HTRU survey (highlighted in yellow, light blue, and purple, respectively). The left panel shows the distribution of these three populations in Galactic latitude, $b$, and longitude, $l$, while the right panel depicts the detected pulsars in the period, $P$, and period derivative, $\dot{P}$, plane. In the latter, we also give lines of constant spin-down power, $|\dot{E}_{\rm rot}|$, and constant dipolar surface magnetic field, $B$ (estimated via Equation (10) for an aligned rotator). Data taken from the ATNF Pulsar Catalogue (Manchester et al. 2005, https://www.atnf.csiro.au/research/pulsar/psrcat/, v1.69). Observational filters are described in detail in the text.

cutoff in period ($P > 0.01$ s) and period derivative ($\dot{P} > 10^{-19}$ s s$^{-1}$; for those objects with measured $\dot{P}$ values because the above counts also include a small number of pulsars without $\dot{P}$ measurements) to remove those objects that have (likely) been spun up by accretion from a companion star and cannot be modeled with the framework discussed so far. The locations of those objects with known period and period derivatives are shown in the $P$–$\dot{P}$ plane in the right panel of Figure 3. The distribution of mean flux densities, $S_{\rm mean}$, measured at 1400 MHz for isolated Galactic pulsars in our three surveys as recorded in the ATNF catalog is shown in Figure 4. Note that this database does not contain flux measurements for all sources and that uncertainties on reported $S_{\rm mean}$ values can be large. We also note that $S_{\rm mean}$ values in the catalog do not form a homogeneous sample, as there is no standardized way for $S_{\rm mean}$ measurements to be reported in the literature. For example, in some cases $S_{\rm mean}$ is measured by observing a flux calibration source, while other values are estimated using the radiometer equation given by Equation (22), introducing additional systematics due to different prescriptions for the S/N or pulse width. For ease of comparison with our simulated pulsar populations, we also show kernel density estimation (KDE) fits for the corresponding probability density functions obtained with a Gaussian kernel in Figure 4.

### 2.6. Simulation Output

To simulate our *mock* observed pulsar populations, we do not make any assumptions on the neutron star birth rate. Instead, we randomly sample a subset of $10^5$ neutron stars from our dynamical database (see Section 2.2). We subsequently evolve these stars magnetorotationally, as outlined in Section 2.3, and assess how many of them are detected by each of the three surveys (see Sections 2.4 and 2.5), saving their respective properties. We iterate this process until the number of detected stars matches the number of observed objects in all surveys. Note that we adaptively reduce the number of stars we draw from our dynamical database to $10^4$ and $5 \times 10^3$, once we have recovered 90% and 95% of the target values, respectively. The output of a single simulator run, which has a typical computation time of around 1 hr, is a data frame containing the properties of those pulsars we can detect with PMPS, SMPS, and HTRU, respectively.

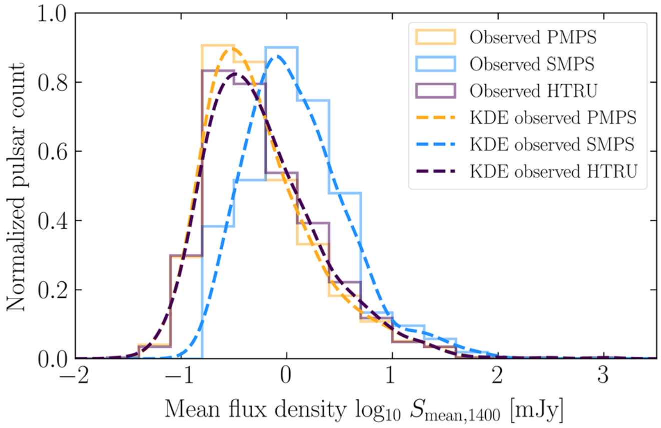

**Figure 4.** Distributions of mean radio flux density measurements, $S_{\rm mean,1400}$, at 1400 MHz for the populations of isolated Galactic radio pulsars detected with PMPS, SMPS, and the low- and mid-latitude HTRU survey (in yellow, light blue, and purple, respectively). We show the normalized number of stars as a function of radio flux density as solid lines. Dashed lines represent the corresponding probability density functions obtained via KDE using a Gaussian kernel. Data taken from the ATNF Pulsar Catalogue (Manchester et al. 2005, https://www.atnf.csiro.au/research/pulsar/psrcat/, v1.69).

The location of the resulting synthetic population and the shape of the stars' distribution in the $P$–$\dot{P}$ plane are directly controlled by the magnetorotational parameters, $\mu_{\log B}$, $\sigma_{\log B}$, $\mu_{\log P}$, $\sigma_{\log P}$, and $a_{\rm late}$, the five parameters we want to infer. Three examples of synthetic $P$–$\dot{P}$ diagrams are shown in the top row of Figure 5.

We note that our prescription does not rely on a *by-hand* implementation of a *pulsar death line* (e.g., Bhattacharya et al. 1992; Chen & Ruderman 1993; Rudak & Ritter 1994; Zhang et al. 2000), beyond which pulsar emission ceases, as done in most previous population synthesis studies (e.g., Faucher-Giguère & Kaspi 2006; Bates et al. 2014; Cieślar et al. 2020). We opt for this approach owing to significant uncertainties around the radio emission process generally associated with the production of electron−positron pairs in pulsar magnetospheres above the polar caps (Ruderman & Sutherland 1975).

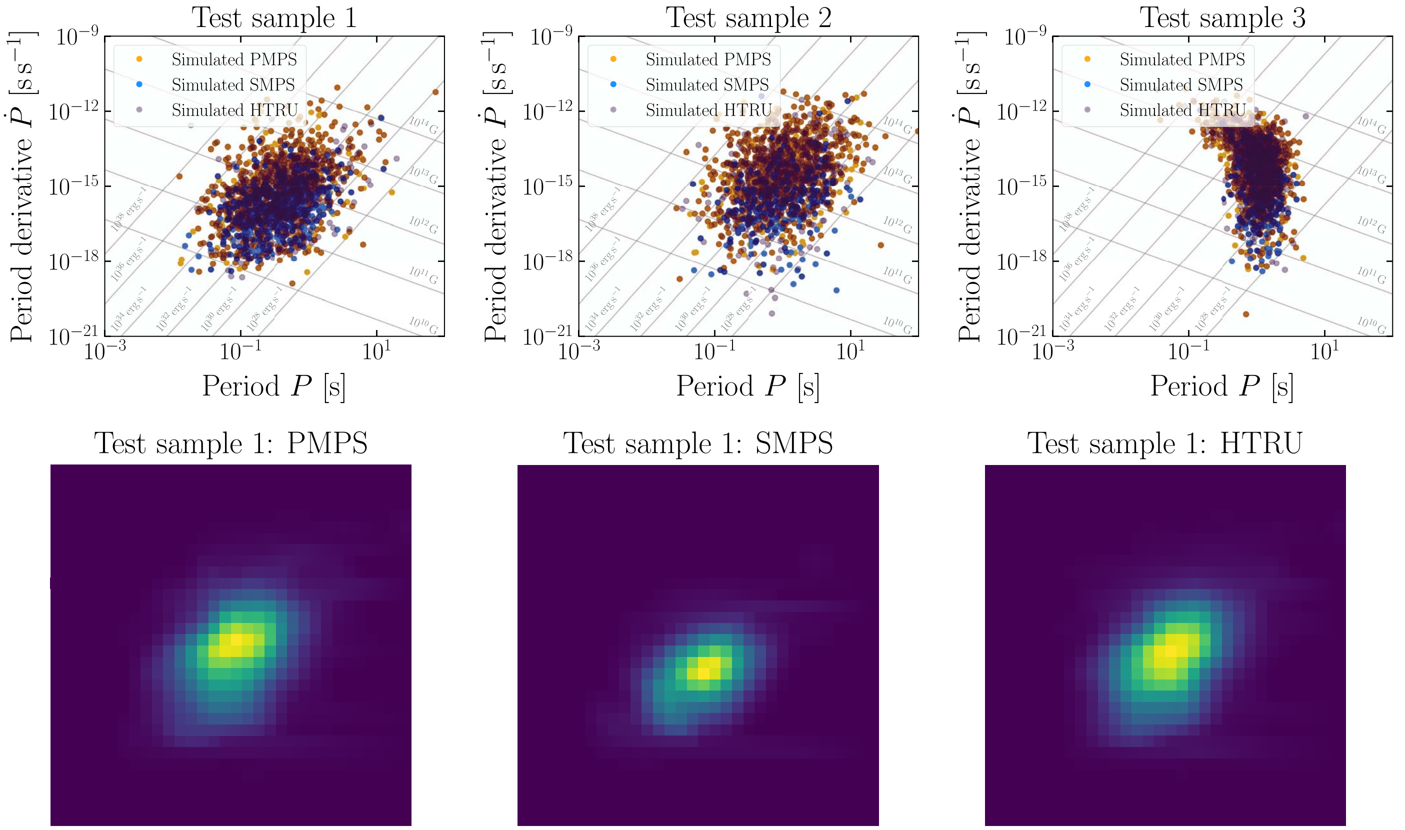

**Figure 5.** Examples of simulated pulsar populations and the corresponding density maps, which are fed into the SBI pipeline. The top row shows synthetic $P$–$\dot{P}$ diagrams for the three surveys considered in this study generated from three random sets of magnetorotational parameters. In particular, test sample 1 (top left) is the result of a simulation with $\mu_{\log B} \approx 13.19$, $\sigma_{\log B} \approx 0.96$, $\mu_{\log P} \approx -0.85$, $\sigma_{\log P} \approx 0.51$, and $a_{\rm late} \approx -0.86$, while test sample 2 (top middle) was generated with $\mu_{\log B} \approx 13.86$, $\sigma_{\log B} \approx 0.88$, $\mu_{\log P} \approx -0.42$, $\sigma_{\log P} \approx 0.61$, and $a_{\rm late} \approx -1.71$. Finally, test sample 3 (top right) corresponds to $\mu_{\log B} \approx 13.35$, $\sigma_{\log B} \approx 0.24$, $\mu_{\log P} \approx -1.25$, $\sigma_{\log P} \approx 0.60$, and $a_{\rm late} \approx -2.38$. The bottom row shows the three density maps (one for each survey) generated with a resolution of 32 from the $P$–$\dot{P}$ diagram for test sample 1. Here dark blue encodes regions where no neutron stars are present, while yellow bins represent the largest density for the binned pulsar distribution.

In particular, different assumptions on magnetic field strengths and geometries, pair production, and stellar properties (like mass and radius) lead to different death lines, effectively expanding into a *death valley*. We thus avoid adopting a somewhat arbitrary choice for a single death line. In our simulations, pulsars instead become undetectable naturally if they approach the bottom right of the $P$–$\dot{P}$ plane. This is due to the evolution toward (i) smaller misalignment angles, $\chi$, resulting in smaller beaming fractions, and (ii) smaller $\dot{P}$ (and thus lower $|\dot{E}_{\rm rot}|$), ultimately leading to sources that are too faint to be detected.

At this point, we also highlight that our approach provides information on the number of total stars generated over a timescale of $10^8$ yr (the oldest possible age for stars in our dynamical database), implying that we can directly determine the birth rate required to reproduce observations for a given survey. Although not the primary focus of this work, we note two things here: first, the number of detectable neutron stars per iteration step described above and, thus, the birth rate (as well as the distribution of stars in the $P$–$\dot{P}$ plane) depend strongly on the five magnetorotational parameters. For some parameter combinations, reaching the counts in Equation (23) requires unrealistically large birth rates, and thus extensive computation time. To mitigate this issue, we stop our iterative simulation approach once the birth rate exceeds a conservative limit of five neutron stars per century (Keane & Kramer 2008; Rozwadowska et al. 2021), even though this implies that we do not reach the numbers of observed objects in these simulations. We, however, still use these simulations in the following to assess whether our inference approach can identify those parameter combinations that require birth rates $\gtrsim 5$ as unreasonable from the distribution of stars in the $P$–$\dot{P}$ plane alone. Second, for a single simulation run, we generally do not obtain the same birth rate for all three surveys, and estimates can differ by a factor of ∼1–3 neutron stars per century. In principle, we only expect the *correct* physical simulator to produce the observed distributions of pulsars across different surveys. The correct simulation framework is, however, not known, and constraining the relevant physics is the main goal of our analysis. To explore this behavior, we thus produce neutron stars until the target values in all three surveys are reached (or exceeded). While this implies that the number of detected objects in some simulations can be larger than the observed number of stars for a given survey (by up to a factor of ∼3), our focus on the location and shape of the distribution of pulsars in $P$ and $\dot{P}$ and not their total number (see below) circumvents this issue. We will, however, return to the issue of the birth rate in the discussion in Section 5.6, once we have explained our inference approach and provided results for our best estimates.

To provide a broad range of synthetic $P$–$\dot{P}$ diagrams for our inference pipeline, we explore the ranges outlined in Section 2.3 and uniformly sample random combinations of the five parameters as follows:

$$\begin{aligned}&\mu_{\log B} \in \mathcal{U}(12,\ 14),\\ &\sigma_{\log B} \in \mathcal{U}(0.1,\ 1),\\ &\mu_{\log P} \in \mathcal{U}(-1.5,\ -0.3),\\ &\sigma_{\log P} \in \mathcal{U}(0.1,\ 1),\\ &a_{\rm late} \in \mathcal{U}(-3,\ -0.5).\end{aligned} \tag{24}$$

We generate a total of 360,000 parameter combinations (which we refer to as our input parameters, labels, or ground truths below) and simulate the corresponding synthetic populations in parallel over the course of 6 weeks.

To represent the discrete output of our simulator in a way that can be processed by a neural network, we convert a single $P$–$\dot{P}$ diagram for three surveys as seen in the top row of Figure 5 into three two-dimensional density maps (one for each survey) by counting the number of stars within a given bin. In particular, we set the limits $P \in [0.001,\ 100]$ s and $\dot{P} \in [10^{-21},\ 10^{-9}]$ s s$^{-1}$ and test our inference procedure for a resolution of 32 and 64 bins. To avoid sharp edges in our binned distributions, we apply a smoothing Gaussian filter (with radius $4\sigma$ and $\sigma = 1$), which will also improve the stability during the training of our machine-learning pipeline. An example of the resulting density maps is shown in the bottom row of Figure 5 for one of our test simulations.

The final preprocessing stage for our simulated data is either a normalization or a standardization step (depending on the choice of setup discussed below) to provide the neural network with signals and labels of similar magnitude. In the former case, the bins in each individual density map are rescaled such that they contain continuous values between 0 and 1. The same holds for the corresponding labels, which are normalized over the entire parameter ranges given in Equation (24). On the other hand, standardization is achieved by using $z$-scores, so that the resulting information in each map has a mean of 0 and standard deviation of 1. The same method is applied to the labels across our entire set of simulations.

