Global Journal of Multidisciplinary Research
Volume 3, Issue 1, January–June 2026 | E-ISSN: 2476-1540
Original Research Article | DOI: 10.5281/zenodo.20302176

Received: March 10, 2026 | Revised: March 12, 2026 | Accepted: March 15, 2026 | Published: March 18, 2026

Bayesian Leave-One-Out Cross-Validation for Astrophysical Model Comparison Using Gravitational-Wave Background Observations

Shreyas Tiruvaskar [1], **Chris Gordon** [2]

[1,2] *School of Physical and Chemical Sciences, University of Canterbury, Christchurch, New Zealand*

Corresponding author: Shreyas Tiruvaskar

Abstract

Prior investigations demonstrated that ultralight dark matter solitons are capable of generating dynamical friction acting on supermassive black-hole binaries, thereby attenuating low-frequency power in the pulsar-timing-array gravitational-wave background and placing constraints on both the particle mass and the effective ultralight-dark-matter abundance. The present study extends that framework by evaluating the predictive performance of four distinct models: a simplified and a physically detailed ultralight-dark-matter formulation, a phenomenological environmental-hardening description, and a gravitational-wave-driven-only model. Predictive performance is assessed through Bayesian leave-one-out cross-validation applied to the five lowest pulsar-timing-array frequency bins. The phenomenological model yields the highest expected log predictive density; however, its margin over the remaining models is modest relative to the associated standard errors. Consequently, the available data do not decisively favor any single model. The most unambiguous pairwise outcome arises within the ultralight-dark-matter framework itself: the simplified formulation surpasses the physically detailed implementation across all five frequency bins. Present pulsar-timing-array observations are therefore consistent with ultralight-dark-matter-driven low-frequency suppression, yet they do not sufficiently discriminate ultralight dark matter from broader environmental characterizations of supermassive-black-hole-binary evolution.

Keywords: Pulsar Timing Arrays, Gravitational-Wave Background, Ultralight Dark Matter, Bayesian Model Comparison, Leave-One-Out Cross-Validation, Supermassive Black-Hole Binaries

1. Introduction

Pulsar timing arrays (PTAs) have established an unprecedented observational channel for detecting nanohertz-frequency gravitational waves [1–5]. The stochastic gravitational-wave background (GWB) recorded by PTAs is broadly consistent with predictions for a cosmological ensemble of inspiralling supermassive black-hole binaries (SMBHBs) [6, 7]. Nevertheless, the precise spectral morphology, especially at the lowest PTA frequencies, is sensitive to the astrophysical mechanisms governing SMBHB orbital evolution prior to the onset of gravitational-wave-dominated inspiral. These mechanisms are therefore directly pertinent to the interpretation of PTA data and to the longstanding final-parsec problem [8, 9].

In Ref. [10], the possibility that ultralight dark matter (ULDM) might serve as a supplementary source of dynamical friction for SMBHBs was systematically explored. In that analysis, ULDM halos hosting central solitonic cores were employed to alter the binary hardening rate, and the resulting characteristic strain spectra were compared against the NANOGrav 15 yr GWB dataset. Two distinct ULDM parameterizations were investigated: a simplified model relying on an analytic soliton profile, and a more realistic prescription in which the soliton geometry is perturbed by the gravitational influence of the SMBHB. The study demonstrated that ULDM-induced drag can suppress the low-frequency portion of the GWB spectrum and yields tractable constraints on the ULDM particle mass and its fractional contribution to the total dark matter density.

Nevertheless, Ref. [10] was fundamentally a parameter-estimation investigation. It addressed which regions of ULDM parameter space are consistent with PTA observations under the adopted modeling assumptions, but it did not conduct a rigorous predictive comparison between the ULDM models and alternative descriptions of the SMBHB population. In particular, it left unquantified whether the simplified ULDM formulation, the more realistic ULDM description, a phenomenological environmental model, or a gravitational-wave-only model is preferred in out-of-sample predictive accuracy. This distinction carries practical significance: a model may produce plausible parameter constraints and visually acceptable spectra without being demonstrably favored over competing descriptions once predictive uncertainty is properly accounted for.

