Accelerating LISA inference with Gaussian processes

Jonas El Gammal ⓘ,[1, 2, *] Riccardo Buscicchio ⓘ,[3, 4, 5] Germano Nardini ⓘ,[1] and Jesús Torrado ⓘ[6, 7, 8]

[1] *Department of Mathematics and Physics, University of Stavanger, NO-4036 Stavanger, Norway*
[2] *Como Lake Center for Astrophysics, Department of Science and High
Technology, University of Insubria, via Valleggio 11, I-22100, Como, Italy*
[3] *Dipartimento di Fisica "G. Occhialini", Università degli Studi di Milano-Bicocca, Piazza della Scienza 3, 20126 Milano, Italy*
[4] *INFN, Sezione di Milano-Bicocca, Piazza della Scienza 3, 20126 Milano, Italy*
[5] *Institute for Gravitational Wave Astronomy & School of Physics and
Astronomy, University of Birmingham, Birmingham, B15 2TT, UK*
[6] *Dipartimento di Fisica e Astronomia "G. Galilei", Università degli Studi di Padova, via Marzolo 8, I–35131 Padova, Italy*
[7] *INFN, Sezione di Padova, via Marzolo 8, I–35131 Padova, Italy*
[8] *Instituto de Estructura de la Materia, CSIC, Serrano 121, 28006 Madrid, Spain*
(Dated: March 31, 2025)

Source inference for deterministic gravitational waves is a computationally demanding task in LISA. In a novel approach, we investigate the capability of Gaussian Processes to learn the posterior surface in order to reconstruct individual signal posteriors. We use `GPry`, which automates this reconstruction through active learning, using a very small number of likelihood evaluations, without the need for pretraining. We benchmark `GPry` against the cutting-edge nested sampler `nessai`, by injecting individually three signals on LISA noisy data simulated with `Balrog`: a white dwarf binary (DWD), a stellar-mass black hole binary (stBHB), and a super-massive black hole binary (SMBHB). We find that `GPry` needs $\mathcal{O}(10^{-2})$ fewer likelihood evaluations to achieve an inference accuracy comparable to `nessai`, with Jensen-Shannon divergence $D_{\rm JS} \lesssim 0.01$ for the DWD, and $D_{\rm JS} \lesssim 0.05$ for the SMBHB. Lower accuracy is found for the less Gaussian posterior of the stBHB: $D_{\rm JS} \lesssim 0.2$. Despite the overhead costs of `GPry`, we obtain a speed-up of $\mathcal{O}(10^2)$ for the slowest cases of stBHB and SMBHB. In conclusion, active-learning Gaussian process frameworks show great potential for rapid LISA parameter inference, especially for costly likelihoods, enabling suppression of computational costs without the trade-off of approximations in the calculations.

I. INTRODUCTION

In the last decade, the direct detection of gravitational waves (GWs) has transformed from a remarkable, singular accomplishment into a routine procedure. Currently, the LIGO-Virgo-KAGRA collaboration has observed approximately a hundred systems emitting GWs in the 10 – 1000 Hz frequency range [1]. Moreover, pulsar timing array experiments are possibly on the verge of gathering enough statistics to announce the first direct detection of GWs in the nHz range [2–5]. Gravitational waves in the mHz frequency range remain unobserved. LISA, with construction commissioned now and launch scheduled in a decade, is set to delve into this uncharted territory [6].

LISA poses data analysis challenges that are radically different from those of the other GW experiments, as it is a signal-dominated one, expected to observe a multitude of Galactic binaries, supermassive BHBs, EMRIs, and stellar-mass BHBs constantly populating the datastreams [6]. A primordial stochastic gravitational wave background (SGWB) as loud as the astrophysical sources might also be present [7]. To further complicate things, the zoology of LISA signals does not admit a common detection and reconstruction strategy: within the experiment's lifetime, some sources are monochromatic, others slowly drift, and others move fast outside the LISA frequency sensitivity. Analyzing the data in time chunks makes the identification of the long-duration sources more difficult, while keeping the whole datastream a priori makes the likelihood evaluations too heavy. On the other hand, splitting the data into frequency intervals is suitable for monochromatic sources [8], less so for those that are largely chirping. Despite their difficulty, these challenges must be solved to achieve the groundbreaking science promised by LISA [7, 9, 10].

Concerning the likelihood evaluation cost, several improvements are conceivable (see e.g., [11] for a review): speeding up waveform evaluation (e.g., through hardware acceleration or approximation schemes) [12–14], bypassing the likelihood evaluation (e.g., using simulation-based inference methods (SBI) [15–22]), or building surrogates of the likelihood function itself [23, 24]. In this paper, we focus on the latter, while agnostically retaining the waveform content and the signal Bayesian model intact. Thus, no approximation is made to either the Fourier transforms of the modeled signals or the likelihood computation.

Instead, we adopt the machine learning framework implemented in `GPry` [25, 26]. Within it, we interpolate the posterior with a Gaussian process [27], trained on a small number of evaluations that are sequentially proposed in an optimal way to minimize their number [28, 29]. As we will see, this approach can produce accurate inference with $\mathcal{O}(10^{-2})$ fewer likelihood evaluations than traditional Monte Carlo approaches. This translates into a speed-up of inference by a factor of 100 in the regime in

* jonas.e.elgammal@uis.no

which the overhead of `GPry` is subdominant, i.e., when the likelihood evaluation time is over a few seconds, and the dimensionality of the problem is $\mathcal{O}(10)$. The output is a surrogate model for the posterior that can be sampled with a Monte Carlo algorithm at virtually zero cost.

Within the general context of machine-learning accelerated inference of GW sources, the likelihood-based, *active-learning* approach taken by `GPry` differs from the likelihood-free, *amortized* approaches such as SBI in a number of ways: a) amortized approaches are much faster at the point of inference, in exchange for some costly pre-training, whereas the more expensive active-learning frameworks can be run with no upfront costs for variations of data or waveform modelling; b) likelihood-based approaches do not necessitate the simulated data to contain stochastic noise, and possess a direct way to evaluate goodness-of-fit. Both approaches are complementary and can thus coexist in the LISA parameter inference pipeline. To estimate the benefits for LISA of a machine learning framework similar to the one just described, we use `GPry` to perform parameter inference on some benchmark LISA signals simulated through `Balrog`, and compare the speed and accuracy of the results to those obtained with the state-of-the-art nested sampler `nessai`.

The paper is organized as follows: in Sec. II A we describe the target sources of our study: a supermassive black-hole binary (SMBHB), a stellar mass black hole binary (stBHB) and a Galactic double white dwarf (DWD) system; in Sec. II B we compare the waveform modeling available in literature; in Sec. II C we briefly describe the inference scheme, for individual source parameter estimation in the three source scenarios previously mentioned; in Sec. III A we present previous approaches for exact or approximate inference, and how they can be used as a starting point for our pipeline; in Sec. III B we briefly introduce how our algorithm models a posterior using a Gaussian process interpolator; in Sec. III C we detail how we perform, evaluate and compare our inference runs; in Sec. IV we present our results for the three source types above, and compare `GPry` and `nessai` on performance and accuracy; in Sec. V we draw conclusions and outline possible future developments.

II. SOURCES

A. Source types

We explore three different source classes, roughly categorized by their signal spectral content. Double white dwarfs (DWDs) are observed by LISA during their early inspiral, emitting quasi-monochromatic GWs largely detectable within the Galactic neighborhood [30–32]. Each signal persists in the LISA datastream for the entire mission, Doppler modulated by the satellite-constellation orbital motion within a very narrow frequency band ($\Delta f/f \leq 10^{-4}$). For DWDs emitting above approximately $2\,\mathrm{mHz}$, the GW-driven orbital tightening reaches a frequency evolution $\dot{f} \gtrsim 10^{-15}\,\mathrm{Hz}^2$ which LISA can measure over the nominal mission duration $T_{\mathrm{LISA}} = 4\,\mathrm{yr}$. As many as 10^7 DWD sources are expected to emit in the LISA band, with up to 1% individually detectable. They are unambiguously the most numerous deterministic sources expected for LISA, and their collective brightness makes its datastream strongly signal dominated below a few mHz. The brightest of these sources are identifiable after a few months [33], once a sufficient phase coherence emerges from the noisy datastream. In most of the available literature, a phenomenological parametrization of their signal is preferred over a physically-motivated one. Waveform models accurately taking into account the LISA response [34, 35] are typically fast, often leveraging frequency domain representation and heterodyning. It is uncertain whether such advantages will be retained in more realistic data analysis setups (see, e.g., [36], for recent developments on a frequency domain treatment of gaps).

stBHBs are the second class of sources we consider. They are expected to populate the whole LISA spectrum, the largest majority slowly drifting in frequency within the LISA mission duration. Only a handful of them will exit the band on its upper end in less than a year and eventually merge in the ground-based detector frequency band ($10\,\mathrm{Hz}$ to $1\,\mathrm{kHz}$) [37, 38]. Their waveforms are comparatively more complex than the DWD ones, with a parameter space equipped to describe eccentric, precessing, unequal-mass binaries [39, 40]. The in-band persistence of DWD and stBHB signals allows for a coherent integration of the data over millions of cycles during the nominal mission duration, making their detection heavily phase dominated.

Finally, we consider SMBHBs as the third category of GW sources: they are the most massive binaries expected to emit GWs in the LISA band. In the lifetime of the mission, SMBHBs will be detected as transient signals reaching signal-to-noise ratios (SNRs) as large as $\sim 10^3$, therefore being the loudest individual sources among the LISA ones. SMBHBs rapid evolution towards merger in band makes them prototypical to excite GW higher multipole modes, spin-precession [41]. However, orbit circularization [42] prevents from measuring large orbital eccentricities. State-of-the-art waveform models are phenomenological ones, calibrated against numerical relativity simulations with mass-ratios up to 1:18 [43]. Their computational efficiency is granted by decades of waveform developments for ground-based detectors, though the signal brightness questions the level of accuracy required to achieve unbiased parameter estimation [44]. The broadband nature of SMBHBs waveforms makes them the most expensive to compute in frequency domain (only second to extreme mass ratio inspirals), with up to 10^5 datapoints required for the lightest, most distant sources merging at about $10\,\mathrm{mHz}$. Even though time-domain truncation may reduce the number of frequencies to evaluate the waveform at, advanced global inference schemes (e.g. Gibbs-like sampling or SBI tech-

niques) may require the usage of conditional data with full-resolution frequency series.