## 3. Simulation-based Inference

### 3.1. Overview

The pulsar population synthesis pipeline summarized in Section 2 is a typical example of a stochastic forward model that aims to emulate real-world observations. We specifically introduced stochasticity by sampling relevant variables from underlying probability distributions using Monte Carlo techniques. In particular, given the input parameter, $\boldsymbol{\theta} = \{\theta_1,\ \theta_2, ...\}$, our simulator generates a synthetic realization of the observed data, $\boldsymbol{x}$. The key challenge is then to constrain our model parameters in such a way that they are consistent with true observations, $\boldsymbol{x}_0$, and our prior knowledge, encoded in the prior distribution, $\mathcal{P}(\boldsymbol{\theta})$. To this end, we want to compute the posterior distribution, $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})$, using Bayes's theorem

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\mathcal{P}(\boldsymbol{\theta})\mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta})}{\mathcal{P}(\boldsymbol{x})}, \tag{25}$$

where $\mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta})$ is the likelihood of our data, $\boldsymbol{x}$, given the parameter, $\boldsymbol{\theta}$, and

$$\mathcal{P}(\boldsymbol{x}) \equiv \int \mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta}')\mathcal{P}(\boldsymbol{\theta}')\, d\boldsymbol{\theta}' \tag{26}$$

denotes the evidence obtained by marginalizing over all $\boldsymbol{\theta}$. However, for complex simulators like ours, we typically cannot write down an explicit form of the likelihood function, so $\mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta})$ is essentially intractable. In addition, even if the likelihood were tractable, Equation (26) involves an integral over $\boldsymbol{\theta}$, which becomes challenging for simulators with high-dimensional parameter spaces.

SBI circumvents these issues by taking advantage of the fact that our simulator encodes the likelihood function implicitly (see Cranmer et al. 2020, for a recent review). These approaches have been particularly successful in combination with deep-learning techniques because neural networks can be used to learn a probabilistic association between a given simulation outcome, $\boldsymbol{x}$, and the input parameters, $\boldsymbol{\theta}$. This allows an approximation of the posterior distribution, $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})$, without the need to explicitly compute the likelihood. Three approaches exist to achieve this goal:

1. Neural posterior estimation (NPE): The network learns to directly map the simulator output, $\boldsymbol{x}$, onto the posterior distribution, $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})$, for the underlying parameters, $\boldsymbol{\theta}$. This requires the use of a flexible neural density estimator such as a *normalizing flow* or a mixture density network (MDN; e.g., Papamakarios & Murray 2016; Lueckmann et al. 2017; Greenberg et al. 2019; Dax et al. 2021; Mishra-Sharma & Cranmer 2022; Hahn et al. 2023; Vasist et al. 2023).
2. Neural likelihood estimation (NLE): The network emulates the simulator by learning an association between $\boldsymbol{\theta}$ and $\boldsymbol{x}$, thus providing direct access to an approximation of the likelihood, $\mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta})$. Because the prior is known, the posterior can then be obtained by an additional MCMC sampling step (e.g., Papamakarios et al. 2018; Alsing et al. 2019).
3. Neural ratio estimation (NRE): Here the network learns the likelihood-to-evidence ratio, $r(\boldsymbol{\theta},\ \boldsymbol{x}) \equiv \mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta})/\mathcal{P}(\boldsymbol{x})$, which is equivalent to $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})/\mathcal{P}(\boldsymbol{\theta})$ using Bayes's theorem (Equation (25)). Once $r(\boldsymbol{\theta},\ \boldsymbol{x})$ is known, the posterior can be recovered through MCMC by sampling the prior weighted by the ratio, $r(\boldsymbol{\theta},\ \boldsymbol{x})$ (e.g., Hermans et al. 2019; Miller et al. 2021; Bhardwaj et al. 2023).

For the following study, we choose an NPE approach to directly learn the posterior conditional on our simulated data (avoiding the additional sampling step required for NLE and NRE) and take advantage of the corresponding implementation in the open-source Python package `sbi` (Tejero-Cantero et al. 2020).[6]

### 3.2. Deep-learning Setup

For NPE, we approximate the posterior using a family of densities, $q_\psi$, characterized by the distribution parameters, $\boldsymbol{\psi}$. For our SBI pipeline, we then use a neural network, $F$, to learn these $\boldsymbol{\psi}$ for our simulator output, $\boldsymbol{x}$, by adjusting the network weights, $\boldsymbol{\phi}$. In particular, we aim to optimize the neural density estimator such that $q_{F(\boldsymbol{x},\boldsymbol{\phi})}(\boldsymbol{\theta}) \approx \mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})$. This can be achieved by minimizing the Kullback–Leibler divergence, $D_{\rm KL}(\mathcal{P}_1||\mathcal{P}_2)$, which is a measure of the difference between two probability distributions, $\mathcal{P}_1$ and $\mathcal{P}_2$ (Kullback & Leibler 1951). Papamakarios & Murray (2016) showed that this is equivalent to

[6] https://github.com/sbi-dev/sbi

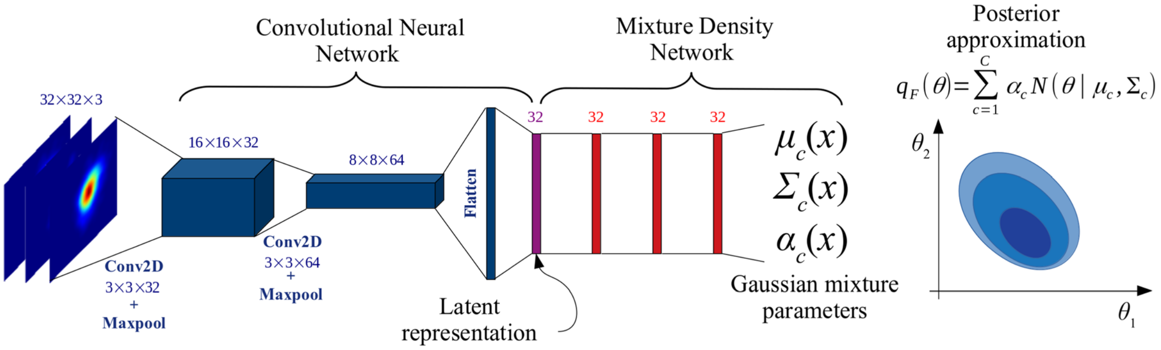

**Figure 6.** Schematic representation of our inference pipeline for three input $P$–$\dot{P}$ maps (one for each survey) with resolution 32 × 32. A CNN is first used to extract features from our images and produce a compressed representation of our simulation output, $\boldsymbol{x}$. We then train a Gaussian MDN, a flexible neural density estimator, on this latent representation to approximate the posterior distribution of the simulation input parameters, $\boldsymbol{\theta}$.

minimizing the expectation value of the loss function

$$\mathcal{L}(\phi) = -\sum_{i=1}^{N} \log q_{F(\boldsymbol{x}_i,\phi)}(\boldsymbol{\theta}_i) \tag{27}$$

over a training data set $\{\boldsymbol{\theta}_i, \boldsymbol{x}_i\}$ of size $N$, provided that $N$ is large and the density estimator is sufficiently flexible. In practice, we maximize the negative of $\mathcal{L}(\phi)$, i.e., the total log-posterior. A key advantage of the resulting posterior approximation is that the evaluation of $q_{F(\boldsymbol{x},\phi)}(\boldsymbol{\theta})$ corresponds to a simple forward pass through a neural network (without the need to simulate additional data), which is very fast. We will take advantage of this *amortized nature* of the posterior to assess the quality of our inferences below.

For our pulsar study, we have drawn the model parameters $\boldsymbol{\theta}_i = \{\mu_{\log B}, \sigma_{\log B}, \mu_{\log P}, \sigma_{\log P}, a_{\rm late}\}$ from uniform priors as defined previously in Equation (24). The corresponding output, $\boldsymbol{x}_i$, of a single run through the simulator are the three $P$–$\dot{P}$ density maps (one for each survey) illustrated in the bottom row of Figure 5. In the following, we stack these maps together to form a three-channel input for our neural network. Of the 360,000 synthetic simulations produced, we use 90% for training and validation, reserving the remaining 10% for testing purposes. The former data set is further split into 90% for training (291,600 populations) and 10% for validation (32,400 populations). We note that as each population is represented by three density maps, we train the following inference pipeline on roughly 875,000 images. Performance results for the unseen test samples quoted in the following are computed for 10% of the full test set (3600 populations) for computational reasons. The full workflow is illustrated schematically in Figure 6.

Due to the complexity of these data, we do not train a neural density estimator directly on the density maps. We instead first apply a convolutional neural network (CNN) to extract features from our images and embed the corresponding information in a lower-dimensional latent vector. We choose the following baseline architecture for our embedding network:

1. Two-dimensional convolution layer with kernel size 3 × 3, 3 input channels, 32 output channels, stride 1, padding 1.
2. Two-dimensional Max pooling layer with size 2 × 2, stride 2, no padding.
3. Two-dimensional convolution layer with kernel size 3 × 3, 32 input channels, 64 output channels, stride 1, padding 1.
4. Two-dimensional Max pooling layer with size 2 × 2, stride 2, no padding.
5. Fully connected linear layer with the flattened output from the second pooling layer as input and 32 output neurons encoding the latent representation.

After each convolution and the fully connected layer, we apply a rectified linear unit (ReLU) activation function. The weights for the CNN are initialized using the Kaiming prescription (He et al. 2015) to avoid exploding or vanishing gradients during the training process.

We subsequently pass the latent vector generated by the CNN to a neural density estimator. We implement an MDN and specifically opt for a Gaussian mixture model in five dimensions to approximate the posterior, $q_{F(\boldsymbol{x},\phi)}(\boldsymbol{\theta})$, for our five free magnetorotational parameters. This implies

$$q_{F(\boldsymbol{x},\phi)}(\boldsymbol{\theta}) = \sum_{c=1}^{C} \alpha_c\, \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \tag{28}$$

where $C$ denotes the total number of Gaussian components used, $\alpha_c$ is the mixture weight, and $\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ is the multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}_c$ and covariance matrix $\boldsymbol{\Sigma}_c$ for the $c$th component.

For our MDN, we follow `sbi`'s default implementation and use the following:

1. Three fully connected layers with 32 neurons each.
2. Four fully connected output layers that encode the Gaussian mixture weights, $\alpha_c$, means, $\boldsymbol{\mu}_c$, and diagonal and upper triangular components of the covariance matrices, $\boldsymbol{\Sigma}_c$. These contain $c$, $5c$, $5c$, and $10c$ neurons, respectively.

We again apply the ReLU activation function after each hidden layer, while weights are now initialized with PyTorch's default initialization (Glorot & Bengio 2010).

We subsequently train the entire pipeline using the gradient descent optimizer Adam (Kingma & Ba 2014). At each epoch the network undergoes a series of optimization steps based on the information provided in the entire training data set before epoch-averaged training and validation metrics are computed based on the negative losses defined in Equation (27), i.e., we maximize our metrics. Note that we also set an early stop of 20 to prevent overfitting, which implies that the training process is interrupted (and the weights of the best validation epoch recorded) once the validation metric has not improved for 20 epochs.

**Table 2**
Information for the 22 Machine-learning Experiments Conducted for This Study

| No. | Res | Surveys | Frac (%) | Input | Comp | BS | LR | CNN | VM | Epochs | Time (s) | TM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 32 | PMPS, SMPS, HTRU | 100 | std | 10 | 8 | 0.0005 | baseline | 3.65 | 38 | 9,373 | 3.64 |
| **2** | 32 | PMPS, SMPS, HTRU | 100 | std | 10 | 8 | 0.0005 | **deep** | 3.71 | 49 | 14,292 | 3.71 |
| **3** | **64** | PMPS, SMPS, HTRU | 100 | std | 10 | 8 | 0.0005 | baseline | 3.55 | 55 | 78,837 | 3.54 |
| **4** | **64** | PMPS, SMPS, HTRU | 100 | std | 10 | 8 | 0.0005 | **deep** | 3.64 | 89 | 128,119 | 3.64 |
| **5** | 32 | PMPS, SMPS, HTRU | **75** | std | 10 | 8 | 0.0005 | baseline | 3.74 | 71 | 13,232 | 3.78 |
| **6** | 32 | PMPS, SMPS, HTRU | **50** | std | 10 | 8 | 0.0005 | baseline | 3.56 | 58 | 7,000 | 3.55 |
| **7** ⋆ | 32 | PMPS, SMPS, HTRU | 100 | **norm** | 10 | 8 | bf 0.01 | baseline | 3.47 | 30 | 7,445 | 3.73 |
| **8** | 32 | PMPS, SMPS, HTRU | 100 | **norm** | 10 | 8 | **0.001** | baseline | 9.66 | 54 | 13,015 | 9.60 |
| **9** | 32 | PMPS, SMPS, HTRU | 100 | std | **8** | 8 | 0.0005 | baseline | 3.74 | 52 | 12,389 | 3.73 |
| **10** | 32 | PMPS, SMPS, HTRU | 100 | std | **5** | 8 | 0.0005 | baseline | 3.83 | 118 | 27,973 | 3.86 |
| **11** | 32 | PMPS, SMPS, HTRU | 100 | std | 10 | **16** | 0.0005 | baseline | 3.99 | 85 | 10,476 | 3.97 |
| **12** | 32 | PMPS, SMPS, HTRU | 100 | std | 10 | **32** | 0.0005 | baseline | 4.11 | 79 | 5,346 | 4.06 |
| **13** | 32 | PMPS, SMPS, HTRU | 100 | std | 10 | 8 | **0.001** | baseline | 3.36 | 61 | 14,785 | 3.33 |
| **14** | 32 | PMPS, SMPS, HTRU | 100 | std | 10 | 8 | **0.0001** | baseline | 4.22 | 75 | 18,369 | 4.22 |
| **15** | 32 | **HTRU** | 100 | std | 10 | 8 | 0.0005 | baseline | 3.43 | 63 | 15,568 | 3.42 |
| **16** | 32 | **SMPS, HTRU** | 100 | std | 10 | 8 | 0.0005 | baseline | 3.58 | 40 | 9,979 | 3.59 |
| **17** | 32 | **PMPS, SMPS** | 100 | std | 10 | 8 | 0.0005 | baseline | 3.41 | 69 | 16,937 | 3.41 |
| **18** ⋆ | **64** | PMPS, SMPS, HTRU | **50** | std | 10 | 8 | 0.0005 | baseline | 3.45 | 47 | 5,766 | 3.44 |
| **19** | 32 | PMPS, SMPS, HTRU | 100 | **norm** | 10 | **32** | **0.001** | baseline | 10.05 | 44 | 2,864 | 10.20 |
| **20** | 32 | PMPS, SMPS, HTRU | 100 | **norm** | 10 | **32** | **0.0001** | baseline | 10.31 | 90 | 5,815 | 10.49 |
| **21** | 32 | PMPS, SMPS, HTRU | 100 | **norm** | 10 | **16** | **0.001** | baseline | 9.82 | 77 | 9,901 | 9.98 |
| **22** ⋆ | 32 | PMPS, SMPS, HTRU | 100 | **norm** | 10 | **16** | **0.0001** | baseline | 10.45 | 124 | 15,603 | 10.55 |

**Note.** The columns summarize the specific training data and hyperparameters, as well as the resulting metrics: the experiment number; the resolution for our $P$–$\dot{P}$ density maps; the different surveys and the fraction of the 291,600 populations in the training set used for training; information on whether we standardized (std) or normalized (norm) the input; the number of Gaussian components in our MDN; the batch size; the learning rate; the CNN architecture (we distinguish our baseline setup and a deeper network; see Sections 3.2 and 3.3 for details); the best metric computed over the validation set; the number of training epochs; the time it took to train the network in seconds; and the average metric computed over our 3600 test samples. In bold, we highlight those parameters that we have varied with respect to the baseline experiment #1. Experiments with an asterisk (⋆) are removed from the following analysis owing to training irregularities.