The objective of the present work is to address this missing comparison. Bayesian leave-one-out cross-validation (LOO-CV) is employed to evaluate the predictive performance of the ULDM models from Ref. [10] relative to two benchmark SMBHB models, a phenomenological environmental model and a gravitational-wave-only model, following the modeling framework adopted in the NANOGrav 15 yr SMBHB analysis [7]. The phenomenological model provides a flexible characterization of environmental influences on binary hardening, while the gravitational-wave-only model constitutes the limiting scenario in which binary orbital decay is driven exclusively by gravitational-wave emission.

Bayesian LOO-CV offers a direct route to predictive accuracy assessment. For each observation, the model is refitted after excluding that data point, and the predictive density of the withheld point is evaluated under the corresponding leave-one-out posterior. Aggregating these pointwise contributions yields the expected log predictive density (ELPD), with higher values indicating superior predictive performance [11, 12]. In this analysis, the LOO units correspond to the five lowest PTA frequency bins entering the likelihood. The limited number of LOO terms necessitates cautious interpretation of the resulting standard errors; nevertheless, the procedure provides a quantitative and transparent comparison of predictive performance on the available data.

A computationally efficient approximation to exact LOO-CV is the Pareto-smoothed importance-sampling LOO estimator (PSIS-LOO), which recycles posterior samples from the full model fit [12, 13]. However, the Pareto-$\hat{k}$ diagnostics obtained below reveal that this approximation is unreliable for certain

frequency-bin omissions in the present setting. Exact LOO-CV is therefore adopted as the primary analysis. For each model and each of the five PTA frequency bins, the model is refit with that bin excluded from the likelihood, and the predictive density of the withheld bin is evaluated under the leave-one-out posterior. This avoids any reliance on importance-sampling weights in cases where a held-out frequency bin is influential.

The principal finding is that the phenomenological model achieves the largest total exact-LOO ELPD among all models examined. However, the margins separating this model from the others are not substantial relative to the estimated standard errors, so the present data do not yield compelling evidence for a statistically significant predictive preference among all four models. The clearest pairwise result is the comparison between the two ULDM formulations: the simplified ULDM model exhibits a larger pointwise predictive contribution than the realistic counterpart across all five frequency bins. This provides a quantitative extension of Ref. [10], demonstrating that the two ULDM implementations differ not only in their parameter constraints but also in their predictive performance for the observed GWB spectrum.

The remainder of the paper is organized as follows. Section II summarizes the astrophysical models employed in the comparison. Section III describes the exact and approximate LOO-CV methodology. Section IV presents the model comparison results, encompassing the pointwise ELPD contributions, the PSIS-LOO diagnostics, and the pairwise comparison between the two ULDM formulations. Section V offers concluding remarks.

2. Astrophysical Models

We compare four models for the stochastic gravitational-wave background arising from a cosmological population of SMBHBs. The models differ solely in the physical prescription adopted for binary hardening. Two are the ULDM models introduced in Ref. [10], and the remaining two are benchmark SMBHB models drawn from the NANOGrav 15 yr analysis [7].

2.1 Ultralight Dark Matter Models

ULDM is composed of extremely light bosonic particles and can aggregate into solitonic cores at the centers of dark-matter halos. An SMBHB traversing such a core experiences dynamical friction that extracts orbital energy from the binary, modifying its residence time within the PTA frequency band and consequently altering the predicted GWB spectrum.

Two ULDM implementations from Ref. [10] are considered. The first, designated "ULDM simplified", adopts a streamlined prescription for the dynamical friction force. It reproduces the leading-order scaling of the ULDM drag with particle mass, soliton density, and binary parameters, while neglecting any deformation of the soliton induced by the binary's gravitational field.

The second, designated "ULDM realistic", employs soliton profiles that have been modified by the gravitational influence of the SMBHB. This soliton compression elevates the central density and can alter the hardening rate relative to the simplified prescription. The construction of these modified profiles and their incorporation into the SMBHB strain calculation are detailed in Ref. [10], building on the ULDM soliton–binary dynamics developed in Ref. [14]. Both ULDM models share identical parameter classes: SMBHB population parameters, black-hole–galaxy scaling-relation parameters, and ULDM parameters specifying the particle mass and effective ULDM fraction.

2.2 Phenomenological Environmental Model

The phenomenological model provides a flexible description of environmental hardening without invoking any specific physical mechanism. The effect of the environment on binary orbital evolution is parametrized through a double-power-law functional form, following the implementation of Ref. [7]. This model is designated "Phenom".