B. Signal model

We model LISA data d as the linear superposition of noise n and signal s. Observations are collected through time-delay-interferometric variables, synthetic time series constructed from suitable delayed combinations of single-link inter-spacecraft laser phase measurements [45]. For simplicity, we assume the three LISA satellites orbiting in an equilateral triangular configuration with constant armlength of 2.5×10^9 m. Under such an approximation, the three interferometric variables, often referred to in literature as X, Y, Z, are linearly combined into the A, E, T variables, such that the respective noises are uncorrelated.

We model the GW strain emitted from a distant DWD as a quasi-monochromatic signal. Its two polarizations are described by

$$h_+(t; \boldsymbol{\theta}) = A(1 + \cos^2 \iota) \cos(2\pi f_{\mathrm{GW}}(t)/\mathrm{Hz} - \phi), \quad (1)$$

$$h_\times(t; \boldsymbol{\theta}) = -2A(\cos \iota) \sin(2\pi f_{\mathrm{GW}}(t)/\mathrm{Hz} - \phi), \quad (2)$$

where A denotes the GW amplitude, ι represents the source inclination with respect to the line-of-sight, $f_{\mathrm{GW}}(t)$ is the instantaneous GW frequency measured in the solar system barycenter frame, and ϕ is the binary orbital phase at the time t_0 at which LISA observations start. The amplitude A can be expressed as

$$A = \frac{2(G\mathcal{M}_c)^{5/3}}{c^4 d_L} (\pi f)^{2/3}, \quad (3)$$

while, to leading order, $f_{\mathrm{GW}}(t)$ reads

$$f_{\mathrm{GW}}(t) = f + \dot{f}(t - t_0), \quad (4)$$

with f and $\dot{f}$ being the orbital frequency and its (solar system barycenter frame) time derivative at time t_0, respectively. In Eq. (3), d_L denotes the source luminosity distance whose redshift is z, and $\mathcal{M}_c$ denotes the chirp mass

$$\mathcal{M}_c = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}}, \quad (5)$$

for a binary system of two component masses m_1 and m_2. Eq. (5) is frame-invariant, however we consider only solar system barycenter frame quantities hereafter.

The LISA detector response introduces an additional dependence upon the source position in the sky. We parametrize it by the source Ecliptic latitude b and longitude λ, and an overall polarization angle ψ. For inference purposes, we also reparameterize ϕ, ψ with two circular initial phases $\phi_L = \phi + \psi$ and $\phi_R = \phi - \psi$, respectively.

We assume the DWD source orbital evolution to be GW driven when injecting it into LISA data. Therefore,

the injected f and $\dot{f}$ must satisfy the constraint

$$\dot{f} = \frac{96}{5} \frac{(G\mathcal{M}_c)^{5/3}}{\pi c^5} (\pi f)^{11/3}. \quad (6)$$

However, we infer f and $\dot{f}$ as free independent parameters. Due to the signal being narrowband and at a much lower frequency than the LISA data sampling rate ($f_s = 0.2$ Hz), we speed up likelihood evaluations through heterodyning, filtering, and downsampling, resulting in a few hundred of datapoints per waveform evaluation.

Concerning stBHBs, we model their GW signal by following [46] where the waveform is computed through an adiabatic inspiral post-Newtonian expansion. As stBHBs drift much faster than DWDs in the LISA frequency band, their waveform exhibits (mild) sensitivity to the component masses m_1, m_2 and dimensionless spins χ_1, χ_2. For simplicity, we consider aligned-spin systems in circular orbits only, leaving the investigation of eccentric, precessing ones for future work. We reduce inference correlations with a convenient physical parameterization through the binary chirp mass $\mathcal{M}_c$, reduced mass ratio $\delta\mu = (m_1 - m_2)/(m_1 + m_2)$, and component dimensionless spin magnitudes $\chi_{1,2}$; its initial orbital frequency f_0, and left- and right-handed phases ϕ_L, ϕ_R. The extrinsic parameters are decomposed as follows: the source position and inclination are parameterized by the square root of two circular amplitudes $A_{L,R} = (1 \pm \cos \iota)/\sqrt{2d_L}$, the sin-ecliptic latitude $\sin \beta$, and longitude λ. TDIs are constructed through a rigid adiabatic approximation [47]. In previous work [46], waveforms were evaluated only at a few hundreds of points, employing Clenshaw-Curtis quadrature to approximate the likelihood in Eq. (7). This was made possible by analyses of noiseless data, whose smoothness allows for such an integration scheme. In turn, in this work we focus on noisy data, and hence we use the full GW frequency content, resulting in around 10^4 data points per waveform.

Finally, SMBHBs signals are described through phenomenological, numerical-relativity calibrated waveforms, as implemented in `IMRPhenomXHM` [48]. This waveform family smoothly captures the inspiral-merger-ringdown structure of a binary merger signal in frequency domain, accounting for higher-modes emission. Despite being extremely fast, thanks to decades-long optimization for current and future ground-based detectors [49], the LISA frequency resolution makes the waveform array typically long: in this study we consider a system emitting up to 3.5 mHz, reaching its merger $\tau_m = 4.138 \times 10^6$s after the start of the mission. We do not consider any time-domain truncation scheme, and model the signal at the highest frequency resolution available. The signal is parameterized by the binary chirp mass, its reduced mass ratio, the component dimensionless spin magnitudes (assumed aligned with respect to the angular momentum), the time-to-merger τ_m, the luminosity distance d_L, the sine-ecliptic latitude $\sin \beta$ and ecliptic longitude λ, the cosine inclination $\cos \iota$, the initial orbital phase ϕ_0^{orb}, and the polarization angle ψ. All quantities are defined in the

solar system barycenter frame, and non-conserved ones are defined at a reference frequency $f_{\text{ref}} = 10^{-4}$ Hz.

Throughout this work, we assume perfectly known, Gaussian, instrumental noise [50], superimposed on likewise perfectly known, Gaussian, confusion noise, whose level is modeled as in [51] as a function of T_{LISA}. In the larger context of global fit pipelines, this is equivalent to performing inference on the three chosen sources after all resolvable ones have been identified and perfectly subtracted from the data. We further simplify the two noise models assuming both zero mean and perfectly stationary, thus reducing their entire description to simple power spectral densities [50].

C. Likelihood

Given the assumptions detailed in Sec. II B, the likelihood of observed data $d_k = A, E, T$ in frequency domain reads

$$\log \mathcal{L}(d|\boldsymbol{\theta}) = -\sum_k \frac{\langle d_k - s_k(\boldsymbol{\theta}) \mid d_k - s_k(\boldsymbol{\theta})\rangle_k}{2} + \text{const.} \quad (7)$$

where d_k denotes the superposition of noises realizations and each injected signal as described in Table I, II, and III, respectively. Thanks to the stationarity of noise in each datastream and uncorrelatedness across them, the inner product is simply given by

$$\langle x \mid y \rangle_k = 4\text{Re} \int_0^{+\infty} \mathrm{d}f \frac{\tilde{x}(f)\tilde{y}^\dagger(f)}{S_{n,k}(f)} . \quad (8)$$

Finally, $s_k(f; \boldsymbol{\theta})$ denotes a proposed GW signal with parameters $\boldsymbol{\theta}$, as observed in the k-th datastream, and $S_{n,k}$ the noise power spectral density in the same datastream. We characterize the overall source brightness with the SNR, defined as

$$\text{SNR}^2 = \sum_{k=A,E,T} \langle s_k(f; \boldsymbol{\theta}) \mid s_k(f; \boldsymbol{\theta})\rangle_k . \quad (9)$$

In this study, we present two approaches to obtain posterior samples for each inference, according to

$$p(\boldsymbol{\theta}|d) \propto \mathcal{L}(d|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) , \quad (10)$$

where $\pi(\boldsymbol{\theta})$ denotes the prior assumption $\boldsymbol{\theta}$. In Sec. III A we detail the construction of priors for each source category, which we assume to be uniform over the prescribed ranges.

III. INFERENCE

A. Setting priors

Pre-constraining the parameter space of the inference problem down to a region around the location of the bulk probability mass, henceforth referred to as "mode", makes surrogate-posterior approaches such as `GPry` significantly faster and more robust. This can usually be achieved with methods that avoid the evaluation of the expensive posterior, via e.g. approximations in the likelihood, template matching with an approximate waveform or machine-learning forward modeling [12, 14, 52]. These methods can produce rough estimates of the location and span of the posterior mode at a very low computational cost.

The DWDs live in a narrow frequency band and can be initially constrained using frequentist triggers with a sliding-window method that scans the frequency domain. Additionally, by using an optimizer, it is possible to obtain a maximum likelihood estimate (MLE), and an estimate of the Fisher information matrix. In combination, these methods allow the setting of priors that sufficiently encapsulate the mode of the posterior distribution, as done in [8]. Since the DWD are mostly a test case for our study, we set conservative priors by hand, encapsulating $\sim 10\sigma$ for each unbounded parameter.

or stBHBs our approach to pre-constraining the parameter space is the one introduced in [53] and successfully applied to LISA data in [54]. This method employs a semi-coherent search combined with Particle Swarm Optimization (PSO) to efficiently scan the large parameter space involved. The semi-coherent approach divides the data into frequency-domain segments, analyzing each individually, and then combining the results. This technique balances sensitivity and computational efficiency by widening the posterior distribution over the parameter space, thus helping to locate the posterior bulk.