### 3.3. Experiments

Table 2 summarizes the 22 different experiments that we have conducted for this study to assess the performance of SBI for pulsar population synthesis. For this purpose, we varied aspects of the training data, as well as the hyperparameters of our deep-learning pipeline. In particular, for the input we explored two different resolutions for the $P$–$\dot{P}$ maps, 32 and 64, assessed the network performance when all three density maps or only two/one are provided, and examined whether normalization or standardization during preprocessing leads to different results. We further studied the impact of using the full training data set or smaller subsets. Moreover, for the network we varied the number of Gaussian mixture components in our neural density estimator, the batch size, and the learning rate, and we explored two different CNNs for our embedding net. In addition to the baseline architecture described in Section 3.2, we also conducted two experiments with a deeper network composed of four convolutional blocks. Here the two convolutional layers introduced previously are followed by an additional layer with 32 and 64 input/output channels, respectively. Kernel size, stride, padding, subsequent pooling, and fully connected layers were kept as above.

Due to the computational cost of each training experiment, a full grid search over all relevant configurations was beyond the scope of this work. We therefore opted to produce a representative set of experiments that provide sufficient information to study the variation of our inferred posteriors in Section 4. Finally, note that almost all of our optimizations are performed on a Tesla V100 SXM2 GPU with 32GB memory.

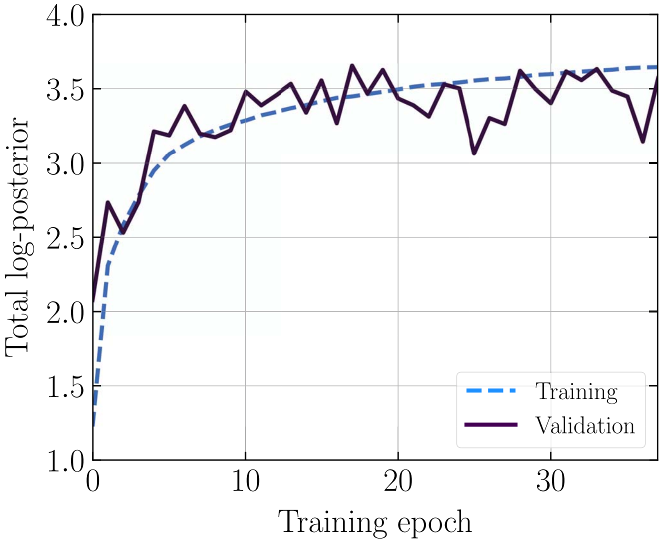

**Figure 7.** Training behavior for baseline experiment #1. We show the training metric (light-blue dashed line) and the validation metric (purple solid line) as a function of the training epoch. We seek to maximize the total log-posterior, $\sum_{i=1}^{N} \log q_{F(x_i,\phi)}(\theta_i)$, over the training and validation data sets, respectively, as the network learns. Both metrics increase as expected, and the validation curve closely tracks the training curve, i.e., we see little overfitting. The best validation metric is reached at epoch 17, and thus the early stop criterion halts the training after 37 training epochs.

We only trained experiments #3 and #4, for which the full training data set with a resolution of 64 was too large to be optimized on the GPU, on a CPU with 32GB RAM. In those

**Table 3**
Magnetorotational Parameters for Three Random Test Samples and the Observed Pulsar Population

| Parameters | | Test Sample 1 | Test Sample 2 | Test Sample 3 | Observed Population |
|---|---|---|---|---|---|
| Ground truths, $\theta$ | $\mu_{\log B}$ | 13.19 | 13.86 | 13.35 | ⋯ |
| | $\sigma_{\log B}$ | 0.96 | 0.88 | 0.24 | ⋯ |
| | $\mu_{\log P}$ | −0.85 | −0.42 | −1.25 | ⋯ |
| | $\sigma_{\log P}$ | 0.51 | 0.61 | 0.60 | ⋯ |
| | $a_{\rm late}$ | −0.86 | −1.71 | −2.38 | ⋯ |
| 95% CI experiment #1 | $\mu_{\log B}$ | $13.28^{+0.18}_{-0.18}$ | $13.73^{+0.15}_{-0.15}$ | $13.33^{+0.05}_{-0.04}$ | $13.07^{+0.07}_{-0.08}$ |
| | $\sigma_{\log B}$ | $0.95^{+0.08}_{-0.08}$ | $0.79^{+0.07}_{-0.07}$ | $0.23^{+0.02}_{-0.02}$ | $0.43^{+0.03}_{-0.03}$ |
| | $\mu_{\log P}$ | $-0.90^{+0.13}_{-0.13}$ | $-0.35^{+0.19}_{-0.18}$ | $-1.17^{+0.33}_{-0.34}$ | $-0.98^{+0.25}_{-0.29}$ |
| | $\sigma_{\log P}$ | $0.49^{+0.10}_{-0.09}$ | $0.73^{+0.20}_{-0.15}$ | $0.73^{+0.25}_{-0.31}$ | $0.54^{+0.33}_{-0.25}$ |
| | $a_{\rm late}$ | $-0.83^{+0.06}_{-0.06}$ | $-1.88^{+0.35}_{-0.35}$ | $-2.47^{+0.43}_{-0.43}$ | $-1.77^{+0.35}_{-0.38}$ |
| 95% CI ensemble | $\mu_{\log B}$ | $13.29^{+0.20}_{-0.20}$ | $13.74^{+0.19}_{-0.16}$ | $13.34^{+0.05}_{-0.05}$ | $13.10^{+0.08}_{-0.10}$ |
| | $\sigma_{\log B}$ | $0.96^{+0.07}_{-0.08}$ | $0.78^{+0.09}_{-0.08}$ | $0.24^{+0.02}_{-0.02}$ | $0.45^{+0.05}_{-0.05}$ |
| | $\mu_{\log P}$ | $-0.92^{+0.16}_{-0.15}$ | $-0.40^{+0.20}_{-0.27}$ | $-1.23^{+0.33}_{-0.34}$ | $-1.00^{+0.26}_{-0.21}$ |
| | $\sigma_{\log P}$ | $0.49^{+0.10}_{-0.09}$ | $0.74^{+0.20}_{-0.17}$ | $0.67^{+0.30}_{-0.28}$ | $0.38^{+0.33}_{-0.18}$ |
| | $a_{\rm late}$ | $-0.84^{+0.06}_{-0.07}$ | $-1.76^{+0.39}_{-0.43}$ | $-2.34^{+0.43}_{-0.45}$ | $-1.80^{+0.65}_{-0.61}$ |

**Note.** The first five rows show the ground truths, $\boldsymbol{\theta}$, used to simulate the test populations. The second block gives medians and 95% CIs obtained from inferences with the neural network from experiment #1. The final block contains medians and 95% CIs determined from the ensemble posterior combining 19 experiments.

two cases, training the network thus took markedly longer than for the other experiments (see below).

## 4. Results

### 4.1. Training

Several metrics for our experiments are summarized in the last four columns of Table 2. We observe that the optimization of our neural networks takes ∼1–8 hr on the GPU and on the order of a day on a CPU, completing ∼30–124 training epochs. In general, we find good training behavior, with the validation metric closely tracking the training metric and little or no overfitting. This is also evident in the network's generalization ability, illustrated by the average metrics computed over the unseen test set of 3600 simulations. The evolution of the training and validation metrics for experiment #1 is shown in Figure 7 as an example. We remind the reader that we aim to maximize the total log-posterior. After visual inspection of all training curves, we remove experiment #7 owing to irregularities in the training behavior and experiments #18 and #22 owing to a slight tendency to overfitting. Note that these shortcomings were not directly visible from the training metrics in Table 2. We also highlight that we find systematically larger training, validation, and test metrics in those experiments where our input density maps were normalized. In the following, however, we assess the quality of the corresponding posteriors and find that these do not result in better inferences. Beyond this difference, we cannot identify any significant variation in the metrics between the remaining configurations. Hence, we proceed with an analysis of all experiments apart from numbers #7, #18, and #22.

### 4.2. Benchmark Inferences

As a first assessment of our approximated posteriors, we focus on inferring the five magnetorotational parameters, $\mu_{\log B}$, $\sigma_{\log B}$, $\mu_{\log P}$, $\sigma_{\log P}$, $a_{\rm late}$, for simulated populations where we know the input parameters, $\boldsymbol{\theta}$. We specifically look at the three simulations, whose $P$–$\dot{P}$ diagrams were illustrated in the top row of Figure 5. Corresponding ground truths, $\boldsymbol{\theta}$, are summarized in the top five rows in Table 3. In Figures 8 and 9, we show the resulting one- and two-dimensional marginal posterior distributions obtained by repeatedly sampling from the neural network optimized during experiment #1. For all three cases, the posteriors are well-defined, significantly smaller than our prior ranges (Equation (24)) shown along the axes, and centered around the ground truths, $\boldsymbol{\theta}$, highlighted in light blue. To quantify this, we calculate the $1\sigma$, $2\sigma$, and $3\sigma$ credible regions, shown as contours in the two-dimensional posteriors. In the one-dimensional posterior panels, the corresponding 95% credible intervals (CIs) are given as black dashed lines, while medians are illustrated as purple solid lines. Their numerical values are given in Table 3. We observe that the ground truths, $\boldsymbol{\theta}$, are typically contained within the $2\sigma$ credible regions, which we interpret as evidence that our NPE approach is capable of producing reasonable posterior distributions. In general, the credible regions for the two parameters characterizing the initial magnetic field distribution are narrower than those for the initial period distribution and the late-time magnetic field decay. We confirm that this behavior is qualitatively similar for the remaining $P$–$\dot{P}$ simulations in our test set.

We next compare the inferences for our various training experiments. To visualize corresponding differences, we plot the one-dimensional marginalized posteriors for all 19 experiments for the three test samples in gray in Figure 10. Ground truths, $\boldsymbol{\theta}$, are shown as light-blue dashed lines. We observe that the widths of individual posterior approximations and their medians can vary somewhat between different test samples and magnetorotational parameters. Compared across the full test set, this behavior is again more dominant

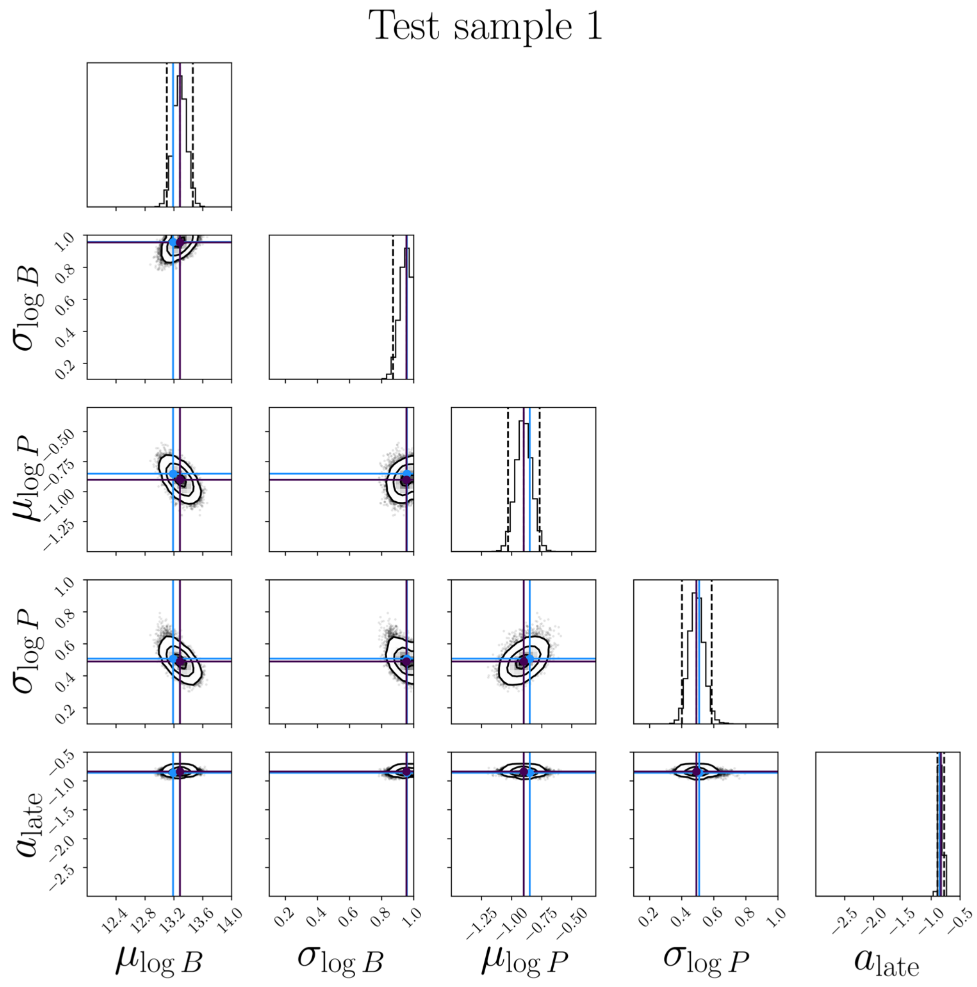

**Figure 8.** Benchmark inference for test simulation 1 using the network from experiment #1. The corner plot shows one- and two-dimensional marginal posterior distributions for the five magnetorotational parameters. We also show corresponding ground truths, $\boldsymbol{\theta}$, in light blue and the medians in purple. We observe that the posteriors cover the $\boldsymbol{\theta}$ well. Corresponding 95% CIs are summarized in Table 3.

for the period and late-time magnetic field parameters than for the initial $B$-field properties. However, no individual NPEs stand out by exhibiting either particularly good or poor posteriors. Further note that we also do not see any differences for those experiments with normalized input maps that showed systematically better metrics than those experiments trained on standardized data. This highlights that training behavior alone does not provide sufficient information on the quality of the resulting inference.