This model serves as a generic environmental benchmark. Unlike the ULDM models, it is not tied to a particular dark-matter scenario; instead, it can represent a broad family of low-frequency spectral modifications driven by non-gravitational-wave hardening mechanisms.

2.3 Gravitational-Wave-Only Model

The gravitational-wave-only model posits that SMBHB evolution is driven exclusively by gravitational-wave emission, with no contribution from stellar, gaseous, dark-matter, or other environmental hardening. This model is designated "GW Only" and provides the baseline against which all environmental models are assessed. Differences between GW Only and the other models quantify the degree to which PTA data favor supplementary hardening or low-frequency spectral modification.

2.4 Model Priors

Following Refs. [7, 10], the astrophysical parameters governing binary population density and the black-hole–stellar-mass relations are assigned normal prior distributions. The ULDM particle mass and effective ULDM fraction are assigned log-uniform priors. Parameters of the double-power-law in the Phenom model carry uniform priors. The exact prior specifications for the ULDM models are provided in Table 1 of Ref. [10], and those for the GW Only and Phenom models appear in Table B1 (Astrophysical Priors) of Ref. [7].

3. Statistical Methods

3.1 Exact Leave-One-Out Cross-Validation

Predictive performance is assessed via exact Bayesian LOO-CV. In this procedure, one data point is withheld, the model is refit to the remaining observations, and the predictive density for the omitted point is evaluated. Iterating over all data points yields a pointwise measure of out-of-sample predictive accuracy [11, 12].

Let $y = (y1, ..., yn)$ denote the observed data and let θ denote the model parameters. Given a prior distribution $p(\theta)$ and likelihood $p(y \mid \theta)$, the posterior distribution is:

$$p(\theta \mid y) = p(\theta)\, p(y \mid \theta) / p(y) \propto p(\theta)\, p(y \mid \theta) \quad (1)$$

where $p(y)$ is the marginal likelihood:

$$p(y) = \int p(\theta)\, p(y \mid \theta)\, d\theta \quad (2)$$

Although the marginal likelihood in Eq. (2) can be used to form Bayes factors or posterior odds ratios, it is not adopted as the primary model comparison statistic here. The marginal evidence is sensitive to the prior volume assigned to parameters that are only weakly constrained by the data (cf. Sec. 7.4 of Ref.

[11] and Sec. 2 of Ref. [15]). This concern is directly relevant here because certain astrophysical and ULDM parameters are partly prior-limited, in particular, the effective ULDM fraction remains only weakly constrained [10]. An odds ratio would therefore conflate physical model performance with the specific prior volumes assigned to poorly constrained parameters.

For these reasons, predictive accuracy quantified by Bayesian LOO-CV is adopted as the primary comparison criterion. This approach asks how well each model anticipates held-out frequency-bin data after being calibrated on the remaining observations.

For a future or withheld observation $\tilde{y}$, the posterior predictive density is [11]:

$$p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \theta)\, p(\theta \mid y)\, d\theta \quad (3)$$

For exact LOO-CV, let y_{-i} denote the dataset with the i-th data point removed. The leave-one-out predictive density for y_i is:

$$p(y_i \mid y_{-i}) = \int p(y_i \mid \theta)\, p(\theta \mid y_{-i})\, d\theta \quad (4)$$

This is the posterior predictive density of the withheld data point under the model fit to the remaining $n - 1$ observations [11, 12].

The pointwise contribution to the leave-one-out expected log predictive density is:

$$\mathrm{elpd_loo},i = \log p(y_i \mid y_{-i}) \quad (5)$$

and summing over all data points yields the exact LOO expected log predictive density:

$$\mathrm{elpd_loo} = \Sigma_i\, \mathrm{elpd_loo},i = \Sigma_i \log p(y_i \mid y_{-i}) \quad (6)$$

If

$$\theta_i^{(s)} \sim p(\theta \mid y_{-i}), \quad s = 1, \ldots, S \quad (7)$$

are S MCMC draws from the leave-one-out posterior for the i-th omission, then Eq. (4) is approximated by:

$$p(y_i \mid y_{-i}) \approx (1/S)\, \Sigma_s\, p(y_i \mid \theta_i^{(s)}) \quad (8)$$