The path traced by the particles in the PSO can then be used to find regions in the parameter space with high posterior density values. For our stBHB analysis, we use a subset of 5000 samples from the PSO paths, obtained from 256 data segments, and evaluate the posterior in Eq. (10) at these locations. We then restrict the prior to the smallest hyper-rectangle containing PSO posterior samples within a 10σ confidence region from the peak, assuming a multivariate Gaussian distribution as the posterior distribution (for a detailed discussion, see App. A of [25]). In addition, we use a small set of these samples close to the top of the mode as an initial training set for `GPry`. Together, the shrunken prior and the initial training set eliminate the need to explore the parameter space and let `GPry` focus on mapping the mode, thus combining the strengths of both approaches: a fast initial exploration of the parameter space by the PSO followed by `GPry` which maps the mode with very few evaluations of the relatively slow-to-evaluate posterior distribution. The PSO search takes $\mathcal{O}(10\,\text{min})$, adding only very little overhead to our pipeline. We perform the initial PSO on noiseless data to introduce an additional bias beyond the one arising from their segmentation.

For the less explored case of SMBHBs, we assume that a similar PSO approach, a neural-network or a frequentist one can be used to approximate the mean and covariance

of the posterior mode (see, e.g., [52, 55]).

Due to the simpler structure of the posterior, we can set a larger uniform prior covering $> 10\sigma$ in each dimension. As a proxy for a search, we generate Monte Carlo samples of the noiseless posterior distribution and use its mean and covariance to draw a set of 35 samples from a multivariate Gaussian distribution. `GPry` is initialized with these samples which are close to the top of the mode but biased. If multimodalities are present, as is expected in the sky location for low latency searches [56–58], `GPry` would be initialized with points from all modes.

B. Gaussian process posterior interpolation

The inference algorithm employed in this study uses a Gaussian process regressor (GPR) to create an approximate model of the posterior density function, using a small set of evaluations performed at optimal locations. This approximation is then used as a *surrogate* model from which we can draw, at very low computational cost, Monte Carlo (MC) samples that very closely resemble samples from the true posterior. Contrary to amortized machine learning-based approaches such as SBI, our surrogate model is built sequentially at runtime (an approach known as *active learning*), and does not rely on previous training. `GPry`'s approach is more closely related to variational inference (see, e.g., [59]), with the difference that `GPry` does not need derivatives of the posterior. As we will see, the necessary number of evaluations of the GW signal likelihood is at least $\mathcal{O}(10^{-2})$ smaller than those needed by `nessai` (which is already more efficient than traditional Nested Sampling implementations).

We use `GPry` [25, 26, 60] to construct such a surrogate model. In this subsection, we adopt the notation most commonly used in the context of Gaussian processes where $\mathbf{x}$ refers to a vector in the sampling space (equivalent to θ above) and y is the value of the target function. `GPry` iteratively proposes points $\mathbf{x}$ in parameter space at locations where the expected gain in information about the posterior is maximized. With them, at every iteration, it builds an approximation of the posterior log-density function $\log p(\mathbf{x}|\mathcal{D})$ under some data $\mathcal{D}$ as the mean of a Gaussian process conditioned on the current set of *training* samples $(\mathbf{X}, \mathbf{y})$, where $\mathbf{X} = \{\mathbf{x}^{(i=1,\cdots)}\}$ and $\mathbf{y} = \{\log p(\mathbf{x}^{(i=1,\cdots)})\}$:

$$\log p(\mathbf{x}|\mathcal{D}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')|\mathbf{X}, \mathbf{y}) . \quad (11)$$

Here $k(\mathbf{x}, \mathbf{x}')$ represents the covariance function, for which `GPry` uses a d-dimensional inverse-squared Radial Basis Function (RBF) kernel allowing for a different length-scale in each dimension of the sampled parameter space:

$$k(\mathbf{x}, \mathbf{x}') = C^2 \prod_{i=1}^{d} \exp\left(\frac{(x_i - x_i')^2}{2l_i^2}\right) , \quad (12)$$

with C and l representing, respectively, the output and length scales of the Gaussian process. The null mean of the Gaussian process prior in Eq. (11) applies to a transformed set of $\mathbf{y}$ log-posterior values so that they have null mean and unit standard deviation. Hereon we drop the explicit dependence on the training data $\mathbf{X}, \mathbf{y}$.

The mean of the conditioned Gaussian process, with which we approximate the log-posterior density, is computed as

$$\mu(\mathbf{x}_*) = \mathbf{k}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} , \quad (13)$$

where $(\mathbf{k}_*)_i = k(\mathbf{x}^{(i)}, \mathbf{x}_*)$, $(\mathbf{K})_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$, and σ_n^2 is an estimate of the numerical uncertainty of log-posterior values. The standard deviation of the conditioned Gaussian process, used in the acquisition function defined below, is $\sigma(\mathbf{x}) = \sqrt{\text{diag}(\Sigma(\mathbf{x}))}$, where

$$\Sigma(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_*. \quad (14)$$

The hyperparameters of the kernel, i.e., its output and length scales, hereon denoted collectively as $\mathbf{\Lambda}$, are determined by maximizing their marginalized likelihood [27]. The mean and standard deviation of a Gaussian process conditioned on a set of training samples can be seen in the upper row of Fig. 1.

Optimizing the kernel hyperparameters $\mathbf{\Lambda}$ eventually dominates the overhead of the algorithm, as it requires multiple kernel matrix inversions that scale as $\mathcal{O}(N^3)$, with N being the number of training samples. In order to mitigate this, we only perform a full re-fit of the hyperparameters at every few iterations of the algorithm (see Appendix B). In general, overhead costs start making `GPry` an impractical approach for dimensionalities larger than a few tens, depending on the cost of the likelihood. The number of training samples, which drives the overhead costs, needed for accurate posterior reconstruction depends on the dimensionality of the problem. In exchange for this overhead, `GPry` reduces the number of posterior evaluations required with respect to traditional samplers by a factor of $\mathcal{O}(10^2)$. Therefore, `GPry`'s advantage in performance increases for low dimensions and large costs per posterior evaluation. An approximate rule of thumb is that `GPry` is faster for dimensionalities lower than a few tens when the posterior evaluation time is $\mathcal{O}(1\,\text{s})$ or higher.

As a further refinement of the surrogate model, we multiply the GPR by a Support Vector Machine (SVM) classifier, of comparatively negligible computational cost, trained both on the evaluations used for the GPR, and those rejected because their log-posterior density is either negative infinity or very low with respect to the best training point. This SVM is used to partition the parameter space into regions in which the true log-posterior is expected to return a finite, or negative infinity value; the latter is used as an exclusion region where future candidates are automatically rejected without evaluating their true log-posterior.
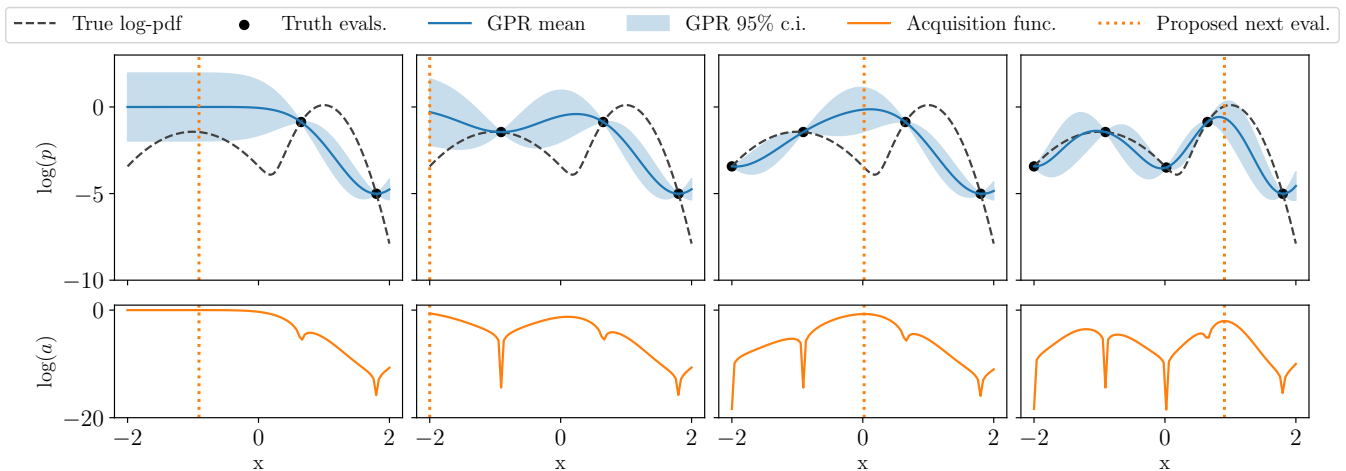
FIG. 1. Simplified illustration of the `GPry` algorithm on a 1-dimensional Gaussian mixture test function. Each column, corresponding to consecutive iterations, shows on the top the true target log-pdf (dashed), the current set of evaluations (black points), and the current GPR model mean from Eq. (13) (blue, solid) and 95% confidence interval defined by Eq. (14) (blue, shaded); the bottom panel shows the current acquisition function values from Eq. (15), whose maximum (dotted orange) will be proposed for evaluation for the next iteration. Not illustrated are more complex aspects of the algorithm, such as the batch proposal of points [25], and the procedure to obtain approximate maxima of the acquisition function [26].