In light of this, we also determine the combined posterior for all 19 experiments. We calculate the corresponding *ensemble posterior*, $\bar{q}(\boldsymbol{\theta})$, as the weighted average of the individual posteriors (Hermans et al. 2021):

$$\bar{q}(\boldsymbol{\theta}) = \sum_{j}^{19} \boldsymbol{w_j q_{F_j}}(\boldsymbol{\theta}), \tag{29}$$

where $w_j$ represents the weight of the $j$th component. Giving equal importance to each experiment in the ensemble, we choose $w_j = 1/19$. The corresponding one-dimensional marginalized ensemble posteriors for $\mu_{\log B}$, $\sigma_{\log B}$, $\mu_{\log P}$, $\sigma_{\log P}$, and $a_{\rm late}$ for the three test simulations are illustrated as purple histograms in Figure 10. As expected, they fall within the individual posteriors. The corresponding 95% CIs for the three test samples, which are typically comparable to or slightly wider than those calculated for experiment #1 posteriors alone, are summarized in the bottom five rows of Table 3.

### 4.3. Posterior Validation

To further assess whether posterior estimates are well calibrated, we determine their *coverage*. As outlined in detail in Appendix B, the coverage probability measures the fraction of test samples for which (for a given credibility level $1 - \alpha$) the ground truths, $\boldsymbol{\theta}$, fall within the corresponding $1 - \alpha$ region of their respective posteriors, $q_{F(\boldsymbol{x},\phi)}(\boldsymbol{\theta})$. For a well-calibrated posterior distribution and a sufficiently large number of test samples, this fraction should equal $1 - \alpha$. This implies that the coverage probability as a function of the credibility level is diagonal. In contrast, for a conservative posterior that is wider than the true posterior, we would recover a fraction larger than $1 - \alpha$. Conversely, for a narrower (overconfident) posterior, the corresponding fraction of test samples is less than $1 - \alpha$. In terms of the coverage, this corresponds to curves above and below the diagonal, respectively, and can therefore be used to assess the quality of approximate posteriors.

We show the coverage probabilities for our different posterior estimates as a function of the credibility level,

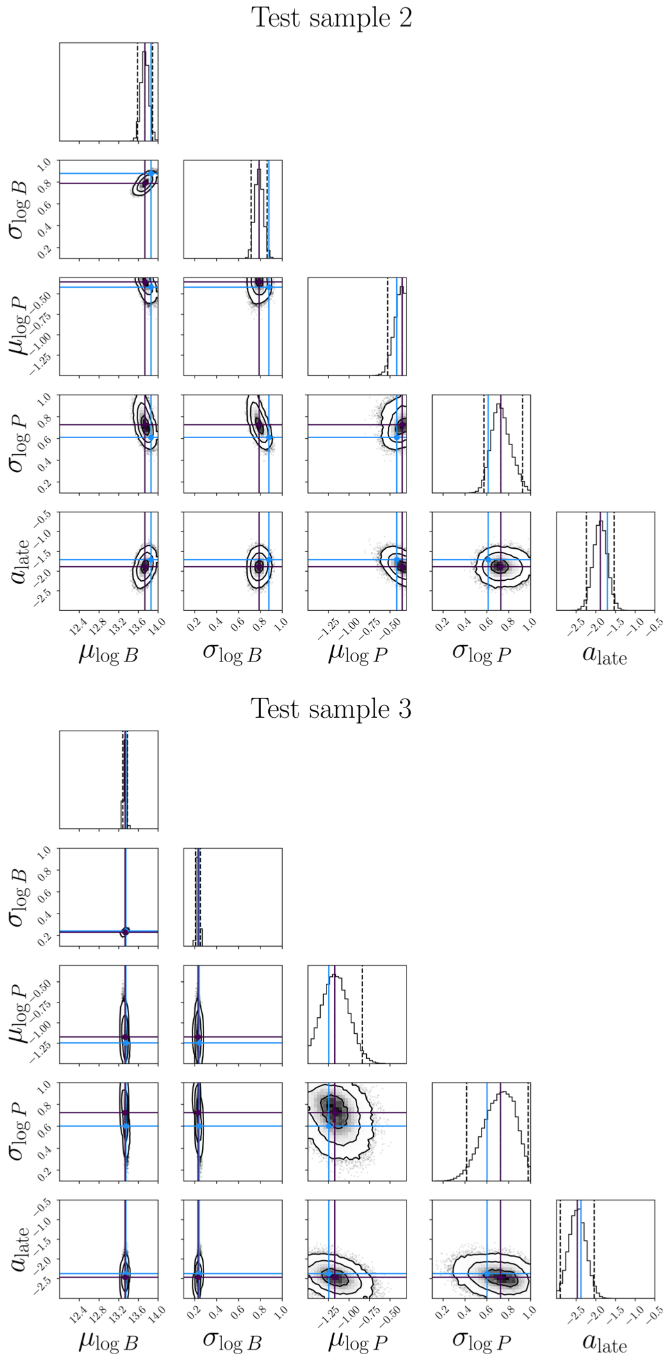

**Figure 9.** Same as Figure 8, but for test simulations 2 and 3.

$1 - \alpha$, in Figure 11. We single out the coverage for the posterior from experiment #1 (light-blue dashed line) and the ensemble posterior (purple solid line). All remaining experiments are shown in gray. We observe that the approximate posteriors for individual experiments closely follow the diagonal, exhibiting either slightly conservative or slightly overconfident behavior. As expected, the most conservative estimate is given by our ensemble posterior, which incorporates variations in the inference for 19 different machine-learning configurations across all 3600 test samples. These results provide additional support that our neural posteriors are trustworthy and have indeed learned to accurately infer magnetorotational parameters from simulated $P$–$\dot{P}$ density maps.

### 4.4. Inference on the Observed Population

Following the benchmark experiments and the coverage determination, we now turn our attention to the true pulsar populations observed with PMPS, SMPS, and the low- and mid-latitude HTRU survey. The corresponding $P$–$\dot{P}$ diagram was shown in the right panel of Figure 3. We represent these populations as three density maps, as outlined in Section 2.6, and subsequently feed them through our trained neural networks to infer the five parameters, $\mu_{\log B}$, $\sigma_{\log B}$, $\mu_{\log P}$, $\sigma_{\log P}$, and $a_{\mathrm{late}}$, assuming that our simulation framework provides a realistic description of the underlying physics.

We show the corresponding one-dimensional marginal posterior distributions for individual experiments (gray histograms) and the ensemble (purple histograms) in Figure 12. Additionally, a corner plot for the one- and two-dimensional ensemble posteriors is illustrated in Figure 13. Corresponding medians (shown in purple in the corner plot) and 95% CIs for experiment #1 and the ensemble are also summarized in the last column of Table 3.

The general trend (already observed for the simulated populations) that the initial magnetic field parameters, $\mu_{\log B}$ and $\sigma_{\log B}$, are much better constrained by our NPE framework than the remaining three values also holds for the observed population. As seen in the first two panels of Figure 12, all 19 experiments recover narrow posteriors around similar medians. For the initial period distribution parameters, $\mu_{\log P}$ and $\sigma_{\log P}$ (see third and fourth panel, respectively), we obtain wider posteriors and a larger variety of median values between different experiments. These posteriors, however, cover similar regions within our prior ranges and are comparable to what we observed for the test samples. In contrast, the inferred posteriors for $a_{\mathrm{late}}$ (the final panel of Figure 12) exhibit different behavior from our benchmark experiments. In particular, posteriors vary significantly in width between different experiments, with those at the larger (smaller) end of the $a_{\mathrm{late}}$ range generally exhibiting narrower (larger) widths. Moreover, several distributions do not overlap at all. This is manifest as a relatively wide posterior in the ensemble that also shows a second peak, primarily driven by the rightmost individual posterior resulting from experiment #2. Note that this configuration did not cause irregularities during the network optimization or unusual posteriors for our test samples. Therefore, we do not associate this behavior with the network itself. The corresponding bimodality is also visible in the final row of the corner plot in Figure 13. We will discuss our interpretation of this below.

## 5. Discussion and Conclusions

In this study, we have successfully developed a new machine-learning pipeline that combines pulsar population synthesis with SBI for the first time and tested the corresponding approach by inferring magnetorotational properties of neutron stars.

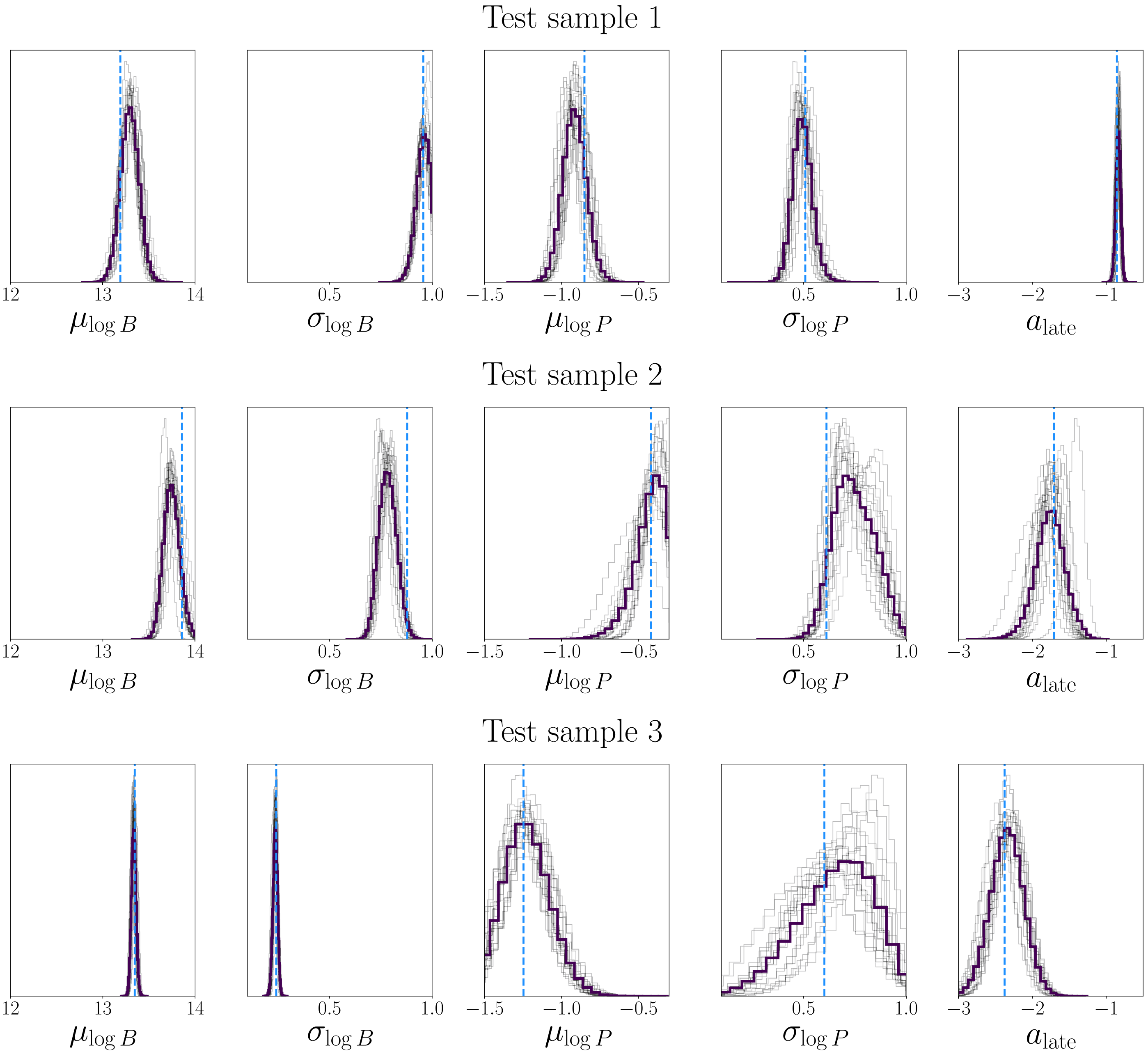

**Figure 10.** One-dimensional marginal posteriors for the five magnetorotational parameters for the three test simulations inferred using 19 different NPE experiments shown in gray. The horizontal axes represent the parameters' prior ranges. The ground truths are shown as vertical light-blue dashed lines. We observe variation between the experiments, specifically for $\mu_{\log P}$, $\sigma_{\log P}$, and $a_{\rm late}$. We also plot the ensemble posteriors (purple) obtained as a weighted average of the individual posteriors.

### *5.1. Simulation Framework*

We first discussed our implementation of the forward model, i.e., the prescription for simulating the dynamical and magnetorotational properties of the Galactic population of isolated radio pulsars, modeling their radio emission and subsequently mimicking observational limitations for PMPS, SMPS, and the low- and mid-latitude HTRU survey. We followed earlier frameworks (e.g., Faucher-Giguère & Kaspi 2006; Bates et al. 2014; Gullón et al. 2014, 2015; Cieślar et al. 2020) but implemented several key differences, as compared in detail in Table 4. In particular, we sampled the birth positions of our pulsars from the Galactic electron distribution (Yao et al. 2017) instead of following the typical approach of combining a spiral arm model with a radial pulsar distribution like that of Yusifov & Küçük (2004). The latter is deduced for the observed, evolved pulsar sample and not the initial population. Moreover, we have included the (rigid) rotation of the Galaxy to treat the pulsar birth positions more consistently compared to earlier analyses. For the magnetic field evolution, we used a similar approach to that of Gullón et al. (2014, 2015), taking advantage of the newest two-dimensional magnetothermal simulations (Viganò et al. 2021), and solved for the coupled evolution of the spin period, $P$, and the misalignment angle, $\chi$, for a plasma-filled magnetosphere. To capture the field changes at

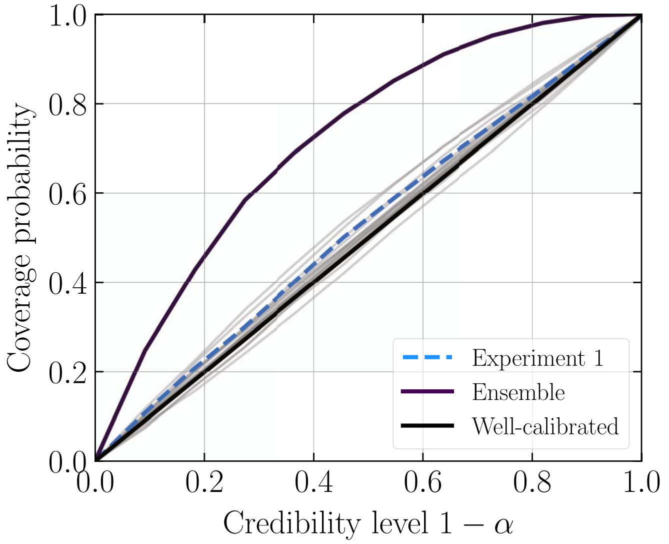

**Figure 11.** Coverage probability as a function of the credibility level, $1-\alpha$, for our approximate posteriors calculated for 3600 test simulations. We specifically highlight the coverage for experiment #1 as a light-blue dashed line and that for the ensemble as a purple solid line. All remaining experiments are given in gray. For a well-calibrated posterior, the coverage follows the diagonal shown in black.

late times, we developed a new physically motivated prescription in which the magnetic field, $B$, decays according to a power law captured by the index, $a_{\mathrm{late}}$. Together with the means, $\mu_{\log B}$, $\mu_{\log P}$, and standard deviations, $\sigma_{\log B}$, $\sigma_{\log P}$, which characterize the normally distributed logarithms of the initial periods and the initial fields, we hence obtained five parameters that control the neutron stars' magnetorotational evolution.