The Monte Carlo estimator of the pointwise LOO contribution is accordingly:

$$\widehat{\mathrm{elpd}}_\mathrm{loo},i \approx \log[\, (1/S)\, \Sigma_s\, p(y_i \mid \theta_i^{(s)})\,] \quad (9)$$

and the corresponding estimator of the total LOO score is:

$$\widehat{\mathrm{elpd}}_\mathrm{loo} = \Sigma_i\, \widehat{\mathrm{elpd}}_\mathrm{loo},i \approx \Sigma_i \log[\, (1/S)\, \Sigma_s\, p(y_i \mid \theta_i^{(s)})\,] \quad (10)$$

This is the standard Monte Carlo estimator of the Bayesian exact LOO quantity [11, 12].

In the present analysis, the existing MCMC framework implemented in holodeck [7] is utilized. Following Refs. [7, 10, 16], the five lowest-frequency bins of the NANOGrav 15 yr gravitational-wave background free-spectrum dataset are employed, corresponding to the HD-w/MP+DP+CURN analysis, so $n = 5$. For each LOO run, one of these five data points is withheld from the total likelihood and the model is refit to the remaining four points, producing draws $\theta_i^{(s)} \sim p(\theta \mid y_{-i})$ for that omission. Given that only five pointwise terms enter the LOO sum, model-comparison uncertainty should be interpreted with appropriate caution.

For each stored draw, the log-likelihood contribution of the omitted point is also retained:

$$\log p(y_i \mid \theta_i^{(s)}) \quad (11)$$

Using these stored values, the pointwise Monte Carlo estimator may be rewritten as:

$$\widehat{\mathrm{elpd}}_\mathrm{loo},i \approx \log[\, (1/S)\, \Sigma_s \exp(\, \log p(y_i \mid \theta_i^{(s)})\,)\,] \quad (12)$$

This expression is numerically equivalent to Eq. (9). This procedure constitutes exact LOO-CV in the sense that the model is refit independently for each omitted data point, and is therefore distinct from importance-sampling approximations such as raw importance sampling and PSIS-LOO discussed by Vehtari et al. [12], which instead reuse draws from the full posterior $p(\theta \mid y)$. For each omission, $S \sim 10^5$ posterior draws are employed.

3.2 Approximate LOO-CV

The exact procedure described in Section 3.1 requires n separate model refits, one per omitted data point. In the present analysis, with $n = 5$ PTA frequency-bin measurements as the LOO units, exact LOO-CV is computationally more demanding than an approximation that recycles draws from the full posterior.

An approximate LOO-CV can be constructed via importance sampling [12, 13]. Under the same pointwise factorization used in Section 3.1, the leave-one-out posterior satisfies:

$$p(\theta \mid y_{-i}) \propto p(\theta \mid y) / p(y_i \mid \theta) \quad (13)$$

Consequently, instead of sampling from $p(\theta \mid y_{-i})$ separately for each omission, draws $\theta^{(s)} \sim p(\theta \mid y)$ from the full posterior may be reused and reweighted to approximate the leave-one-out predictive density.

In raw importance sampling, the importance ratios are proportional to:

$$r_i^{(s)} = 1 / p(y_i \mid \theta^{(s)}) \quad (14)$$

These weights can become unstable when a held-out data point is influential [12, 13]. Vehtari et al. [12] therefore recommend Pareto-smoothed importance sampling (PSIS), which regularizes the largest importance weights by smoothing their upper tail. Using PSIS weights $w_i^{(s)}$, the approximate leave-one-out score is:

$$\widehat{\mathrm{elpd}}_\mathrm{psis\text{-}loo} = \Sigma_i \log(\, \Sigma_s\, w_i^{(s)}\, p(y_i \mid \theta^{(s)}) / \Sigma_s\, w_i^{(s)}\,) \quad (15)$$

This approximation eliminates the n separate refits required by exact LOO-CV and is therefore substantially faster in practice. PSIS-LOO was implemented using the Python package ArviZ [17]. Its reliability is assessed via the Pareto-$\hat{k}$ diagnostic: values $\hat{k} < 0.5$ indicate a reliable approximation, values between

approximately 0.5 and 0.7 warrant some caution, and values $\hat{k} > 0.7$ indicate that the approximation may be unreliable and that exact refitting is preferable [12]. In the present analysis, these diagnostics are evaluated independently for each of the five PTA frequency bins.