To enable active sampling, we introduce an acquisition function, denoted as $a(\mathbf{x})$, which guides the sampling process by quantifying the expected utility of sampling the true posterior at each point in the parameter space:

$$a(\mathbf{x}) = \exp\left(2\zeta \cdot \mu\left(\mathbf{x}\right)\right)\left(\sigma(\mathbf{x}) - \sigma_n\right) , \qquad (15)$$

ζ is a scaling factor that balances exploration and exploitation. The learning efficiency is maximized when this scaling factor is made dimensionality-dependent, increasingly encouraging exploration for larger dimensionalities: $\zeta = d^{-c}$, with $c > 0$ [25]. In this paper, we empirically set this scaling factor to $\zeta = d^{-0.65}$, promoting exploitation slightly more than the value derived in [25] for Gaussian distributions. The effect of evaluating sequentially at the optimum of the acquisition function can be seen in Fig. 1.

The acquisition function is optimized through the NORA active sampling strategy described in [26]: we draw MC samples from the mean $\mu(\mathbf{x})$ of the GPR using a Nested Sampler (NS), in our study `PolyChord` [61, 62]. The acquisition function is then evaluated at the resulting NS samples, and its value is used to produce a pool of candidate points. This pool is ranked using the Kriging believer [63] prescription so that the n-th point is assigned a conditioned acquisition function value assuming a true posterior evaluation at the $n-1$ points above it. The optimal batch size is approximately equal to d [25], making the `GPry` algorithm efficiently parallelizable up to d processes using MPI.

The use of a NS at the acquisition step, that explores the full surrogate posterior (as opposed to directly maximizing the acquisition function), makes it easier for `GPry` to map a multimodal posterior, such as those expected in the sky localization parameters for low-latency signals, as demonstrated in [26]. This ability can be further boosted by making the acquisition function more exploratory (lower scaling factor ζ in Eq. (15)), and the exploration of the posterior more thorough (larger number of *live points* of the NS).

At the end of every iteration, convergence is checked and considered reached as soon as one of two criteria is fulfilled at least twice consecutively: the value of the likelihood at the new proposed sampling locations is close enough to their GPR-predicted value (see [25] for clarification), or the Gaussian-approximated Kullback-Leibler divergence[1] between consecutive NORA NS runs is small enough.

After convergence of the Bayesian optimization loop has been reached, MCMC samples of the surrogate model are generated. This typically only takes a few seconds since the evaluation of the surrogate model is very fast at $\mathcal{O}(10^{-5}\,\mathrm{s})$. To do so, we use `Cobaya`'s implementation of the MCMC sampler of `CosmoMC` [64, 65].[2] A flow chart of the algorithm is shown in Fig. 2.

---

[1] I.e., the Kullback-Leibler divergence, see Eq. (17), when distributions are approximated as multivariate Gaussians defined by their respective empirical means and covariance matrices.

[2] Notice that any other sampler, including those that require gradients, could be used without changing our conclusions.
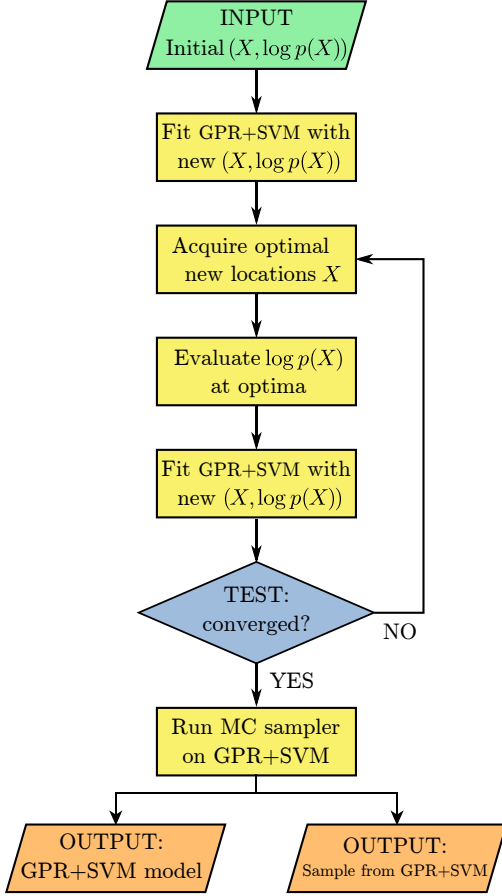
FIG. 2. Simplified flow chart of the `GPry` algorithm. Looking at Fig. 1, the GPR at the top of its first column presents the initial stage, where the GPR has been fit to an initial set of two samples. The main loop (*"acquire→evaluate→fit"*) corresponds sequentially to finding the location of the maximum of the acquisition function in the bottom row (dotted vertical line), evaluating the log-posterior there, and fitting the GPR to obtain the new model at the top of the following column.

C. Inference strategy and methodology for validating the results

`GPry` adopts default values for the parameters controlling the some of the aspects of the algorithm mentioned above, based on test runs on typical scenarios [25, 26]. The peculiarities of the problem at hand motivate changing some of these defaults in each case, as detailed in Table IV in Appendix B, and summarized below:

- For all three sources, especially for the stBHB and SMBHB, the log-likelihood presents significant numerical noise with respect to small changes in the waveform parameters. If not correctly accounted for, `GPry` interprets this sizable noise contribution as physically meaningful, which may lead to overfitting. We alleviate this problem by choosing large values of the expected noise scale σ_n in Eq. (14).

Away from the mode, for very low likelihood values, the numerical noise dominates, so we raise the SVM classifier cutoff to exclude low-valued regions from the GPR.

- For the sources with the slowest likelihood, the stBHB and especially the SMBHB, it makes sense to increase the overhead of the algorithm in exchange for reducing the number of necessary true posterior evaluations for convergence. Hence, we increase the frequency and the number of restarts for the GPR hyperparameters optimization. Similarly, we update the set of NS samples from mean GPR more often, and, for the SMBHB, reduce the number of Kriging steps.

- Since the set of initial points for the stBHB and SMBHB is very informative (see Sec. III A), it is advantageous to define a *trust region* around the current training set restricting the area where new evaluations are proposed. This region is the minimal hyper-rectangle containing training samples with posterior density above some cutoff with respect to the best one.

For each source type, we consider a high-SNR source signal as a noiseless LISA data stream, then inject it in multiple simulated noise realizations The three easier noiseless inference problems are used for consistency checks (e.g., robustness with respect to initialization) and are not presented below.

In order to benchmark `GPry`'s performance, both in terms of computational cost and inference accuracy, we pair every `GPry` run in each noise realization with a similar run with the machine-learning-enhanced nested sampler `nessai`, which has proven to be an efficient and reliable sampler in the context of GW data analysis [66–68].

We perform two tests on the two sets of runs. The first focuses on the accuracy of the full pipeline, from signal and noise generation to MC sampling. In literature, this test is often referred to as a *pp*-plot [69, 70]. For a given sampler choice and source category, we perform N inference runs on independent noise realizations, and compute the empirical quantiles $\{q_i\}_i^N$ corresponding to the injected parameters for the inferred posterior. In the limit $N \to \infty$ the cumulative distribution function of quantiles across runs is theoretically expected to approach that of the uniform distribution over the unit interval. Deviations from the asymptotic distribution due to finite N can be estimated numerically, and confidence intervals constructed accordingly. We present results of this test across source categories in Figs. 3a, 4a and 5a, respectively.

In the second test, we focus instead on a direct comparison between posteriors obtained through inference with `nessai` and `GPry` in paired runs on the same noise realizations. To do so, we evaluate the Jensen-Shannon (JS) divergence D_{JS} between each `nessai` posterior distribution P and the `GPry` surrogate model P_{GP} over the

parameter space [71]

$$D_{\text{JS}}(P||P_{\text{GP}}) = \frac{1}{2}\left(D_{\text{KL}}(P||M) + D_{\text{KL}}(P_{\text{GP}}||M)\right), \quad (16)$$

where $M = \frac{1}{2}(P + P_{\text{GP}})$ is the mixture distribution of P and P_{GP}. The Kullback-Leibler (KL) divergence between two continuous probability distributions P, M with densities $p(x)$, $m(x)$ is defined as

$$D_{\text{KL}}(P||M) = \int p(x) \log\left(\frac{p(x)}{m(x)}\right) \, \mathrm{d}x . \quad (17)$$

In practice, we compute the KL divergence as a Monte Carlo sum of the samples from GPry and nessai. In this paper, we use natural logarithms for the divergence calculations. The JS-divergence $D_{\text{JS}}(P||Q)$ approaches zero if and only if P and Q describe the same distribution and is upper-bounded by $\log 2$. For inference purposes, values of $D_{\text{JS}} \lesssim 0.05$ would make GPry as accurate as traditional samplers, whereas values up to $D_{\text{JS}} = 0.1$ could be considered precise enough, given the large computational trade-off. We show the distribution of D_{JS} for different source categories in Figs. 3b, 4b and 5b, respectively.

Following the formulas in Appendix A, for the dimensionality of our problems $D_{\text{JS}} = 0.05$ ($D_{\text{JS}} = 0.1$) would translate into a mean deviation in each parameter of $\approx 0.08\sigma$ ($\approx 0.11\sigma$) if assuming similar covariances, or alternatively a misestimation of the error of $\sim 15\%$ ($\sim 25\%$) if assuming similar means.

IV. RESULTS

A. Double white dwarf system

For a single injection of a DWD system, the waveform and subsequent likelihood computations are fast ($\sim 10^{-3}$ s). Hence, we do not expect significant savings in wall clock computation time between GPry and nessai. We therefore use it to test the GPry algorithm and gain some insight on its reliability.

Parameter	Symbol	Value
Ecliptic longitude	λ	2.0 rad
Ecliptic sine-latitude	$\sin\beta$	0.479
Amplitude	A	$2 \cdot 10^{-23}$
Frequency	f	0.00377 Hz
Frequency derivative	$\dot{f}$	2×10^{-18} Hz^2
Cosine-inclination	$\cos\iota$	0.4
Left phase	ϕ_L	1.3 rad
Right phase	ϕ_R	1.5 rad
SNR		23.64

TABLE I. Injected values for the sampled parameters of the DWD system and total source SNR.