To simulate the detection of our synthetic pulsars, we make the following changes compared to earlier studies: First, we do not model the pulsars' pseudoluminosity, defined as $L_{\mathrm{ps}} \equiv S_{f,\mathrm{obs}} d^2$ (where $S_{f,\mathrm{obs}}$ is the detected flux at frequency, $f$, and $d$ is the pulsar distance), but instead assume that the intrinsic neutron star luminosity, $L_{\mathrm{int}}$, is proportional to the spin-down power, $\dot{E}_{\mathrm{rot}}$. In particular, we considered $L_{\mathrm{int}} \propto |\dot{E}_{\mathrm{rot}}|^{1/2}$ to determine the bolometric radio flux and subsequently propagate the corresponding pulsed emission toward Earth. We also used a geometry-based description to determine the pulsars that are beamed toward us, which earlier works typically treat in an empirical manner. In addition, we do not implement a pulsar death line to quench radio emission but instead let pulsars become undetectable naturally. Finally, we not only looked at PMPS and SMPS but also incorporated the HTRU survey for the first time. Using the resulting simulation framework, we then produced 360,000 synthetic $P$–$\dot{P}$ diagrams, which we converted to one density map per survey in preparation for the neural networks. A total of 90% of these simulations were used for training and validation, and the remaining 10% were reserved for testing.

### 5.2. Inference Procedure

The second part of this study is centered on the implementation of the SBI approach, specifically focusing on NPE, to learn a probabilistic association between our simulator output and the input parameters, $\boldsymbol{\theta} = \{\mu_{\log B}, \sigma_{\log B}, \mu_{\log P}, \sigma_{\log P}, a_{\mathrm{late}}\}$. To do so, we first used a CNN to extract features from our high-dimensional $P$–$\dot{P}$ maps and obtain a compressed representation, which was then transferred into a flexible neural density estimator. By taking advantage of the open-source Python package `sbi` (Tejero-Cantero et al. 2020),[7] we specifically opted for a Gaussian mixture density model in five dimensions to approximate our posterior. To study the sensitivity of the NPE results on the representation of our input data and the network hyperparameters, we conducted 22 distinct experiments. An inspection of the corresponding training metrics led us to discard three experiments owing to irregular training behavior or overfitting. The remaining 19 trained neural networks were analyzed further, and we found no significant differences in the resulting inferences when benchmarked on three random test simulations. The same was observed when validating the posteriors through a coverage calculation over the test set with 3600 samples, highlighting that all 19 posterior estimates are well calibrated. From this we concluded, in particular, that the training behavior is a poor identifier of subsequent inference quality, because normalization of input maps led to systematically better training, test, and validation metrics compared to standardizing the input but comparable inferences. Learning rate and batch size played a negligible role in both setups.

We also point out that the use of smaller training data sets did not affect the inference quality either. While we expect that training sets of $\lesssim$10% (i.e., 30,000 simulations) will eventually have an effect on this, databases of 50% (i.e., 150,000 simulations) are sufficient when inferring five parameters. For comparable studies, this would imply a significant reduction in simulation time, the most costly part of these analyses. Similar performances further justify optimizing our networks for density maps with a resolution of $32 \times 32$ bins instead of $64 \times 64$ and the shallower baseline CNN to speed up the training process. Additionally, we highlight that the use of different numbers of Gaussian mixture components also led to comparable optimization metrics and inference results. Extracting the corresponding mixture weights, $\alpha_c$, after the optimization, we find that across the entire test data set we only require two or three Gaussians to approximate our posteriors. However, we point out that training with a larger number of components was faster owing to fewer training epochs. Finally, note that the use of fewer surveys (i.e., one or two density maps only) did not change the inference results for our five magnetorotational parameters. Naively, one might think that complementary information on the pulsar population as, e.g., provided by SMPS, which is sensitive to older stars at higher Galactic latitudes, would help the network learn better posteriors. However, we do not observe such behavior in our experiments. Although this might suggest that using single surveys in the future could be sufficient to constrain neutron star parameters through population synthesis, we caution that different surveys, in principle, provide additional information on the neutron star birth rate (see below) that was not supplied to our neural networks, i.e., we focused on the location and shape of the pulsar population in the $P$–$\dot{P}$ plane only.

Due to the variations in our inference results, and because we could not identify a single neural network as the best posterior estimator, we also determined the ensemble

[7] https://github.com/sbi-dev/sbi

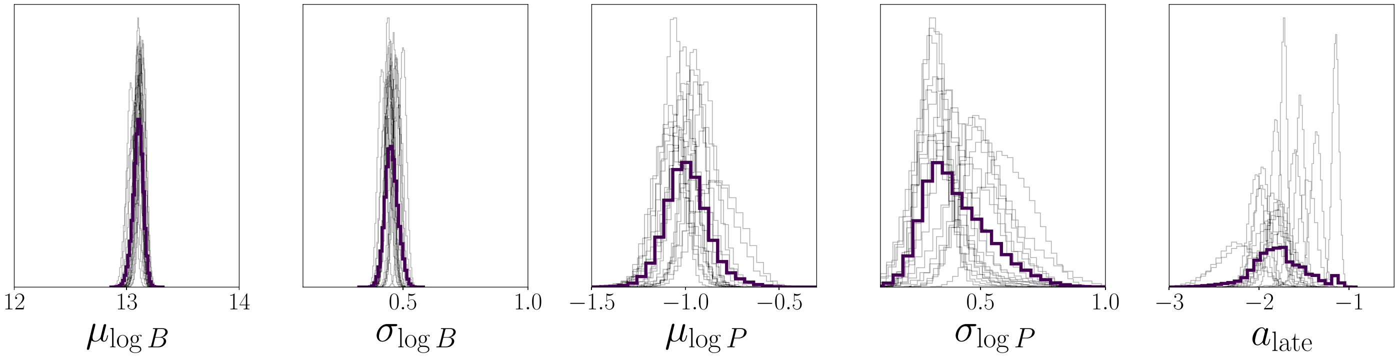

**Figure 12.** One-dimensional marginal posteriors for the five magnetorotational parameters for the observed pulsar population. We show inference results for 19 different NPE experiments in gray and the ensemble posterior in purple.

**Table 4**
Comparison between This Work and Several Population Synthesis Studies in the Literature

| | Faucher-Giguère & Kaspi (2006) | Bates et al. (2014) | Gullón et al. (2014, 2015) | Cieślar et al. (2020) | This Work |
|---|---|---|---|---|---|
| $\mathcal{P}(r, \phi)$ | spiral arms,<br>$\mathcal{P}(r)$ | spiral arms,<br>$\mathcal{P}(r)$ | spiral arms,<br>$\mathcal{P}(r)$ | spiral arms,<br>$\mathcal{P}(r)$ | *e*-density model<br>Yao et al. (2017) |
| $\mathcal{P}(z)$ | exponential | exponential | exponential | exponential | exponential |
| **Galactic rotation** | … | … | … | … | $T \approx 250$ Myr |
| $\mathcal{P}(\boldsymbol{v}_{\rm k})$ | exponential | exponential, normal | exponential | Maxwell | Maxwell |
| $\mathcal{P}(\boldsymbol{B_0})$ | lognormal | lognormal | lognormal | lognormal | lognormal |
| $\mathcal{P}(\boldsymbol{P_0})$ | normal | normal, lognormal | normal | normal | lognormal |
| $\boldsymbol{B(t)}$ | … | … | magnetothermal models<br>Viganò et al. (2013) | exponential<br>decay | magnetothermal models<br>Viganò et al. (2021),<br>late-time power law |
| $\boldsymbol{P(t)}$ | vacuum dipole | vacuum dipole | plasma-filled dipole | vacuum dipole | plasma-filled dipole |
| $\boldsymbol{\chi(t)}$ | … | exponential | $P$–$\chi$ coupled | … | $P$–$\chi$ coupled |
| **Beaming** | empirical | empirical,<br>geometry dependent | empirical | empirical | geometry dependent |
| **Luminosity** | pseudo,<br>$\propto\lvert\dot{E}_{\rm rot}\rvert^{\epsilon}$ | pseudo,<br>$\propto P^{\alpha}\dot{P}^{\beta}$ | pseudo,<br>$\propto\lvert\dot{E}_{\rm rot}\rvert^{\epsilon}$ | pseudo,<br>$\propto\lvert\dot{E}_{\rm rot}\rvert^{\epsilon}$ | intrinsic,<br>$\propto\lvert\dot{E}_{\rm rot}\rvert^{\epsilon}$ |
| **Surveys** | PMPS, SMPS | PMPS, SMPS | PMPS, SMPS<br>+ X-ray pulsars (2015 study) | PMPS | PMPS, SMPS, HTRU |
| **Comparison** | K-S test,<br>by eye | K-S test | annealing method,<br>K-S test | MCMC with<br>Gaussian likelihood | SBI |

**Note.** We compare the following ingredients, which are given as individual table rows: the distributions of sources in the Galactic plane, $\mathcal{P}(r, \phi)$, and along Galactic heights, $\mathcal{P}(z)$, in cylindrical galactocentric coordinates; the inclusion of Galactic rotation and, if present, the corresponding rotation period, $T$; the distribution of neutron star kick velocities, $\mathcal{P}(v_k)$; the distributions of initial dipolar magnetic field strengths and initial periods, i.e., $\mathcal{P}(B_0)$ and $\mathcal{P}(P_0)$, as well as the prescriptions for their evolution (denoted as $B(t)$ and $P(t)$, respectively); the treatment of the misalignment angle evolution, $\chi(t)$; the description of the radio beaming, where pulsars that intercept our line of sight are determined either with an *empirical* relation between the beaming fraction and the period obtained from polarization data (Tauris & Manchester 1998) or with a *geometry-dependent* approach that considers the radio beam aperture and the inclination angle, $\chi$. We further provide information on the luminosity (distinguishing between *pseudo-* and *intrinsic* luminosities), the respective surveys used for comparison, and, finally, the method used to contrast simulated and observed populations (where K-S denotes the Kolmogorov–Smirnov test).

posterior through an equally weighted average of the individual experiments. The resulting posterior behaved as expected and showed more conservative behavior than the ensemble members. For the next section, we will hence follow the recommendation by Hermans et al. (2021) and use our (most conservative) ensemble posterior to analyze the observed pulsar population.

### 5.3. Inference Results on the Observed Population

Following the validation of our NPE approach, we subsequently used the ensemble posterior estimator to infer the five magnetorotational parameters for the true population of isolated Galactic radio pulsars observed with our three surveys. In particular, we found the following best estimates at the 95%

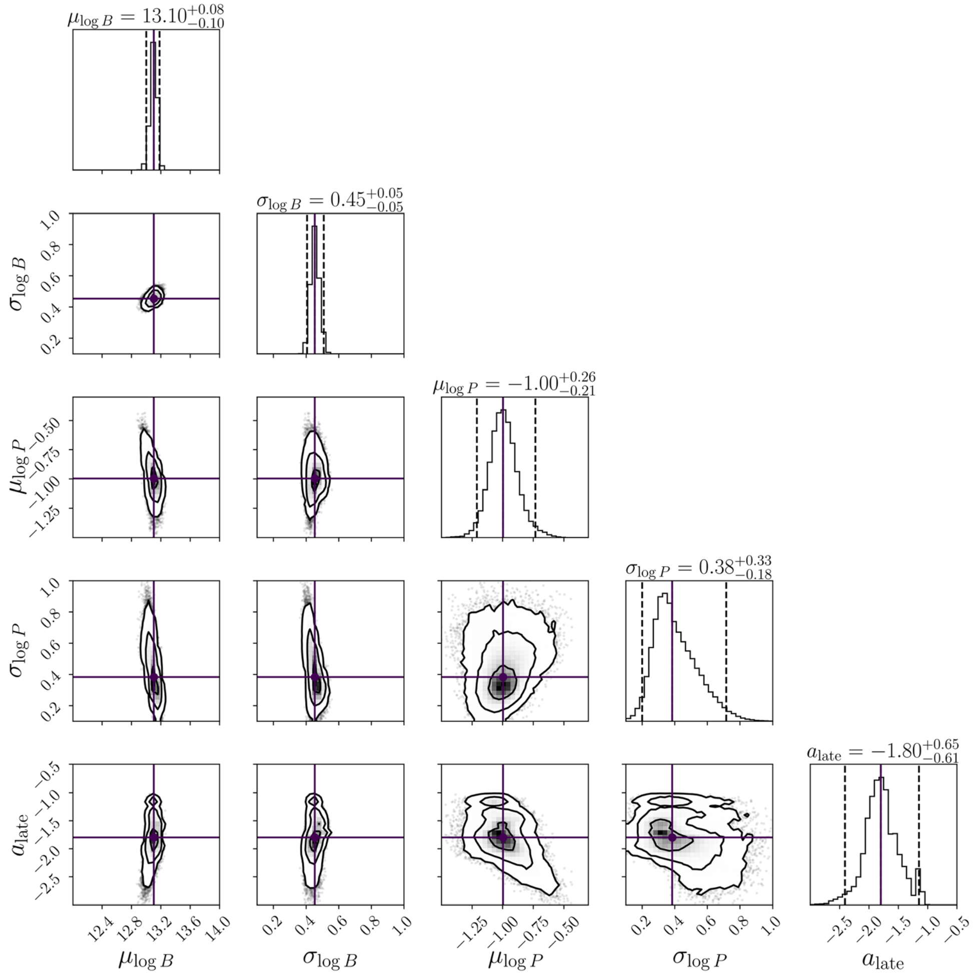

**Figure 13.** Inference results for the observed pulsar population using the ensemble posterior of 19 different NPEs. The corner plot shows one- and two-dimensional marginal posterior distributions for the five magnetorotational parameters. We highlight the medians in purple. Corresponding values and 95% CIs are summarized above the panels and in Table 3.

credible level:

$$\begin{aligned} \mu_{\log B} &= 13.10^{+0.08}_{-0.10}, \\ \sigma_{\log B} &= 0.45^{+0.05}_{-0.05}, \\ \mu_{\log P} &= -1.00^{+0.26}_{-0.21}, \\ \sigma_{\log P} &= 0.38^{+0.33}_{-0.18}, \\ a_{\rm late} &= -1.80^{+0.65}_{-0.61}. \end{aligned} \tag{30}$$

The corresponding corner plot was illustrated in Figure 13, while we show the resulting distributions for the initial magnetic field and period as black solid lines in Figure 14.