4. Results

This section presents the outcomes of model comparison and the computed ELPD values.

4.1 Model Comparison via Approximate LOO-CV

For the approximate method, the Pareto-$\hat{k}$ diagnostic must be assessed as discussed in Section 3.2. In Figure 1, these diagnostics are presented for all models across the five PTA frequency bins.

As evident from Figure 1, frequency bin 1 yields Pareto-$\hat{k}$ values exceeding 0.7 for several models, indicating that the PSIS approximation is unreliable for that bin. In contrast, all models produce acceptable Pareto-$\hat{k}$ values for the final frequency bin. Among the four models, only the ULDM simplified model maintains acceptable Pareto-$\hat{k}$ values across all five frequency bins. These elevated diagnostics motivate the use of exact LOO-CV as a more reliable basis for model comparison.

For the approximate method, it is important to assess the Pareto-$\hat{k}$ diagnostic, as discussed in Sec. III B. The Pareto-$\hat{k}$ values provide a quantitative measure of the reliability of the PSIS approximation for each pointwise term. In Figure 1, we present these diagnostics for the approximate model comparison.

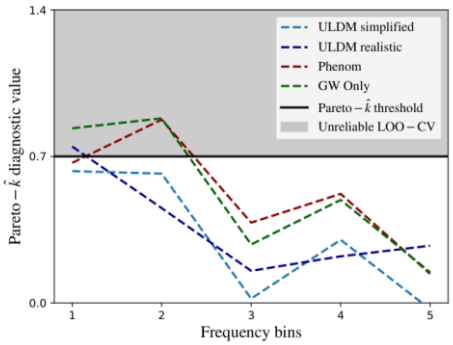


FIG. 1: Pareto-$\hat{k}$ diagnostic values for all four models across all five frequency bins. Values exceeding the threshold of 0.7 (horizontal line) indicate unreliable PSIS-LOO approximations.

4.2 Model Comparison via Exact LOO-CV

Table I presents the pointwise exact-LOO contributions for each model at the five PTA frequency bins, their aggregate over all bins, the difference in total ELPD relative to the best-performing model, and the corresponding estimated standard errors. Following Vehtari et al. [12], the standard error of the total LOO score is estimated from the dispersion in the pointwise contributions:

$$\mathrm{SE}(\mathrm{elpd}\hat{}_\mathrm{loo}) = \sqrt{[\, n \cdot \mathrm{Var}_\{i=1\}\hat{}\{n\}(\, \mathrm{elpd}\hat{}_\mathrm{loo},i\,)\,]} \quad (16)$$

where Var denotes the sample variance over the n pointwise contributions. Explicitly:

$$\mathrm{Var}(\mathrm{elpd}\hat{}_\mathrm{loo},i) = (1/(n-1))\, \Sigma_i\, (\mathrm{elpd}\hat{}_\mathrm{loo},i - \mathrm{elpd}\hat{}_\mathrm{loo})^2 \quad (17)$$

with

$$\mathrm{elpd}\hat{}_\mathrm{loo} = (1/n)\, \Sigma_i\, \mathrm{elpd}\hat{}_\mathrm{loo},i \quad (18)$$

For comparison against the best-performing model, the pointwise difference is defined as:

$$\Delta\mathrm{elpd}\hat{}_\mathrm{loo},i = \mathrm{elpd}\hat{}_\mathrm{loo},i\hat{}\mathrm{best} - \mathrm{elpd}\hat{}_\mathrm{loo},i \quad (19)$$

so that the total difference is:

$$\Delta\mathrm{elpd}\hat{}_\mathrm{loo} = \Sigma_i\, \Delta\mathrm{elpd}\hat{}_\mathrm{loo},i \quad (20)$$

with estimated standard error:

$$\Delta\hat{\mathrm{SE}} = \mathrm{SE}(\Delta\mathrm{elpd}\hat{}_\mathrm{loo}) = \sqrt{[\, n \cdot \mathrm{Var}_\{i=1\}\hat{}\{n\}(\, \Delta\mathrm{elpd}\hat{}_\mathrm{loo},i\,)\,]} \quad (21)$$

It is also useful to express the ELPD difference in standard-error units. The standardized LOO difference is defined as:

$$z_\mathrm{loo} \equiv \Delta\mathrm{elpd}\hat{}_\mathrm{loo} \,/\, \Delta\hat{\mathrm{SE}} \quad (22)$$

This quantity serves as a descriptive measure of the separation between two models in units of the estimated standard error. It is not interpreted as a formal frequentist significance statistic or as a Gaussian z-score, particularly because only n = 5 pointwise LOO terms enter the comparison.