The injected parameters for the benchmark source are shown in Table I. We draw 200 noise realizations according to our model in Sec. II B. For this source all

parameters are constrained and the posterior distribution exhibits a single, localized nearly-Gaussian mode. For each noise realization, we perform separate inference runs with GPry and nessai, with the nessai runs performed with 2000 live points. We then generate a PP plot comparing the performance of both algorithms (see Fig. 3a), and find a similar accuracy for the reconstruction. Furthermore, we compute the JS divergence, D_{JS}, between GPry and nessai for each noise realization, and show its histogram in Fig. 3b. This comparison shows that both samplers are in excellent agreement for all but one noise realization. In Fig. 8 we show a corner plot overlaying the posterior contours obtained by GPry and nessai for the realization corresponding to the median D_{JS}. There we can observe the clear agreement between the two approaches, achieved with 1/300 fewer likelihood evaluations by GPry compared to nessai. The locations of the GPry evaluations can be seen in the upper triangle of Fig. 8.

In Fig. 9 we show a corner plot and the posterior contours obtained by GPry and nessai corresponding to the highest $D_{\text{JS}} = 0.12$. Although the mode has been found by GPry in this example, it remains underexplored. Tightening the convergence criterion would eliminate this problem in exchange for higher computational costs, but maintaining the two-orders-of-magnitude difference in the number of likelihood evaluations with respect to nessai. Only one of the 200 runs performed shows this behavior with $D_{\text{JS}} > 0.05$ which leads us to conclude that the precision and accuracy of GPry is sufficient in this context.

B. Stellar origin binary black holes

For one injection of a stBHB system, the cost of a single evaluation of the inference pipeline (waveform and likelihood calculations) is $\sim 10^{-1}$ s, which is significantly higher than for DWDs. Thus, the savings here are potentially higher, which makes GPry a worthy approach.

The benchmark source's injected parameters are shown in Table II. Unfortunately, as the semi-coherent search presented in Sec. III A does not provide us with a reliable estimate of the phases, sampling these proves to be difficult with GPry. This is further complicated by the periodic nature of these parameters. We therefore fix the values to the injected ones. Contrary to the DWD case, the resulting posterior is highly non-Gaussian: it is heavy-tailed and has a large curving degeneracy. As discussed in [25], exploring the full posterior in a reasonable amount of time poses a challenge to GPry.

We generate 100 noise realizations and perform inference runs for each of them with GPry and nessai, with the nessai runs performed with 2000 live points. We then generate both a PP plot (see Fig. 4a), and compute the JS divergence for each pair of runs, whose histogram is shown in Fig. 4b. From the PP plot, it is clear that GPry performs worse than nessai, even if both samplers

(a)

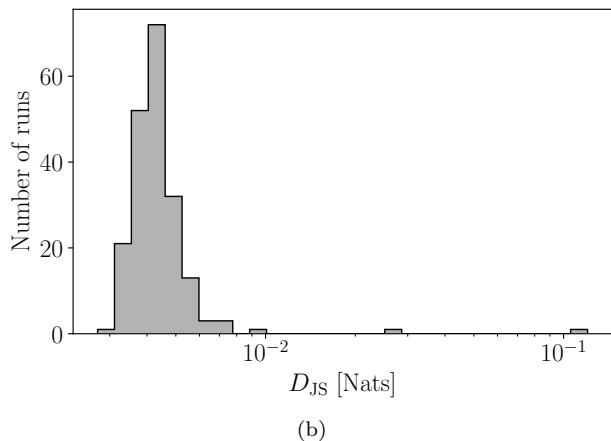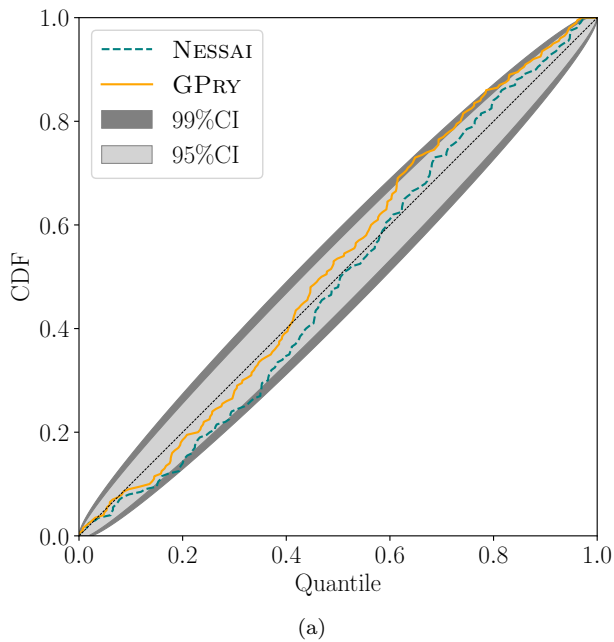(b)

FIG. 3. PP plot **(a)** and Jensen-Shannon divergence **(b)** for 200 DWD runs with different noise realizations. `nessai` and `GPry` show comparable accuracy in the former, consistent at 99% confidence (dark gray shaded area) with the theoretical prediction (dotted black line) across all runs, and at 95% confidence (light gray shaded area) for the largest majority of them. Relatively to `nessai`, `GPry` reconstructs the posterior shape reliably with only one run exceeding the target of $D_{\rm JS} = 0.05$.

show reasonably good performance. This is reflected in the higher JS divergences between `nessai` and `GPry` (see Fig. 4b), localized mostly in the $[0.1, 0.25]$ interval, with a few $D_{\rm JS} \gtrsim 0.3$ outliers.

The effect of the $D_{\rm JS} \sim 0.2$ divergence is illustrated in Fig. 10, which shows the result of the median $D_{\rm JS}$ run with `GPry` and `nessai`. As we can see, although the resulting mode for `GPry` is localized correctly towards the injected value, it fails to explore a fraction of the posterior corresponding to the large-values tail of the $(\mathcal{M}_c, \delta\mu)$ degeneracy. The handful of cases with higher $D_{\rm JS}$ (up to

Parameter	Symbol	Value
Redshifted chirp mass	$\mathcal{M}_c$	$48.618\,\mathrm{M}_\odot$
Reduced mass-ratio	$\delta\mu$	0.5
Ecliptic longitude	λ	$0.19\,\mathrm{rad}$
Ecliptic sine-latitude	$\sin\beta$	$0.82\,\mathrm{rad}$
Initial orbital frequency	f_0	$1.87\,\mathrm{mHz}$
Left phase (fixed)	ϕ_L	$0.97\,\mathrm{rad}$
Right phase (fixed)	ϕ_R	$1.76\,\mathrm{rad}$
Left square-root amplitude	$\sqrt{A_L}$	$12.57 \cdot 10^{-5}$
Right square-root amplitude	$\sqrt{A_R}$	$1.13 \cdot 10^{-5}$
Dimensionless spin	χ_1	0.223
Dimensionless spin	χ_2	0.262
SNR		16.79

TABLE II. Injected values for the parameters of the stBHB system, and total source SNR. All parameters are sampled except for the phases, for which the method described in Sec. III A failed to provide reliable estimates. The detector-frame individual masses are $m_1 = 99.55\,\mathrm{M}_\odot$ and $m_2 = 33.18\,\mathrm{M}_\odot$.

0.6) present the same sort of effect, and small $(< 1\sigma)$ biases for other parameters.

Possible mitigation strategies include fine-tuning of the `GPry` hyperparameters, to increase the chance that it fits this particular problem better (e.g. that it does not converge prematurely), as well as the use of alternative parameterizations whose posterior would not present these strong non-Gaussian features. We leave this endeavor for future work. It must be remarked that this difficulty also affects `nessai`, whose precision we had to increase in order to map this posterior correctly, so that $\mathcal{O}(10^2)$ more evaluations are needed than for the other two test sources (see Fig. 6a).

As explained in Sec. III A, we are seeding the stBHB runs discussed in this section with high-likelihood points from a noiseless Semi-Coherent PSO run. Since a noise realization introduces a bias in the inferred source parameters with respect to the noiseless case, we have investigated whether this under performance may be related to the use of a biased initial set of samples in the `GPry` runs. To do this, we have performed 100 additional paired runs with a noiseless injection, but found the same under-exploration effect with a similar magnitude.

We find that our approach reliably recovers the expected central values, and therefore could be used for source subtraction or fast, preliminary analysis; it is however suboptimal for full statistical source characterization. There is ongoing development of `GPry` addressing more robust inference in highly non-Gaussian distributions such as this stBHB posterior.