As noted during the benchmarking experiments, we generally obtain narrower posterior distributions for the initial magnetic field parameters when compared to the initial period parameters. Difficulties in constraining rotational birth properties are, however, not a shortcoming of our inference approach itself, as this was also noted by earlier population synthesis analyses (e.g., Gullón et al. 2014, 2015). Instead, this has a physical reason that lies in the coupled evolution of the stars' misalignment angle, rotation period, and magnetic field. While the $B$-field initially stays constant (see Figure 2), pulsars move from the top left in the $P$–$\dot{P}$ plane diagonally toward the bottom right, following lines of constant magnetic field (see, e.g., the right panel of Figure 3). As they do, stars with comparable field strengths but different initial periods evolve toward similar $P$ values. In addition, the misalignment angle evolution introduces further degeneracies because all $\chi$ decrease with time. However, as the field decays, spin-down and misalignment evolution slow down and pulsars begin to evolve almost vertically toward smaller $\dot{P}$ values. These processes depend further on $B_0$ and $P_0$, as stronger initial fields and smaller initial periods result in faster spin-down and faster evolution toward alignment. This is especially visible for test sample 3 (top right panel of Figure 5), which is characterized by the smallest period mean, $\mu_{\log P}$, of all three test cases. The combined action of these effects is that stars born with different rotational properties attain similar $P$ at current times. This information loss on the initial period makes it harder to infer corresponding parameters. As expected, test simulation 3 thus shows the largest 95% CIs for $\mu_{\log P}$ and $\sigma_{\log P}$ out of our three test samples (third column in Table 3 and last row in Figure 10).

### 5.4. Comparing Results with Earlier Works

Contrasting the posterior medians from Equation (30) with the results of earlier population synthesis studies summarized in Table 5 and Figure 14, we first note that our $\mu_{\log B}$ estimate is roughly consistent with Gullón et al. (2014, 2015) but

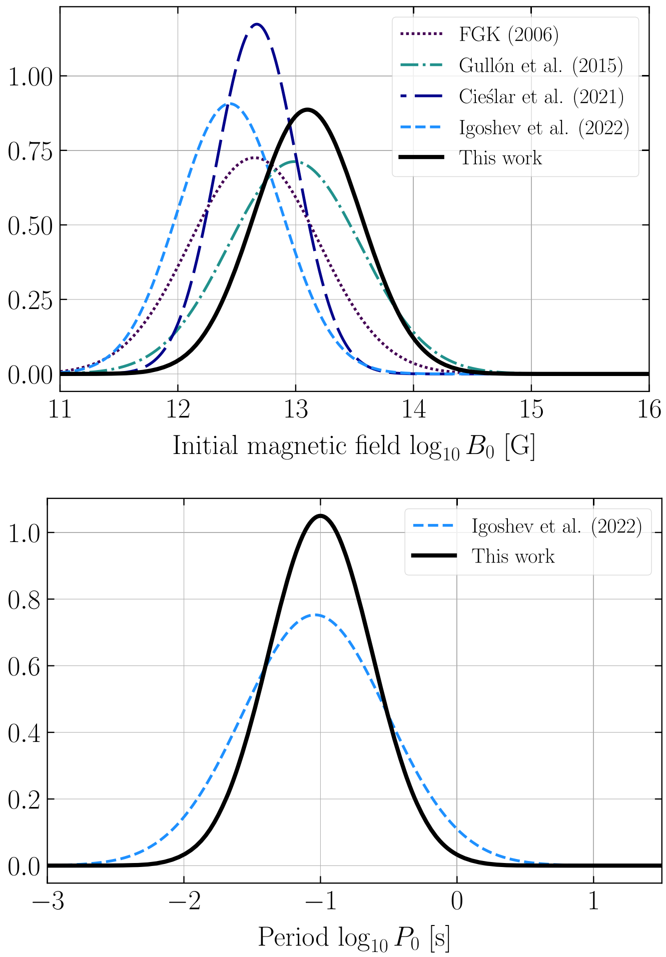

**Figure 14.** A comparison of initial magnetic field, $B_0$ (top), and period, $P_0$ (bottom), distributions for the radio pulsar population. The logarithms of $B_0$ and $P_0$ are normally distributed (see Equations (8) and (9)) and characterized by means, $\mu_{\log B,P}$, and standard deviations, $\sigma_{\log B,P}$, respectively. That is, these distributions are normalized such that the total area under the curves equals 1 for logarithmic abscissa values. Corresponding numerical values are summarized in Table 5. The results of this work are illustrated as black solid lines. Additional studies are shown as detailed in the legends.

somewhat larger than those of Faucher-Giguère & Kaspi (2006), Cieślar et al. (2020), and Igoshev et al. (2022). Moreover, while very close to Igoshev et al. (2022), we obtain a smaller $\sigma_{\log B}$ than Gullón et al. (2014, 2015) and Faucher-Giguère & Kaspi (2006) and a slightly larger estimate than Cieślar et al. (2020). Although these works determine optimal parameter ranges different from ours (see Table 4), we expect the variation in the $B_0$ constraints to be mainly due to our more realistic prescription for the field and the coupled $P$–$\chi$ evolution.

**Table 5**
Comparison between Best Parameters for the Lognormal Initial Magnetic Field and Initial Period Distributions in the Literature

| References | $\mu_{\log B}$ | $\sigma_{\log B}$ | $\mu_{\log P}$ | $\sigma_{\log P}$ |
|---|---|---|---|---|
| Faucher-Giguère & Kaspi (2006) | 12.65 | 0.55 | ⋯ | ⋯ |
| Gullón et al. (2015) | 12.99 | 0.56 | ⋯ | ⋯ |
| Cieślar et al. (2020) | $12.67^{+0.01}_{-0.02}$ | $0.34^{+0.02}_{-0.01}$ | ⋯ | ⋯ |
| Igoshev et al. (2022) | 12.44 | 0.44 | $-1.04^{+0.15}_{-0.20}$ | $0.53^{+0.12}_{-0.08}$ |
| This work | $13.10^{+0.04}_{-0.05}$ | $0.45^{+0.03}_{-0.02}$ | $-1.00^{+0.11}_{-0.10}$ | $0.38^{+0.16}_{-0.10}$ |

**Note.** We provide references and the four relevant parameters. Note that the first three studies use a different prescription for the initial period, which prevents a direct comparison with our study. For Gullón et al. (2015) and Cieślar et al. (2020), we compare with *model D* for the radio pulsar population and the *rotational model*, respectively. The corresponding distributions are illustrated in Figure 14. Where available, we quote CIs at the 68% level (including for this work), but note that these are difficult to compare owing to the difference in inference methods and underlying models and data.

A direct comparison of our initial period parameters and earlier population synthesis literature is not possible because (following recent results by Igoshev et al. 2022; see also Xu et al. 2023) we considered the periods' logarithm and not the periods themselves to be normally distributed. However, we highlight that our inferred $\mu_{\log P}$ is comparable to that of Igoshev et al. (2022), whereas our $\sigma_{\log P}$ is somewhat smaller (see bottom panel of Figure 14). Igoshev et al. (2022) focused on a simplified analysis of 56 young neutron stars in supernova remnants and looked at magnetorotational properties only. The authors were thus able to define an explicit likelihood function and perform statistical inference. Corresponding CIs given in Table 5 are similar to ours, but we highlight that a systematic comparison is complicated owing to the distinct choices of underlying data and inference techniques. In this context, we also point out that although Cieślar et al. (2020) derive relatively narrow posteriors (see Table 5) for a range of pulsar properties using an MCMC analysis, their underlying simulation framework is significantly reduced compared to ours invoking, e.g., (unrealistic) exponential field decay, vacuum magnetospheres, no coupling between periods and misalignment angles, and a simplified prescription for the beamed emission. In addition, they make an explicit assumption on the likelihood that might not accurately capture the complexity of the pulsar population synthesis even for their simplified model. We reiterate the robustness of our SBI approach, which eliminates the need for an explicit expression for the likelihood and is therefore also suitable for more complex simulators like ours. Moreover, as outlined above, the use of a neural density estimator results in amortized posterior distributions that allow fast evaluation and sampling. We used this fact to determine the coverage and validate our posteriors, a procedure that is infeasible in MCMC or nested sampling approaches owing to the time-consuming need for repeated sampling.

### *5.5. Late-time Magnetic Field Decay*

We now turn our attention to the parameter $a_{\mathrm{late}}$, the power-law index for the late-time magnetic field decay. We newly introduced $a_{\mathrm{late}}$ in pulsar population synthesis to account for the highly uncertain, core-dominated field evolution above $10^6$ yr in a phenomenological way. While corresponding inferences were satisfactory for our benchmark experiments, we found that posteriors for $a_{\mathrm{late}}$ inferred from the observed population differed significantly between our 19 experiments, resulting in systematically larger 95% CIs for smaller $a_{\mathrm{late}}$ medians and vice versa (see rightmost panel of Figure 12). In addition, several posteriors did not overlap at all across our prior range, leading to a bimodality in the ensemble posterior. As we did not see anything similar for our synthetic simulations, we do not associate this behavior with the networks' performance or the SBI approach itself. Instead, we hypothesize that this is due to shortcomings in our simulation framework. Put differently, our statistical inferences are only as good as the simulation

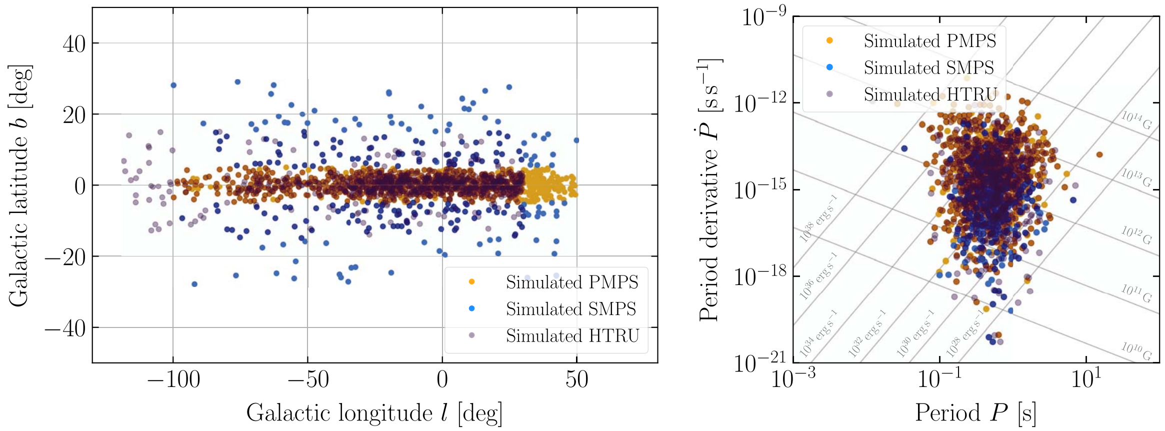

**Figure 15.** Simulated populations of isolated Galactic radio pulsars detected with PMPS, SMPS, and the low- and mid-latitude HTRU survey (highlighted in yellow, light blue, and purple, respectively) for the parameters inferred via SBI from the observed radio pulsar population (see Equation (30)). The left panel shows the distribution of the simulated population in Galactic latitude, $b$, and longitude, $l$, while the right panel depicts the pulsars in the period, $P$, and period derivative, $\dot{P}$, plane. In the latter, we also give lines of constant spin-down power, $|\dot{E}_{\rm rot}|$, and constant dipolar surface magnetic field, $B$ (estimated via Equation (10) for an aligned rotator). Both plots directly compare to the true (observed) population shown in Figure 3.

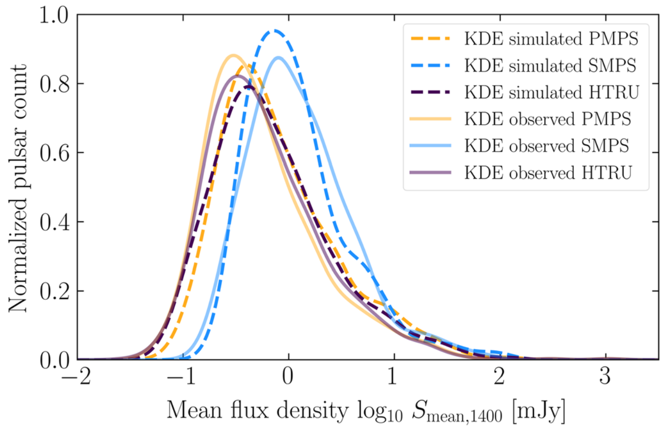

**Figure 16.** Distributions of mean radio flux densities, $S_{\rm mean,1400}$, at 1400 MHz for the populations of isolated Galactic radio pulsars in PMPS, SMPS, and the low- and mid-latitude HTRU survey (in yellow, light blue, and purple, respectively). To avoid overcrowding the plot, we omit the underlying histograms (see Figure 4) and only show the individual probability density functions obtained via KDE using a Gaussian kernel. Estimates for the observed population are shown as solid lines, while one of our best-parameter simulations is shown with dashed lines. Data taken from the ATNF Pulsar Catalogue (Manchester et al. 2005, https://www.atnf.csiro.au/research/pulsar/psrcat/, v1.69).

model used to train our density estimator. Consequently, we see the complications in inferring $a_{\rm late}$ as an indication that our treatment of the late-time field evolution via a power law (albeit physically motivated by the behavior of known magnetic field evolution mechanisms) is insufficient to model the observed pulsar population.

Although further work is needed to better understand the late-time evolution of neutron star fields, we can assure ourselves that our current power-law prescription is not too far from reality. To do so, we rerun our simulator with the best estimates summarized in Equation (30). We show an example of the resulting population in Galactic longitude and latitude and in $P$ and $\dot{P}$ in Figure 15. Both panels are analogous to the respective plots in Figure 3. Moreover, Figure 16 shows a comparison between the estimated probability density functions for the radio flux density distributions for the observed populations (solid lines) and our best-parameter simulation (dashed lines).