As shown in Table I, the Phenom model achieves the highest exact-LOO score and is accordingly ranked first in predictive accuracy. However, the zloo values are not large, implying that the present data do not furnish strong evidence for a decisive predictive difference between Phenom and the remaining models.

TABLE I: Exact LOO-CV model comparison with pointwise ELPD contributions.

Model	i=1	i=2	i=3	i=4	i=5	elpd_loo	Δelpd_loo	ΔSÊ	z_loo
Phenom	-1.906	0.538	-0.234	-0.165	-1.962	-3.728	0.000	0.000	—
ULDM simplified	-1.322	0.135	-0.523	-0.424	-1.919	-4.053	0.324	0.890	0.364
GW Only	-2.693	0.622	-0.341	-0.247	-2.082	-4.741	1.012	0.754	1.342
ULDM realistic	-1.556	-0.339	-0.918	-0.919	-2.256	-5.988	2.259	1.114	2.028

TABLE II: Approximate PSIS-LOO model comparison with pointwise predictive contributions.

Model	i=1	i=2	i=3	i=4	i=5	elpd_psis	Δelpd_psis	ΔSÊ	z_psis
Phenom	-1.907	0.511	-0.232	-0.169	-1.959	-3.757	0.000	0.000	—
ULDM simplified	-1.303	0.142	-0.522	-0.416	-1.919	-4.017	0.260	0.796	0.327
GW Only	-2.767	0.553	-0.344	-0.259	-2.083	-4.900	1.143	0.718	1.592
ULDM realistic	-1.549	-0.335	-0.913	-0.922	-2.256	-5.975	2.218	0.989	2.243

TABLE III: Wall-clock times for approximately 10^5 samples. Measurements obtained on the University of Canterbury Research Cluster (AMD EPYC-Milan, 128 of 200 cores).

Process	Model	t_exact (min)	t_approx (min)
MCMC	ULDM simplified	169.9	35.0
	ULDM realistic	160.2	33.4
	Phenom	84.8	14.4
	GW Only	21.1	4.1
ELPD analysis	all	0.145	0.071

4.3 Comparison Between Exact and Approximate Methods

Table II presents the corresponding results obtained using the PSIS-LOO method. Despite the elevated Pareto-$\hat{k}$ values observed for certain bins, the PSIS-LOO scores show only negligible deviation from the exact LOO-CV results reported in Table I. This close agreement is further illustrated in Figure 2.

The PSIS-LOO approach would ordinarily be preferred on efficiency grounds, given that it is substantially faster than exact LOO-CV, as documented in Table III. However, the elevated Pareto-$\hat{k}$ values necessitated verification via exact refitting. Fortunately, given the scale of the present problem (n = 5 bins), the computational overhead of exact LOO-CV remained manageable.

4.4 Pairwise Comparison: ULDM Simplified vs. ULDM Realistic

Figure 2 displays the frequency-bin-wise predictive contributions for all models using both exact LOO-CV and PSIS-LOO. Based on Tables I and II together with Figure 2, no single model is consistently preferred over all others across every frequency bin. The clearest exception is the comparison between the ULDM simplified and ULDM realistic formulations (blue curves in Figure 2): the ULDM simplified model yields a larger pointwise predictive contribution in all five bins.

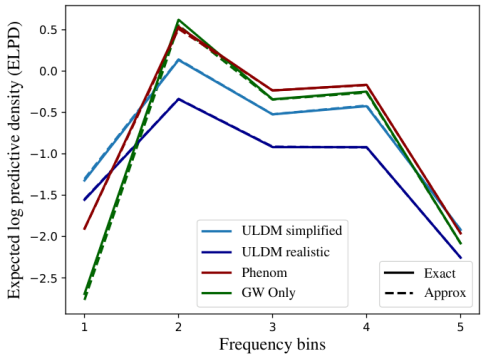


FIG. 2: Pointwise predictive contributions for all models across the five PTA frequency bins, shown for both exact LOO-CV (solid lines) and approximate PSIS-LOO (dashed lines).