C. Supermassive binary black-hole

We now focus on the injection of a single SMBHB in noisy LISA data. Here, the cost per evaluation of the inference pipeline is larger than a few seconds and there-
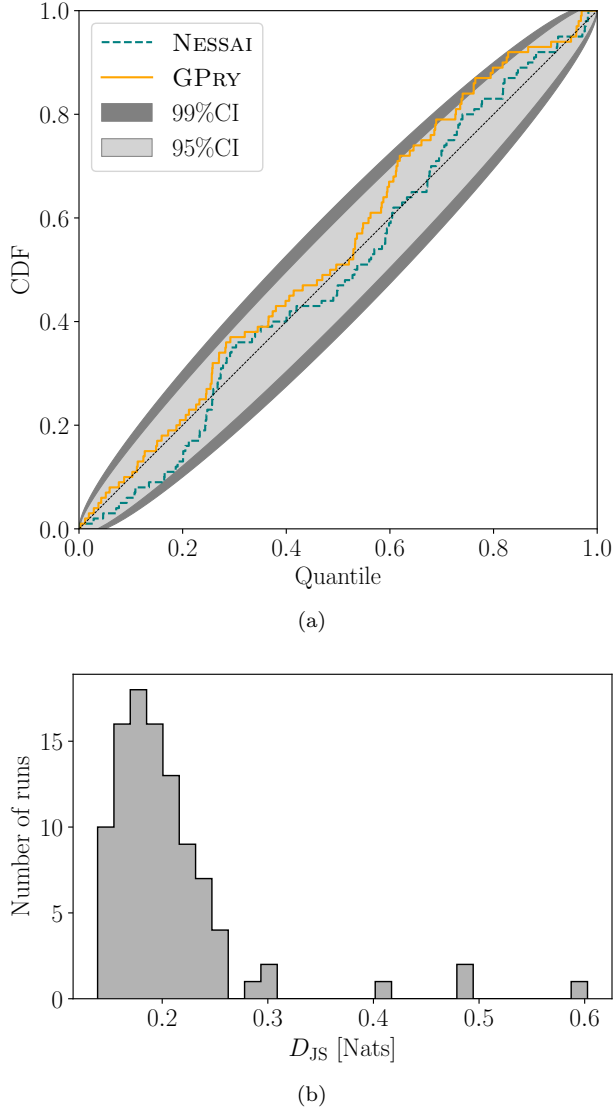
(a)

(b)

FIG. 4. PP plot **(a)** and Jensen-Shannon divergence **(b)** for 100 stBHB runs with different noise realizations. While `nessai` and `GPry` show comparable accuracy in the former, consistent at 99% CL (dark shaded region) with the theoretical prediction (black dotted line), the distribution of D_{JS} that is entirely above the target value of 0.05 indicates insufficient characterization of the posterior mode. Indeed, Fig. 10 shows that `GPry` underestimates the tails, especially in the $\mathcal{M}_c, \delta\mu$ direction which leads to the large discrepancy.

.

fore `GPry` shows great potential: it could turn days- or weeks-long inference runs with `nessai` into hours-long ones.

The injected parameters for the benchmark source are shown in Table III. For this high signal-to-noise case, the posterior is nearly Gaussian.

In this case, we generate 100 noise realizations (of which one was discarded due to an HPC error) and perform inference runs with `GPry` and `nessai`. The `nessai` runs were performed with 500 live points, instead of 2000

Parameter	Symbol	Value
Redshifted chirp mass	$\mathcal{M}_c$	$6.5744 \times 10^6\,\mathrm{M}_\odot$
Reduced mass-ratio	$\delta\mu$	0.12864
Luminosity distance	d_L	$18.7\,\mathrm{Gpc}$
Ecliptic longitude	λ	$2.15\,\mathrm{rad}$
Ecliptic sine-latitude	$\sin\beta$	$-0.34\,\mathrm{rad}$
cosine-inclination	$\cos\iota$	$0.86\,\mathrm{rad}$
Orbital phase	ϕ_0^{orb}	$5.86\,\mathrm{rad}$
Polarization	ψ	$-0.136\,\mathrm{rad}$
Time to merger	τ_m	$4.138 \times 10^6\,\mathrm{s}$ (47.89 days)
Dimensionless spin	χ_1	0.9874
Dimensionless spin	χ_2	0.9876
SNR		1944.8

TABLE III. Injected values for the sampled parameters of the SMBHB system and total source SNR. The detector-frame individual masses are $m_1 = 8.61 \times 10^6\,\mathrm{M}_\odot$ and $m_2 = 6.65 \times 10^6\,\mathrm{M}_\odot$. The reference frequency is $f_{\mathrm{ref}} = 10^{-4}\,\mathrm{Hz}$.

as for the other sources, for reasons of limited computational capacity. The resulting PP plot can be seen in Fig. 5a, and the relative JS divergence for each pair of runs in Fig. 5b. Both `GPry` and `nessai` show very good performance in the PP plot, and agree very well, with a median $D_{JS} \approx 0.05$ and no run with $D_{JS} \geq 0.1$. The runs with the median and highest D_{JS} are shown in Figs. 11 and 12, respectively. Therein we contrast the respective `GPry` runs with two higher-resolution (2000 live points) `nessai` runs performed to show finer contours for comparison.

For the SMBHB runs `GPry` needs $n < 10^3$ evaluations, which amounts to $\approx 30\%$ of the total computation time when the learning overhead is accounted for. In contrast, `nessai`, despite being run with a significantly low resolution for reasons of time, performs $n \sim 10^5$ evaluations.

D. Number of posterior evaluations and speedup

`GPry`'s main advantage compared to more traditional samplers is a drastic reduction in the number of posterior evaluations needed for inference, as shown in Fig. 6a. It is clear that `GPry` consistently performs $\mathcal{O}(10^2) - \mathcal{O}(10^3)$ fewer evaluations than `nessai` to converge to the posterior mode. This, however, comes at the price of the relatively large amount of time required for the acquisition of new optimal sampling locations and fitting the GPR hyperparameters. The size of this overhead depends mainly on the dimensionality of the sampling space and, to a lesser extent, on the Gaussianity of the posterior. The dimensionality scaling of the overhead can be clearly observed in Fig. 7 in Appendix B. Ultimately, the potential speedup with respect to an alternative sampler depends on a combination of a slow-enough posterior and a reasonable dimensionality.

In Fig. 6b we show a comparison between the distribution of wall-clock times of the `GPry` runs in this paper, and an optimistic (assuming no overhead) estimate
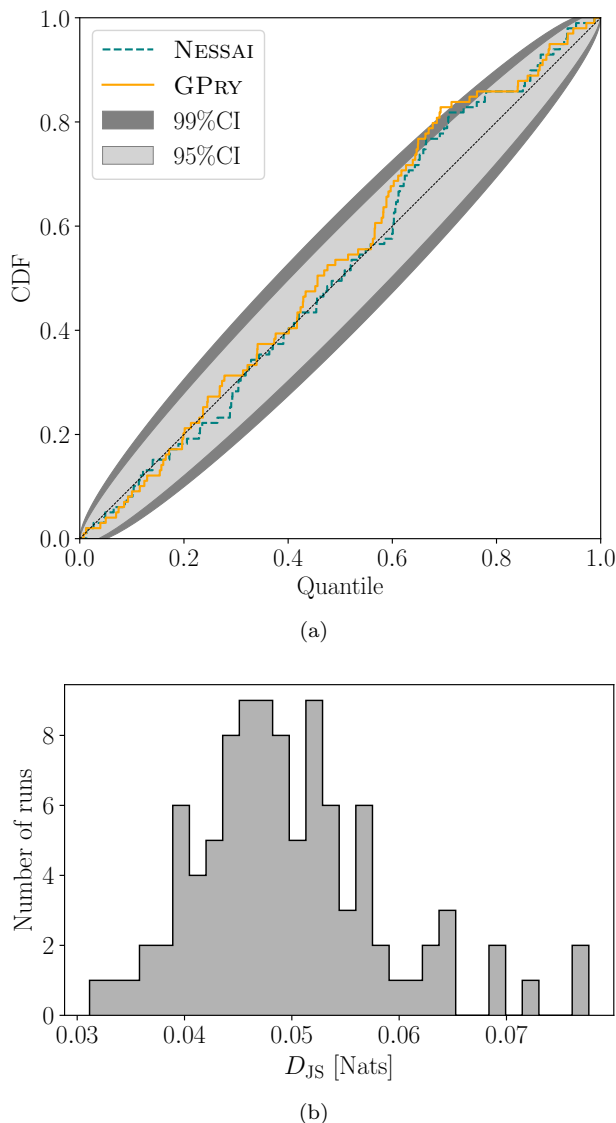
(a)

(b)

FIG. 5. PP plot **(a)** and Jensen-Shannon divergence **(b)** for 99 SMBHB noisy runs. The former shows that both `GPry` and `nessai` show comparable accuracy, consistent at 99% confidence (dark gray shaded area) with the theoretical prediction (black dotted line). The distribution of D_{JS} clusters around our target value of 0.05. This might partially be caused by `nessai` running at low resolution, but also by `GPry` occasionally under-exploring the tails of the posterior.

for the paired `nessai` runs. The quoted clock times are obtained by multiplying the number of their likelihood evaluations by their evaluation on the same hardware as for the `GPry` runs.[3] As we can see there, in the case of the DWD source `GPry` does not outperform `nessai`, needing roughly twice the time on average, due to the very short computation time of the DWD likelihood. However, for

---

[3] The `nessai` runs needed to be performed on a different platform due to limitations in our computing budget.

the stBHBs and SMBHBs, where likelihood computations are more expensive, the speed up is highly significant, reducing the time for inference from $\sim 10^6$ core seconds (around 11 days) to $\sim 10^4$ core seconds (around 3 hours). Of course, both of these numbers can be reduced through parallel processing but the advantage would still be evident.

When taking the reliability and accuracy of the inference into account, it is clear that the biggest potential for speed up is currently in the inference of SMBHBs.

V. CONCLUSIONS

We demonstrated that active sampling methods with Gaussian processes have the ability to produce accurate inference on individual injections of three different GW sources expected in the LISA band, DWDs, stBHBs and SMBHBs, employing $\mathcal{O}(10^{-2})$ fewer evaluations of the GW signal likelihood than a state-of-the-art nested sampler, and with a significant speedup, going up to a $\mathcal{O}(10^{-2})$ wall-clock time reduction for likelihood evaluation times approaching $\mathcal{O}(1\,\mathrm{s})$ and above. They do so with some, but little preconditioning, that can be provided by frequentist searches or other faster but less accurate approximate inference schemes. Crucially, no expensive pretraining is required with these methods as would be in amortized approaches.

Using `GPry` as an active learning framework, we found the advantages with respect to traditional Monte Carlo samplers to be problem-dependent:

- Inference for DWDs can be provided quickly and robustly, but the fast-to-evaluate DWD waveforms mean that the overhead of acquiring samples and fitting the Gaussian process outweigh the time saved by reducing the number of sampling steps. This in turn means that we report no savings in terms of wall-clock time. However, in the presence of gaps in the data, as expected in LISA, the computational cost of the likelihood will go up. In this case our approach could be competitive.