While a detailed comparison between this simulated and the observed population and a study of implications for the neutron star birth rate are beyond the scope of this work, we will highlight a few main aspects. First, we note that the distributions look markedly similar, giving a reasonable level of confidence in our underlying simulation framework. This is particularly true for the Galactic longitude vs. latitude distribution and the mean radio flux densities. We attribute the small remaining differences in Figure 16 primarily to uncertainties in the flux density measurements in the ATNF pulsar catalog discussed previously in Section 2.5, our choice of luminosity function (see Equation (16)), and systematics in the determination of pulsar survey sensitivities (see Table 1). Finally, we do see a slight shift in the SMPS population in the $P$–$\dot{P}$ diagram toward lower $\dot{P}$ values. This might again hint at missing physics at late times because SMPS is sensitive to somewhat older pulsars compared to the other two surveys.

### 5.6. Neutron Star Birth Rate

We can further count the numbers of detected pulsars in all three synthetic surveys for our best-estimate simulation. Running our simulator 10 times to account for its stochastic nature, we obtain average pulsar counts of 1013, 242, and 1298 for PMPS, SMPS, and the HTRU survey, respectively. Comparing these to the true observed counts in Equation (23), we find an equivalent number of objects in PMPS (within the sensitivity limits of our iterative approach of generating and detecting pulsars as summarized in Section 2.6), while we overestimate the SMPS population by ∼11% and the HTRU population by ∼27% on average.

To understand these small discrepancies, we return to our earlier discussion of the neutron star birth rate in Section 2.6. In particular, for our best estimates, we reach the observed target counts  given in Equation (23) for each survey for the following birth rates:

$$\begin{aligned} &\text{PMPS:}\ \sim 2.02 \pm 0.02 \text{ neutron stars per century,} \\ &\text{SMPS:}\ \sim 1.84 \pm 0.03 \text{ neutron stars per century,} \\ &\text{HTRU:}\ \sim 1.66 \pm 0.02 \text{ neutron stars per century,} \end{aligned} \tag{31}$$

where we quote means and standard errors for the 10 runs. These estimates are somewhat smaller than those obtained in earlier population synthesis studies (Faucher-Giguère & Kaspi 2006; Gullón et al. 2014) and very close to the recent core-collapse supernova estimate from Rozwadowska et al. (2021; $1.63 \pm 0.46$ per century). The differences in Equation (31) are sufficient to result in the slight overproduction of objects noted above. We remind that this is because we continue producing neutron stars until we hit the number of observed pulsars in all three surveys. In our specific case, PMPS detections require a slightly larger birth rate than the other two surveys. As mentioned previously, the main reason for this is that we only expect the *correct* physical model to produce the same birth rate across all surveys, again hinting that our simulator is missing some physics. Nonetheless, besides successfully constraining magnetorotational parameters for pulsar population synthesis using SBI for the first time, we do recover birth rate results in Equation (31) that are very similar across all surveys.

### *5.7. Future Directions*

In light of the previous conclusions, we intend to further develop our current approach in a number of ways.

On the simulation side, we will investigate additional luminosity prescriptions that go beyond our assumption, $L_{\rm int} \propto |\dot{E}|^{1/2}$, as this is another quantity that can significantly affect the pulsar distribution. Varying the exponent in our simulations, which was beyond the scope of this study owing to computational limitations, but using SBI to constrain corresponding parameter ranges would be a first step in that direction. Moreover, while we followed Gullón et al. (2014, 2015) and took a significant step forward in incorporating a realistic description of the neutron star magnetic field, we already noted above that further investigations into the field evolution of the neutron star core at late times will be important for future population synthesis frameworks. Finally, new pulsar surveys (in the radio band, as well as in other wavelengths) might hold the key to further constraining the neutron star population. While we did not see a significant improvement in our inferences using information from one, two, or three radio surveys, future studies will benefit from larger numbers of detected pulsars and accurate classification of telescope and detection biases. Furthermore, other wave bands, specifically X-rays or gamma rays, provide complementary information on the neutron star population. Our focus on realistic magnetic field evolution and the expansion of our approach to new three-dimensional magnetothermal simulations (e.g., De Grandis et al. 2021; Dehman et al. 2023) will be particularly crucial to determine realistic X-ray luminosities of the most strongly magnetized neutron stars. As highlighted by Gullón et al. (2015), modeling these so-called magnetars and the isolated radio pulsar population consistently will be crucial to break degeneracies and constrain neutron star physics further.

The increase in simulator complexity associated with these improvements will not only result in more free parameters but also inevitably lead to larger computation times for our forward model. The approach taken here, i.e., simulating a large database for input parameter combinations that cover the entire space sufficiently, will become infeasible. To overcome these hurdles, we will also have to explore new SBI approaches. Sequential methods (e.g., Papamakarios et al. 2018; Deistler et al. 2022; Bhardwaj et al. 2023) that reduce the need for simulations by starting from a relatively small database and adaptively providing additional simulations (generated for those parts of the parameter space that are most useful for a neural density estimator to learn a posterior approximation) seem particularly suited to these tasks.

## Acknowledgments

The authors thank Emilie Parent for useful exchanges on radio pulsar emission and detections, Clara Dehman for providing magnetic field evolution curves, and Jose Pons for insights on late-time magnetic field evolution. V.G., M.R., and C.P.A. also thank Jakob Macke and his group for valuable discussions on SBI. The data production, processing, and analysis tools for this paper have been implemented and operated at the Port d'Informació Científica (PIC) data center. PIC is maintained through a collaboration of the Institut de Física d'Altes Energies (IFAE) and the Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (Ciemat). We particularly thank Christian Neissner and Carles Acosta for their support at PIC. V.G. acknowledges support from a Juan de la Cierva Incorporación Fellowship. The authors are further supported by the ERC via the Consolidator grant "MAGNESIA" (No. 817661) and by the program Unidad de Excelencia María de Maeztu CEX2020-001058-M. We also acknowledge partial support from grant SGR2021-01269 (PI: Graber). M.R.'s and C.P.A.'s work has been carried out within the framework of the doctoral program in Physics at the Universitat Autonoma de Barcelona.

*Software:* Astropy (Astropy Collaboration et al. 2013, 2018), healpy (Górski et al. 2005; Zonca et al. 2019), IPython (Perez & Granger 2007), JupyterLab, Matplotlib (Hunter 2007), Numba (Lam et al. 2015), NumPy (Oliphant 2006; van der Walt et al. 2011; Harris et al. 2020), Pandas (McKinney 2010), PyGEDM, PyTorch (Paszke et al. 2019), sbi (Tejero-Cantero et al. 2020), SciPy (Jones et al. 2020; Virtanen et al. 2020), Sphinx.

## Appendix A
## Magnetic Field Prescription

As outlined in Section 2.3, a key ingredient for the magnetorotational evolution of radio pulsars is a realistic prescription for the evolution of the dipolar magnetic field strength, $B$, up to neutron star ages of $10^8$ yr. While earlier population synthesis studies have typically either neglected magnetic field decay entirely or relied on simplified descriptions invoking decaying exponentials or power laws, we choose a different approach and take advantage of recent progress in modeling the magnetothermal evolution of neutron star crusts. In particular, we use a set of five two-dimensional simulations (Viganò et al. 2021) to fit the early-time magnetic field evolution, which is driven by the combined action of the Hall effect and ohmic dissipation (see, e.g., Pons & Viganò 2019, for details on these mechanisms).

All five curves, shown as solid lines in Figure 2, were simulated with realistic assumptions on relevant physics. In particular, the stellar structure and composition are based on the equation of state SLy4 (Douchin & Haensel 2001) for a neutron star of mass $1.4 M_{\odot}$, resulting in a radius of 11.74 km. The impurity parameter at the highest densities in the inner crust is set to 100 (Pons et al. 2013), representing the presence of resistive nuclear pasta phases (see, e.g., Chamel & Haensel 2008), whereas the impurity profile for other crustal densities matches the results of Carreau et al. (2020, see their Figure 5). Furthermore, the model for the neutron star envelope is taken from Potekhin et al. (2015), while specific parameterization for the superfluid and superconducting energy gaps (SFB for the crustal neutrons, TToa for the core neutrons, and CCDKp for the core protons) were adopted from Ho et al. (2015).

What varies between the different simulations is the initial poloidal magnetic field strength, $B$, taking the values $10^{12}$, $10^{13}$, $10^{14}$, $10^{15}$, and $5 \times 10^{15}$ G, respectively. This also implies different toroidal field strengths, which are typically a factor 10 larger than the poloidal $B$ values. We observe in Figure 2 that those runs with larger magnetic fields decay faster. This is a direct result of the Hall effect, which depends on $B$ and acts to redistribute the magnetic field energy to smaller scales, where it subsequently decays owing to ohmic dissipation. For sources with $B \lesssim 10^{12}$ G and coupled thermal evolution, this Hall cascade does not take place and magnetic fields remain pretty much constant on timescales of the order of $10^6$ yr.

Above this timescale, however, current magnetothermal simulations become unreliable because the implementation of relevant microphysics (Potekhin et al. 2015) is unsuited to old neutron stars with temperatures $\lesssim 10^6$ K. In addition, these simulations focus primarily on the crust and do not include a realistic treatment of the highly uncertain dynamics of the neutron star core, which should become relevant above $\sim 10^6$ yr. As we require a prescription for the field above $10^6$ yr for our population synthesis, we develop a simplified parameterization for the late-time magnetic field evolution that encodes the unknown evolution of the stellar core. As highlighted in Equation (12), we assume that field changes at late times can be captured by a power law characterized by the index, $a_{\rm late}$. This choice is physically motivated because several known magnetic field evolution mechanisms exhibit the same functional form. For example, Hall-like physics are encoded by $a_{\rm late} = -1$ (Aguilera et al. 2008), while ambipolar diffusion follows a power law with $a_{\rm late} = -0.5$ (Goldreich & Reisenegger 1992).

To directly parameterize the behavior of the magnetic field across all relevant $B$ ranges and times $t$, we describe the field evolution with the following broken power laws:

$$B(t) = B_0\left(1 + \frac{t}{\tau_1}\right)^{a_1}\left(1 + \frac{t}{\tau_2}\right)^{a_2 - a_1}\left(1 + \frac{t}{\tau_{\rm late}}\right)^{a_{\rm late} - a_2} \quad \text{for} \quad \tau_1 < \tau_2 < \tau_{\rm late}, \tag{A1}$$

$$B(t) = B_0\left(1 + \frac{t}{\tau_1}\right)^{a_1}\left(1 + \frac{t}{\tau_{\rm late}}\right)^{a_{\rm late} - a_1} \quad \text{for} \quad \tau_1 < \tau_{\rm late} < \tau_2, \tag{A2}$$

$$B(t) = B_0\left(1 + \frac{t}{\tau_{\rm late}}\right)^{a_{\rm late}} \quad \text{for} \quad \tau_{\rm late} < \tau_1 < \tau_2. \tag{A3}$$

Here the two timescales $\tau_1 \equiv A_1 B_0^{b_1}$ and $\tau_2 \equiv A_2 B_0^{b_2}$ depend on the initial magnetic field, $B_0$, while $\tau_{\rm late}$ is a constant. The latter, together with the free parameters $A_{1,2}$ and $b_{1,2}$ and the power-law indices $a_{1,2}$, can be adjusted to closely fit the numerical simulations. Measuring all three timescales in years and $B_0$ in gauss, we then choose $\tau_{\rm late} = 2 \times 10^6$ yr, $A_1 = 10^{14}$yr G$^{-b_1}$, $b_1 = -0.8$, $A_2 = 6 \times 10^8$yr G$^{-b_2}$, $b_2 = -0.2$, $a_1 = -0.13$, and $a_2 = -3.0$.

For particularly steep power-law indices, $a_{\rm late}$, the current prescription, in principle, allows the magnetic field to decay to unrealistically small values in contrast with observations of old millisecond pulsars (Lorimer 2008). To prevent this, we assume that the magnetic field eventually settles at a constant value, $B_{\rm late}$, for very late times. In line with detected old neutron stars, we randomly sample the logarithm of $B_{\rm late}$ from a normal distribution with a mean $\mu_{\log B,\rm final} = 8.5$ and a standard deviation $\sigma_{\log B,\rm final} = 0.5$ as already outlined previously. The result of this magnetic field prescription for $a_{\rm late} = -3.0$ is shown as the dashed lines in Figure 2.

## Appendix B
## Coverage Calculation

To validate our neural posterior estimates, we follow Cook et al. (2006), who demonstrated that for a well-calibrated posterior distribution the smallest volume that contains the ground truth, $\boldsymbol{\theta}$, for a given sample in a test data set follows a uniform distribution. This, in turn, implies that the cumulative distribution function of these quantiles across the entire test set forms a diagonal line. The graphical representation of this cumulative distribution function is commonly referred to as the *coverage plot* (see Figure 11). Put differently, if we consider a credibility level $1 - \alpha$, we expect the ground truth, $\boldsymbol{\theta}$, to fall into this region for a fraction $1 - \alpha$ of test samples if the coverage is diagonal.

To calculate the corresponding coverage for our posteriors and assess how well they are calibrated, we take advantage of the amortized nature of our approximate posterior. In particular, for each of our 3600 test samples, we have access to the ground truth, $\boldsymbol{\theta}$, and the corresponding posterior approximation, $q_{F(\boldsymbol{x},\phi)}(\boldsymbol{\theta})$, where $F(\boldsymbol{x}, \phi)$ represents a trained neural network. To determine the coverage, we need to calculate the quantiles for each $\boldsymbol{\theta}$. In our case, where we infer five magnetorotational parameters and the posterior, $q_{F(\boldsymbol{x},\phi)}(\boldsymbol{\theta})$, is a five-dimensional probability density function (see Equation (28)), we obtain corresponding quantiles by determining the so-called highest-density regions (HDRs), i.e., those regions covering our sample space for a given probability $1 - \alpha$ that have the smallest possible volume (Hyndman 1996). To obtain these HDRs for each of our test samples, we first compute the total log-posterior at the ground truth, $\boldsymbol{\theta}$, i.e., $\log q_{F(\boldsymbol{x},\phi)}(\boldsymbol{\theta})$. From each posterior, we subsequently draw samples, $\boldsymbol{\theta}_s$, with $s \in \{1,...,S\}$, for which we also individually compute the log-posterior, i.e., $\log q_{F(\boldsymbol{x},\phi)}(\boldsymbol{\theta}_s)$. The HDR for a given test sample with ground truth, $\boldsymbol{\theta}$, is now the percentage of samples, $\boldsymbol{\theta}_s$, that satisfy the condition $\log q_{F(\mathbf{x},\phi)}(\boldsymbol{\theta}_s) > \log q_{F(\mathbf{x},\phi)}(\boldsymbol{\theta})$. To compute the cumulative distribution function (coverage) across our test set, we repeat this process iteratively for all 3600 test samples to determine, for a given credibility level $1-\alpha$, the fraction of test samples where the HDR is smaller than or equal to $1-\alpha$.