All six pairwise model comparisons were also examined. The largest separation arises between the ULDM simplified and ULDM realistic models: the exact LOO difference is $\Delta elpd\hat{}_loo = 1.935$ with estimated standard error $\Delta S\hat{E} = 0.238$, placing the ULDM simplified model substantially ahead in predictive accuracy. This finding is also visually apparent in Figure 2, where the light-blue ULDM simplified curve consistently lies above the dark-blue ULDM realistic curve in all five bins.

To understand why the simplified model performs more favorably for the observed GWB data, the MCMC chains for both models were re-examined. For each sampled parameter combination, the trained Gaussian-process interpolator was used to obtain a predictive distribution for the strain spectrum, from which samples were drawn. Repeating this procedure over ~10^5 sampled parameter combinations yields an ensemble of strain spectra across the five PTA frequencies, from which the median and 95% posterior predictive interval for each model are computed and shown in Figure 3.

Figure 3 reveals that the strain predictions from the ULDM simplified model (right panel) are more tightly concentrated around the observed data points (grey violin distributions) compared with the ULDM realistic model. By contrast, the ULDM realistic model (left panel) produces a substantially broader posterior predictive distribution across the five PTA frequencies. This provides a qualitative explanation for the superior predictive performance of the simplified formulation on the observed GWB data.

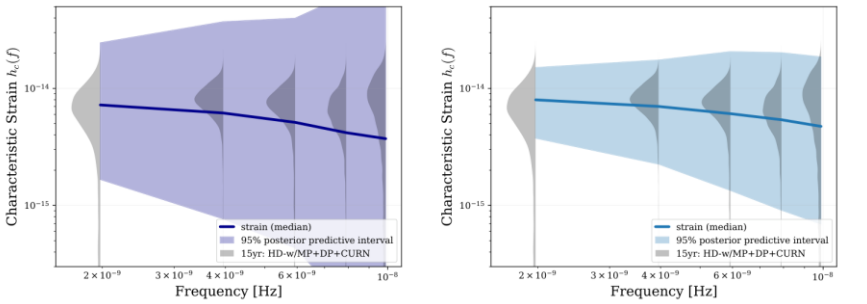


FIG. 3: Characteristic strain spectra for all sampled MCMC parameter combinations. The median and 95% posterior predictive interval are drawn from the Gaussian-process predictive distribution. Left panel: ULDM realistic model. Right panel: ULDM simplified model.

5. Conclusion

This work extends the ULDM analysis of Ref. [10] by performing a formal predictive model comparison. That earlier study demonstrated that ULDM-induced dynamical friction can attenuate low-frequency power in the PTA GWB and yields tractable constraints on the ULDM particle mass and effective fraction. The present work addresses whether the resulting ULDM models are preferred, in terms of predictive accuracy, over alternative SMBHB descriptions.

Four models were compared: the simplified and realistic ULDM formulations of Ref. [10], a phenomenological environmental model, and a gravitational-wave-only model. The comparison employed Bayesian leave-one-out cross-validation, with the five lowest PTA frequency bins serving as the LOO units.

The phenomenological environmental model achieves the highest total exact-LOO ELPD. However, the margins separating it from the other models are not large relative to the estimated standard errors. With the current five-bin dataset, no strong evidence is found for a decisive predictive preference among the four models overall.

The clearest outcome is the pairwise comparison between the two ULDM models. The ULDM simplified model yields a larger exact-LOO contribution than the realistic ULDM model in all five frequency bins, with $z_loo \approx 8$. The PSIS-LOO calculation produces a numerically similar result, but the elevated Pareto-$k\hat{}$ diagnostics necessitated verification via exact refitting.

These results clarify the current status of the ULDM interpretation of PTA data. Present observations are consistent with ULDM-induced low-frequency suppression of the GWB, but do not yet statistically prefer ULDM over more generic environmental characterizations of SMBHB evolution. Within the ULDM framework, the simplified implementation delivers better predictive performance than the physically detailed pinched-soliton formulation for the data considered. This outcome should not be interpreted as evidence that the simplified dynamical model is physically more accurate; rather, it constitutes the better predictive model under the assumptions and data employed in this analysis.