- Inferring the parameters for stBHBs was possible with considerable time savings of 2 orders of magnitude. However, this comes at the cost of underestimating the tails of the distribution, though we retain the ability to reliably recover the central values. There is ongoing development of `GPry` aimed at addressing this shortcoming.

- The best combination of speed-up and accuracy was achieved for the SMBHBs, whose likelihood is very slow to evaluate at $\mathcal{O}(1\,\mathrm{s})$. We report a speed-up of two orders of magnitude compared to nested sampling while retaining a comparable accuracy. This reduces the computational cost of the inference from $\sim 10^6$ core-seconds (~ 11 core-days) to merely $\sim 10^4$ core-seconds (~ 3 core-hours).
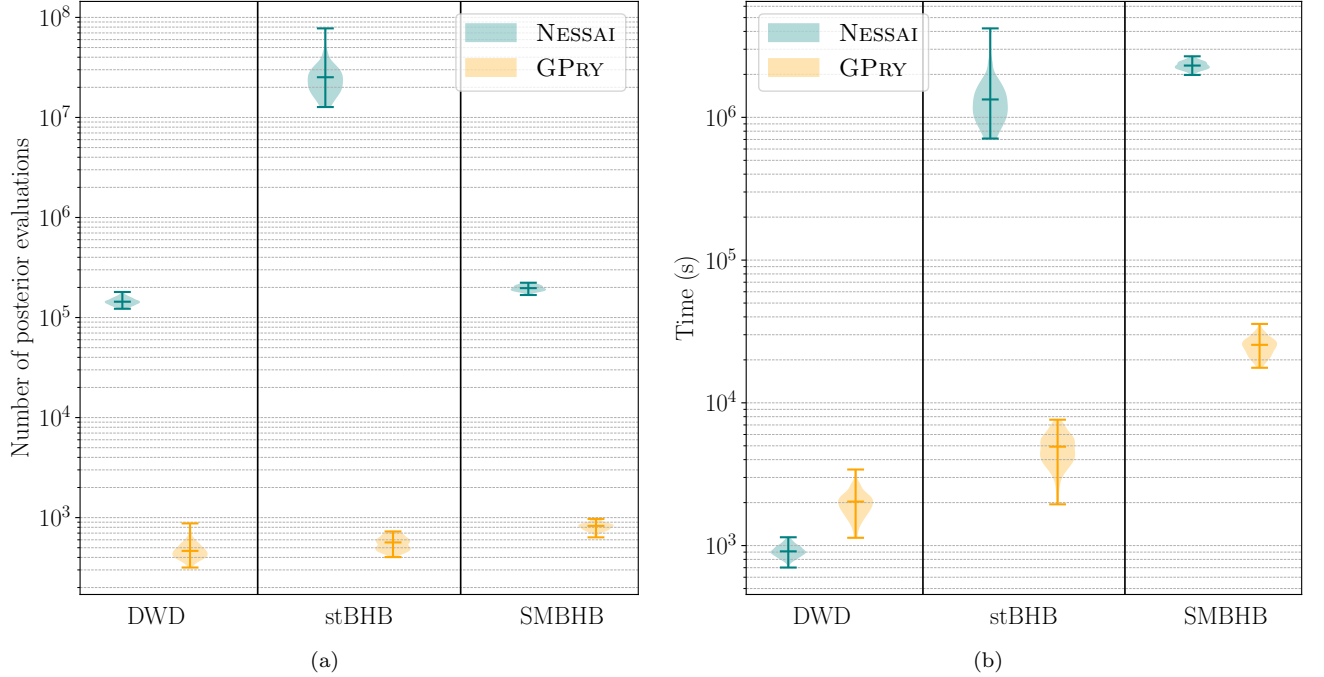
FIG. 6. Violin plots comparing `GPry` (orange) and `nessai` (teal) on each of the three test sources on noisy LISA data, according to **(a)** the total number of posterior evaluations needed, and **(b)** the hypothetical wall-clock time required for inference in a single-core setup (see text for a precise definition of these time estimates). The violins show the distribution the number of posterior samples and times, the minimum, median, and maximum values are marked with horizontal bars.

The integration of `GPry` (or a different active learning approach) into the LISA Global Fit pipeline would allow the characterization of expensive-likelihood signals (such as ones with strong time dependence) with low latency, by spawning it on their conditional likelihoods at any point. Even in cases in which `GPry` would be outperformed at inference time by amortized approaches (such as simulation-based inference), or if the calculation of waveforms or likelihoods are significantly accelerated, `GPry` opens the door to explore new physics (e.g., modified GR at emission or propagation) or characterize exotic signals, for neither of which a pre-trained emulator may be available or cost-effective.

`GPry` can also be a very powerful tool for prototyping waveforms, theoretical models, and the inference pipeline with mock data. It requires no pretraining, and no noise to be present in the data and accounted for in the likelihood. This enables quick forecasting and testing without the need for dedicated computing infrastructure to be in place.

The resulting surrogate posterior can be stored as kB-sized object, a size much smaller than the data necessary to reproduce the inference problem, and can be upsampled at very low computational cost. As an analytic function, it can be easily used as a prior in subsequent searches or constraints.

In the future, we aim to go beyond the results of this paper in parallel with the ongoing development of `GPry`, improving its accuracy in highly non-Gaussian and highly multimodal cases (e.g., extreme mass ratio inspirals), and its performance in larger dimensionalities such as inference problems with multiple sources.

ACKNOWLEDGMENTS

Software: We acknowledge usage of `Mathematica` [72] and of the following `Python` [73] packages for modeling, analysis, post-processing, and production of results throughout: `nessai` [74], `matplotlib` [75], `numpy` [76], `scipy` [77], `scikit-learn` [78], `Cobaya` [79] and `corner` [80].

[1] R. Abbott *et al.* (KAGRA, VIRGO, LIGO Scientific), Phys. Rev. X **13**, 041039 (2023), arXiv:2111.03606 [gr-qc].

[2] A. Afzal *et al.* (NANOGrav), Astrophys. J. Lett. **951**, L11 (2023), [Erratum: Astrophys.J.Lett. 971, L27 (2024), Erratum: Astrophys.J. 971, L27 (2024)], arXiv:2306.16219 [astro-ph.HE].

[3] J. Antoniadis *et al.* (EPTA, InPTA:), Astron. Astrophys. **678**, A50 (2023), arXiv:2306.16214 [astro-ph.HE].

[4] D. J. Reardon *et al.*, Astrophys. J. Lett. **951**, L6 (2023), arXiv:2306.16215 [astro-ph.HE].

[5] H. Xu *et al.*, Res. Astron. Astrophys. **23**, 075024 (2023), arXiv:2306.16216 [astro-ph.HE].

[6] P. Amaro-Seoane *et al.* (LISA), arXiv e-prints (2017), arXiv:1702.00786 [astro-ph.IM].

[7] P. Auclair *et al.* (LISA Cosmology Working Group), Living Rev. Rel. **26**, 5 (2023), arXiv:2204.05434 [astro-ph.CO].

[8] S. H. Strub, L. Ferraioli, C. Schmelzbach, S. C. Stähler, and D. Giardini, Phys. Rev. D **106**, 062003 (2022), arXiv:2204.04467 [astro-ph.IM].

[9] P. A. Seoane *et al.* (LISA), Living Rev. Rel. **26**, 2 (2023), arXiv:2203.06016 [gr-qc].

[10] N. Afshordi *et al.* (LISA Consortium Waveform Working Group), arXiv e-prints (2023), arXiv:2311.01300 [gr-qc].

[11] E. Cuoco *et al.*, Mach. Learn. Sci. Tech. **2**, 011002 (2021), arXiv:2005.03745 [astro-ph.HE].

[12] P. Canizares, S. E. Field, J. Gair, V. Raymond, R. Smith, and M. Tiglio, Phys. Rev. Lett. **114**, 071104 (2015), arXiv:1404.6284 [gr-qc].

[13] R. Smith, S. E. Field, K. Blackburn, C.-J. Haster, M. Pürrer, V. Raymond, and P. Schmidt, Phys. Rev. D **94**, 044031 (2016), arXiv:1604.08253 [gr-qc].

[14] S. E. Field, C. R. Galley, J. S. Hesthaven, J. Kaye, and M. Tiglio, Phys. Rev. X **4**, 031006 (2014), arXiv:1308.3565 [gr-qc].

[15] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, Phys. Rev. Lett. **127**, 241103 (2021), arXiv:2106.12594 [gr-qc].

[16] U. Bhardwaj, J. Alvey, B. K. Miller, S. Nissanke, and C. Weniger, Phys. Rev. D **108**, 042004 (2023), arXiv:2304.02035 [gr-qc].

[17] M. Andrés-Carcasona, M. Martinez, and L. M. Mir, Mon. Not. Roy. Astron. Soc. **527**, 2887 (2023), arXiv:2309.04303 [gr-qc].

[18] D. Chatterjee *et al.*, Mach. Learn. Sci. Tech. **5**, 045030 (2024), arXiv:2407.19048 [gr-qc].

[19] M. Dax, S. R. Green, J. Gair, N. Gupte, M. Pürrer, V. Raymond, J. Wildberger, J. H. Macke, A. Buonanno, and B. Schölkopf, Nature **639**, 49 (2025), arXiv:2407.09602 [gr-qc].

[20] V. Raymond, S. Al-Shammari, and A. Göttel, arXiv e-

[21] I. M. Vílchez and C. F. Sopuerta, arXiv e-prints (2024), arXiv:2406.00565 [gr-qc].

[22] H. Sun, H. Wang, and J. He, "Accelerating bayesian sampling for massive black hole binaries with prior constraints from conditional variational autoencoder," (2025), arXiv:2502.09266 [astro-ph.IM].

[23] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, Nature Phys. **18**, 112 (2022), arXiv:1909.06296 [astro-ph.IM].