Deviations from the diagonal are present when posterior estimates are either too wide (conservative) or too narrow (overconfident). In the former case, ground truths would be enclosed within a given HDR more often than expected for the true posterior, while in the latter scenario the opposite applies. The resulting coverage curves would, thus, lie above and below the diagonal, respectively, highlighting the benefit of the coverage plot in validating our posteriors.

Finally, note that for our ensemble approach we calculate the HDR with the ensemble posterior, $\bar{q}(\boldsymbol{\theta})$, using the condition $\log \bar{q}(\boldsymbol{\theta}_s) > \log \bar{q}(\boldsymbol{\theta})$. The remaining steps are identical to those outlined above.

## ORCID iDs

Vanessa Graber https://orcid.org/0000-0002-6558-1681
Michele Ronchi https://orcid.org/0000-0003-2781-9107
Celsa Pardo-Araujo https://orcid.org/0000-0002-8118-255X
Nanda Rea https://orcid.org/0000-0003-2177-6388

## References

Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017, PhRvL, 119, 161101
Aguilera, D. N., Pons, J. A., & Miralles, J. A. 2008, A&A, 486, 255
Alsing, J., Charnock, T., Feeney, S., & Wandelt, B. 2019, MNRAS, 488, 4440
Ashton, G., Hübner, M., Lasky, P. D., et al. 2019, ApJS, 241, 27
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33
Bachetti, M., Harrison, F. A., Walton, D. J., et al. 2014, Natur, 514, 202
Bates, S. D., Lorimer, D. R., Rane, A., & Swiggum, J. 2014, MNRAS, 439, 2893
Beaumont, M. A., Zhang, W., & Balding, D. J. 2002, Genetics, 162, 2025
Berger, E. 2014, ARA&A, 52, 43
Bhardwaj, U., Alvey, J., Miller, B. K., Nissanke, S., & Weniger, C. 2023, PhRvD, 108, 042004
Bhattacharya, D., Wijers, R. A. M. J., Hartman, J. W., & Verbunt, F. 1992, A&A, 254, 198
Bovy, J. 2015, ApJS, 216, 29
Carreau, T., Fantina, A. F., & Gulminelli, F. 2020, A&A, 640, A77
Chamel, N., & Haensel, P. 2008, LRR, 11, 10
Chen, K., & Ruderman, M. 1993, ApJ, 402, 264
Cheung, D. H. T., Wong, K. W. K., Hannuksela, O. A., Li, T. G. F., & Ho, S. 2022, PhRvD, 106, 083014
Cieślar, M., Bulik, T., & Osłowski, S. 2020, MNRAS, 492, 4043
Coleman, M. S. B., & Burrows, A. 2022, MNRAS, 517, 3938
Conroy, C., Weinberg, D. H., Naidu, R. P., et al. 2022, arXiv:2204.02989
Cook, S., Gelman, A., & Rubin, D. 2006, JCGS, 15, 675
Cordes, J. M., & McLaughlin, M. A. 2003, ApJ, 596, 1142
Cranmer, K., Brehmer, J., & Louppe, G. 2020, PNAS, 117, 30055
Dax, M., Green, S. R., Gair, J., et al. 2021, PhRvL, 127, 241103
De Grandis, D., Taverna, R., Turolla, R., et al. 2021, ApJ, 914, 118
Dean, T. A., Singh, S. S., Jasra, A., & Peters, G. W. 2011, arXiv:1103.5399
Dehman, C., Viganò, D., Pons, J. A., & Rea, N. 2023, MNRAS, 518, 1222
Deistler, M., Goncalves, P. J., & Macke, J. H. 2022, in Advances in Neural Information Processing Systems, ed. S. Koyejo et al., Vol. 35 (New York: Curran Associates, Inc.), 23135
Douchin, F., & Haensel, P. 2001, A&A, 380, 151
Edwards, R. T., Bailes, M., Van Straten, W., & Britton, M. C. 2001, MNRAS, 326, 358
Faucher-Giguère, C.-A., & Kaspi, V. M. 2006, ApJ, 643, 332
Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Frazier, D. T., Maneesoonthorn, W., Martin, G. M., & McCabe, B. P. M. 2017, arXiv:1712.07750
Gangadhara, R. T., & Gupta, Y. 2001, ApJ, 555, 31
Glorot, X., & Bengio, Y. 2010, in PMLR Vol. 9, Proc. 13th Int. Conf. on Artificial Intelligence and Statistics, ed. Y. W. Teh & M. Titterington, 249
Goldreich, P., & Reisenegger, A. 1992, ApJ, 395, 250
Gonthier, P. L., Story, S. A., Clow, B. D., & Harding, A. K. 2007, Ap&SS, 309, 245
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Greenberg, D. S., Nonnenmacher, M., & Macke, J. H. 2019, arXiv:1905.07488
Gullón, M., Miralles, J. A., Viganò, D., & Pons, J. A. 2014, MNRAS, 443, 1891
Gullón, M., Pons, J. A., Miralles, J. A., et al. 2015, MNRAS, 454, 615
Hahn, C., Lemos, P., Parker, L., et al. 2023, arXiv:2310.15246
Harris, C. R., Jarrod Millman, K., van der Walt, S. J., et al. 2020, Natur, 585, 357
Haslam, C. G. T., Klein, U., Salter, C. J., et al. 1981, A&A, 100, 209
Haslam, C. G. T., Salter, C. J., Stoffel, H., & Wilson, W. E. 1982, A&AS, 47, 1
He, K., Zhang, X., Ren, S., & Sun, J. 2015, arXiv:1502.01852
Hermans, J., Begy, V., & Louppe, G. 2019, arXiv:1903.04057
Hermans, J., Delaunoy, A., Rozet, F., et al. 2021, arXiv:2110.06581
Hernquist, L. 1990, ApJ, 356, 359
Hewish, A., Bell, S. J., Pilkington, J. D. H., Scott, P. F., & Collins, R. A. 1968, Natur, 217, 709
Ho, W. C. G., Elshamouty, K. G., Heinke, C. O., & Potekhin, A. Y. 2015, PhRvC, 91, 015806
Hobbs, G., Lorimer, D. R., Lyne, A. G., & Kramer, M. 2005, MNRAS, 360, 974
Hunter, J. D. 2007, CSE, 9, 90
Huppenkothen, D., & Bachetti, M. 2022, MNRAS, 511, 5689
Hyndman, R. J. 1996, TAS, 50, 120
Igoshev, A. P. 2020, MNRAS, 494, 3663
Igoshev, A. P., Frantsuzova, A., Gourgouliatos, K. N., et al. 2022, MNRAS, 514, 4606
Igoshev, A. P., Gourgouliatos, K. N., Hollerbach, R., & Wood, T. S. 2021, ApJ, 909, 101
Jacoby, B. A., Bailes, M., Ord, S. M., Edwards, R. T., & Kulkarni, S. R. 2009, ApJ, 699, 2009
Janka, H.-T., Wongwathanarat, A., & Kramer, M. 2022, ApJ, 926, 9
Jankowski, F., van Straten, W., Keane, E. F., et al. 2018, MNRAS, 473, 4436
Johnston, S., Lyne, A. G., Manchester, R. N., et al. 1992, MNRAS, 255, 401
Johnston, S., Smith, D. A., Karastergiou, A., & Kramer, M. 2020, MNRAS, 497, 1957
Jones, S., Oliphant, T. E., Peterson, P., et al. 2001, SciPy: Open source scientific tools for Python, http://www.scipy.org/
Keane, E. F., & Kramer, M. 2008, MNRAS, 391, 2009
Keith, M. J., Jameson, A., van Straten, W., et al. 2010, MNRAS, 409, 619
Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980
Kramer, M., Wielebinski, R., Jessner, A., Gil, J. A., & Seiradakis, J. H. 1994, A&AS, 107, 515
Krishnakumar, M. A., Mitra, D., Naidu, A., Joshi, B. C., & Manoharan, P. K. 2015, ApJ, 804, 23
Kullback, S., & Leibler, R. A. 1951, Ann. Math. Stat., 22, 79
Lam, S. K., Pitrou, A., & Seibert, S. 2015, in Proc. 2nd Workshop on the LLVM Compiler Infrastructure in HPC (New York: ACM), 1
Lawson, K. D., Mayer, C. J., Osborne, J. L., & Parkinson, M. L. 1987, MNRAS, 225, 307
Lemos, P., Cranmer, M., Abidi, M., et al. 2023, MLS&T, 4, 01LT01
Li, C., Zhao, G., Jia, Y., et al. 2019, ApJ, 871, 208
Lin, K., von wietersheim-Kramsta, M., Joachimi, B., & Feeney, S. 2023, MNRAS, 524, 6167
Lorimer, D. R. 2004, in IAU Symp. 218, Young Neutron Stars and Their Environments, ed. F. Camilo & B. M. Gaensler (San Francisco, CA: ASP), 105
Lorimer, D. R. 2008, LRR, 11, 8
Lorimer, D. R., Bailes, M., Dewey, R. J., & Harrison, P. A. 1993, MNRAS, 263, 403
Lorimer, D. R., Faulkner, A. J., Lyne, A. G., et al. 2006, MNRAS, 372, 777
Lorimer, D. R., & Kramer, M. 2012, Handbook of Pulsar Astronomy (Cambridge: Cambridge Univ. Press)
Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., et al. 2017, arXiv:1711.01861
Maciesiak, K., & Gil, J. 2011, MNRAS, 417, 1444
Maciesiak, K., Gil, J., & Ribeiro, V. A. R. M. 2011, MNRAS, 414, 1314
Manchester, R. N., Hobbs, G. B., Teoh, A., & Hobbs, M. 2005, AJ, 129, 1993
Manchester, R. N., Lyne, A. G., Camilo, F., et al. 2001, MNRAS, 328, 17

Marchetti, T., Rossi, E. M., & Brown, A. G. A. 2019, MNRAS, 490, 157

Margalit, B., Metzger, B. D., Berger, E., et al. 2018, MNRAS, 481, 2407

McKinney, W. 2010, in Proceedings of the 9th Python in Science Conference, ed. S. van der Walt & J. Millman, http://conference.scipy.org/proceedings/scipy2010/mckinney.html

Metzger, B. D., Vurm, I., Hascoët, R., & Beloborodov, A. M. 2014, MNRAS, 437, 703

Miller, B., Cole, A., Forré, P., Louppe, G., & Weniger, C. 2021, Advances in Neural Information Processing Systems, 34 (New York: Curran Associates, Inc.), 129

Mishra-Sharma, S., & Cranmer, K. 2022, PhRvD, 105, 063017

Miyamoto, M., & Nagai, R. 1975, PASJ, 27, 533

Narayan, R., & Ostriker, J. P. 1990, ApJ, 352, 222

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563

Oliphant, T. E. 2006, A guide to NumPy (USA: Trelgol Publishing)

Papamakarios, G., & Murray, I. 2016, arXiv:1605.06376

Papamakarios, G., Sterratt, D. C., & Murray, I. 2018, arXiv:1805.07226

Paszke, A., Gross, S., & Massa, F. 2019, arXiv:1912.01703

Perez, F., & Granger, B. E. 2007, CSE, 9, 21

Petroff, E., Hessels, J. W. T., & Lorimer, D. R. 2022, A&ARv, 30, 2

Philippov, A., Tchekhovskoy, A., & Li, J. G. 2014, MNRAS, 441, 1879

Pichardo, B., Moreno, E., Allen, C., et al. 2012, AJ, 143, 73

Pons, J. A., & Viganò, D. 2019, LRCA, 5, 3

Pons, J. A., Viganò, D., & Rea, N. 2013, NatPh, 9, 431

Popov, S. B., Pons, J. A., Miralles, J. A., Boldin, P. A., & Posselt, B. 2010, MNRAS, 401, 2675

Posselt, B., Karastergiou, A., Johnston, S., et al. 2023, MNRAS, 520, 4582

Potekhin, A. Y., Pons, J. A., & Page, D. 2015, SSRv, 191, 239

Remazeilles, M., Dickinson, C., Banday, A. J., Bigot-Sazy, M. A., & Ghosh, T. 2015, MNRAS, 451, 4311

Ronchi, M., Graber, V., Garcia-Garcia, A., Rea, N., & Pons, J. A. 2021, ApJ, 916, 100

Rozwadowska, K., Vissani, F., & Cappellaro, E. 2021, NewA, 83, 101498

Rubin, D. B. 1984, AnSta, 12, 1151

Rudak, B., & Ritter, H. 1994, MNRAS, 267, 513

Ruderman, M. A., & Sutherland, P. G. 1975, ApJ, 196, 51

Sharma, S. 2017, ARA&A, 55, 213

Skowron, D. M., Skowron, J., Mróz, P., et al. 2019, Sci, 365, 478

Skrzypczak, A., Basu, R., Mitra, D., et al. 2018, ApJ, 854, 162

Speagle, J. S. 2020, MNRAS, 493, 3132

Spitkovsky, A. 2006, ApJL, 648, L51

Tauris, T. M., & Manchester, R. N. 1998, MNRAS, 298, 625

Tejero-Cantero, A., Boelts, J., Deistler, M., et al. 2020, JOSS, 5, 2505

Vallée, J. P. 2017, AstRv, 13, 113

van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, CSE, 13, 22

Vasist, M., Rozet, F., Absil, O., et al. 2023, A&A, 672, A147

Verbunt, F., Igoshev, A., & Cator, E. 2017, A&A, 608, A57

Viganò, D., Garcia-Garcia, A., Pons, J. A., Dehman, C., & Graber, V. 2021, CoPhC, 265, 108001

Viganò, D., Rea, N., Pons, J. A., et al. 2013, MNRAS, 434, 123

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

Wainscoat, R. J., Cohen, M., Volk, K., Walker, H. J., & Schwartz, D. E. 1992, ApJS, 83, 111

Xiang, M., & Rix, H.-W. 2022, Natur, 603, 599

Xu, K., Yang, H.-R., Mao, Y.-H., et al. 2023, ApJ, 947, 76

Yao, J. M., Manchester, R. N., & Wang, N. 2017, ApJ, 835, 29

Yusifov, I., & Küçük, I. 2004, A&A, 422, 545

Zhang, B., Harding, A. K., & Muslimov, A. G. 2000, ApJL, 531, L135

Zonca, A., Singer, L., Lenz, D., et al. 2019, JOSS, 4, 1298