Forthcoming PTA datasets, incorporating more frequency bins and reduced measurement uncertainties, should render this comparison substantially more discriminating. On the theoretical side, more refined modeling of SMBHB evolution within ULDM solitons, as well as improved treatment of the core–halo relation in mixed ULDM–CDM scenarios, will be required before predictive preferences can be reliably translated into robust physical conclusions.

6. Acknowledgments

The authors are grateful to Russell Boey and Richard Easther for stimulating discussions. Financial support from the Marsden Fund Council grant MFP-UOA2131, administered by the Royal Society Te Apārangi with New Zealand Government funding, is gratefully acknowledged. The authors further acknowledge the University of Canterbury Research Cluster for providing computational resources (DOI:10.18124/CANTERBURYNZ-UCRCH, RRID:SCR_027870).

7. References

[1] Agazie, G., Anumarlapudi, A., et al. (2023). The NANOGrav 15 yr data set: Evidence for a gravitational-wave background. Astrophysical Journal Letters, 951, L8. https://doi.org/10.3847/2041-8213/acdac6

[2] Reardon, D. J., Zic, A., et al. (2023). Search for an isotropic gravitational-wave background with the Parkes Pulsar Timing Array. Astrophysical Journal Letters, 951, L6. https://doi.org/10.3847/2041-8213/acdd02

[3] Antoniadis, J., Arumugam, P., et al. (2023). The second data release from the European Pulsar Timing Array: III. Search for gravitational wave signals. Astronomy & Astrophysics, 678, A50. https://doi.org/10.1051/0004-6361/202346844

[4] Xu, H., Chen, S., et al. (2023). Searching for the nano-hertz stochastic gravitational wave background with the Chinese Pulsar Timing Array data release I. Research in Astronomy and Astrophysics, 23, 075024. https://doi.org/10.1088/1674-4527/acdfa5

[5] Miles, M. T., et al. (2024). The MeerKAT Pulsar Timing Array: The first search for gravitational waves with the MeerKAT radio telescope. Monthly Notices of the Royal Astronomical Society, 536, 1489–1500. https://doi.org/10.1093/mnras/stae2571

[6] Begelman, M. C., Blandford, R. D., & Rees, M. J. (1980). Massive black hole binaries in active galactic nuclei. Nature, 287, 307–309. https://doi.org/10.1038/287307a0

[7] Agazie, G., et al. (2023). The NANOGrav 15 yr data set: Constraints on supermassive black hole binaries from the gravitational-wave background. Astrophysical Journal Letters, 952, L37. https://doi.org/10.3847/2041-8213/ace18b

[8] Milosavljević, M., & Merritt, D. (2003). The final parsec problem. AIP Conference Proceedings, 686, 201–210. https://doi.org/10.1063/1.1629432

[9] Milosavljević, M., & Merritt, D. (2003). Long-term evolution of massive black hole binaries. Astrophysical Journal, 596, 860–878. https://doi.org/10.1086/378086

[10] Tiruvaskar, S., Boey, R., Easther, R., & Gordon, C. (2026). Ultralight dark matter constraints from nano-Hertz gravitational waves. Physical Review D, 113, 063541. https://doi.org/10.1103/PhysRevD.113.063541

[11] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis (3rd ed.). CRC Press.

[12] Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, 27, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

[13] Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), Bayesian statistics 4 (pp. 147–167). Oxford University Press.

[14] Boey, R., Kendall, E., Wang, Y., & Easther, R. (2025). Supermassive binaries in ultralight dark matter solitons. arXiv:2504.16348 [astro-ph.CO]. https://arxiv.org/abs/2504.16348

[15] Gordon, C., & Trotta, R. (2007). Bayesian calibrated significance levels applied to the spectral tilt and hemispherical asymmetry. Monthly Notices of the Royal Astronomical Society, 382, 1859–1863. https://doi.org/10.1111/j.1365-2966.2007.12707.x

[16] Tiruvaskar, S., & Gordon, C. (2026). Self-interacting dark-matter spikes and the final-parsec problem: Bayesian constraints from the NANOGrav 15-year gravitational-wave background. Physical Review D, 113. https://doi.org/10.1103/2hqm-qv99

[17] Martin, O. A., et al. (2026). ArviZ: A modular and flexible library for exploratory analysis of Bayesian models. Journal of Open Source Software, 11, 9889. https://doi.org/10.21105/joss.09889