[24] M. Dax, S. R. Green, J. Gair, M. Pürrer, J. Wildberger, J. H. Macke, A. Buonanno, and B. Schölkopf, Phys. Rev. Lett. **130**, 171403 (2023), arXiv:2210.05686 [gr-qc].

[25] J. EL Gammal, N. Schöneberg, J. Torrado, and C. Fidler, JCAP **10**, 021 (2023), arXiv:2211.02045 [astro-ph.CO].

[26] J. Torrado, N. Schöneberg, and J. El Gammal, "Parallelized acquisition for active learning using monte carlo sampling," (2023), arXiv:2305.19267 [stat.ML].

[27] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, Adaptive computation and machine learning (MIT Press, Cambridge, Mass., 2006) pp. XVIII, 248 S.

[28] M. Osborne, R. Garnett, Z. Ghahramani, D. K. Duvenaud, S. J. Roberts, and C. Rasmussen, in *Advances in Neural Information Processing Systems*, Vol. 25, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc., 2012) pp. 46–54.

[29] T. Gunter, M. Osborne, R. Garnett, P. Hennig, and S. Roberts, in *Advances in Neural Information Processing Systems 27* (Curran Associates, Inc., 2014) pp. 2789–2797.

[30] M. Georgousi, N. Karnesis, V. Korol, M. Pieroni, and N. Stergioulas, MNRAS **519**, 2552 (2023), arXiv:2204.07349 [astro-ph.GA].

[31] V. Korol, V. Belokurov, C. J. Moore, and S. Toonen, MNRAS **502**, L55 (2021), arXiv:2010.05918 [astro-ph.GA].

[32] M. Colpi, K. Danzmann, M. Hewitson, K. Holley-Bockelmann, and et al., arXiv e-prints , arXiv:2402.07571 (2024), arXiv:2402.07571 [astro-ph.CO].

[33] E. Finch, G. Bartolucci, D. Chucherko, B. G. Patterson, and et al., MNRAS **522**, 5358 (2023), arXiv:2210.10812 [astro-ph.SR].

[34] C. Cutler, Phys. Rev. D **57**, 7089 (1998), arXiv:gr-qc/9703068 [gr-qc].

[35] M. Katz, "mikekatz04/gbgpu: First official public release!" (2022).

[36] O. Burke, S. Marsat, J. R. Gair, and M. L. Katz, arXiv e-prints , arXiv:2502.17426 (2025), arXiv:2502.17426 [gr-qc].

[37] R. Buscicchio, J. Torrado, C. Caprini, G. Nardini, N. Karnesis, M. Pieroni, and A. Sesana, JCAP **01**, 084 (2025), arXiv:2410.18171 [astro-ph.HE].

[38] A. Klein, G. Pratten, R. Buscicchio, P. Schmidt, and et al., arXiv e-prints , arXiv:2204.03423 (2022), arXiv:2204.03423 [astro-ph.HE].

[39] A. Klein, arXiv e-prints , arXiv:2106.10291 (2021), arXiv:2106.10291 [gr-qc].

[40] G. Morras, G. Pratten, and P. Schmidt, arXiv e-prints , arXiv:2502.03929 (2025), arXiv:2502.03929 [gr-qc].

[41] G. Pratten, P. Schmidt, H. Middleton, and A. Vecchio, Phys. Rev. D **108**, 124045 (2023), arXiv:2307.13026 [gr-qc].

[42] P. C. Peters and J. Mathews, Phys. Rev. **131**, 435 (1963).

[43] C. García-Quirós, M. Colleoni, S. Husa, H. Estellés, and et al., Phys. Rev. D **102**, 064002 (2020), arXiv:2001.10914 [gr-qc].

[44] LISA Consortium Waveform Working Group, N. Afshordi, S. Akçay, P. Amaro Seoane, and et al., arXiv e-prints , arXiv:2311.01300 (2023), arXiv:2311.01300 [gr-qc].

[45] M. Tinto and S. V. Dhurandhar, Living Reviews in Relativity **8**, 4 (2005).

[46] R. Buscicchio, A. Klein, E. Roebber, C. J. Moore, and et al., Phys. Rev. D **104**, 044065 (2021), arXiv:2106.05259 [astro-ph.HE].

[47] L. J. Rubbo, N. J. Cornish, and O. Poujade, Phys. Rev. D **69**, 082003 (2004), arXiv:gr-qc/0311069 [gr-qc].

[48] G. Pratten, C. García-Quirós, M. Colleoni, A. Ramos-Buades, and et al., Phys. Rev. D **103**, 104056 (2021), arXiv:2004.06503 [gr-qc].

[49] M. Pürrer and C.-J. Haster, Physical Review Research **2**, 023151 (2020), arXiv:1912.10055 [gr-qc].

[50] LISA Science Study Team, *LISA Science Requirements Document*, Tech. Rep. 1.0 (ESA, 2018).

[51] N. Karnesis, S. Babak, M. Pieroni, N. Cornish, and T. Littenberg, Phys. Rev. D **104**, 043019 (2021), arXiv:2103.14598 [astro-ph.IM].

[52] H. Sun, H. Wang, and J. He, arXiv e-prints , arXiv:2502.09266 (2025), arXiv:2502.09266 [astro-ph.IM].

[53] D. Bandopadhyay and C. J. Moore, Physical Review D **108** (2023), 10.1103/physrevd.108.084014.

[54] D. Bandopadhyay and C. J. Moore, Physical Review D **110** (2024), 10.1103/physrevd.110.103026.

[55] G. Cabourn Davies, I. Harry, M. J. Williams, D. Bandopadhyay, and et al., Phys. Rev. D **111**, 043045 (2025), arXiv:2411.07020 [hep-ex].

[56] G. Pratten, A. Klein, C. J. Moore, H. Middleton, and et al., Phys. Rev. D **107**, 123026 (2023), arXiv:2212.02572 [gr-qc].

[57] L. Piro, M. Colpi, J. Aird, A. Mangiagli, and et al., MNRAS **521**, 2577 (2023), arXiv:2211.13759 [astro-ph.HE].

[58] A. Mangiagli, C. Caprini, M. Volonteri, S. Marsat, and et al., Phys. Rev. D **106**, 103017 (2022), arXiv:2207.10678 [astro-ph.HE].

[59] M. Vallisneri, M. Crisostomi, A. D. Johnson, and P. M. Meyers, arXiv e-prints (2024), arXiv:2405.08857 [gr-qc].

[60] J. El Gammal, N. Schöneberg, J. Torrado, and C. Fidler, "GPry: Bayesian inference of expensive likelihoods with Gaussian processes," Astrophysics Source Code Library, record ascl:2212.006 (2022), ascl:2212.006.

[61] W. J. Handley, M. P. Hobson, and A. N. Lasenby, Mon. Not. Roy. Astron. Soc. **450**, L61 (2015), arXiv:1502.01856 [astro-ph.CO].

[62] W. J. Handley, M. P. Hobson, and A. N. Lasenby, Mon. Not. Roy. Astron. Soc. **453**, 4384 (2015), arXiv:1506.00171 [astro-ph.IM].

[63] D. Ginsbourger, R. Le Riche, and L. Carraro, "Kriging is well-suited to parallelize optimization," (Springer, 2010) pp. 131–162.

[64] A. Lewis and S. Bridle, Phys. Rev. **D66**, 103511 (2002), arXiv:astro-ph/0205436 [astro-ph].

[65] A. Lewis, Phys. Rev. **D87**, 103529 (2013), arXiv:1304.4473 [astro-ph.CO].

[66] M. J. Williams, "nessai: Nested sampling with artificial intelligence," (2021).

[67] M. J. Williams, J. Veitch, and C. Messenger, Phys. Rev. D **103**, 103006 (2021), arXiv:2102.11056 [gr-qc].

[68] M. J. Williams, J. Veitch, and C. Messenger, Mach. Learn. Sci. Tech. **4**, 035011 (2023), arXiv:2302.08526 [astro-ph.IM].

[69] S. R. Cook, A. Gelman, and D. B. Rubin, Journal of Computational and Graphical Statistics **15**, 675 (2006).

[70] M. B. Wilk and R. Gnanadesikan, Biometrika **55**, 1 (1968).

[71] J. Lin, IEEE Transactions on Information Theory **37**, 145 (1991).

[72] Wolfram Research Inc., "Mathematica," (2022).

[73] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).

[74] M. J. Williams, J. Veitch, and C. Messenger, Phys. Rev. D **103**, 103006 (2021), arXiv:2102.11056 [gr-qc].

[75] J. D. Hunter, Computing in Science and Engineering **9**, 90 (2007).

[76] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, and et al., Nature **585**, 357 (2020), arXiv:2006.10256 [cs.MS].

[77] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, and et al., Nature Methods **17**, 261 (2020), arXiv:1907.10121 [cs.MS].

[78] F. Pedregosa *et al.*, Journal of Machine Learning Research **12**, 2825 (2011).

[79] J. Torrado and A. Lewis, JCAP **05**, 057 (2021), arXiv:2005.05290 [astro-ph.IM].

[80] D. Foreman-Mackey, The Journal of Open Source Software **1**, 24 (2016).

Appendix A: Approximate Jensen-Shannon divergence between multivariate Gaussians

The KL divergence D_{KL}, defined in Eq. (17), has an analytical representation when computed between two d-dimensional multivariate Gaussians, $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma_1})$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma_2})$, with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. Conversely, an analytical representation for the JS divergence, defined in Eq. (16), does not exist in this case. However, we can find an approximate expression if the mixture distribution M (with mean $\boldsymbol{\mu}_M = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$) is a multivariate normal distribution with covariance $\boldsymbol{\Sigma}_M = \frac{1}{2}(\sigma_1^2 + \sigma_2^2)\mathbb{I}$. The approximation holds if the multivariate Gaussians are sufficiently similar, i.e. $|\Delta\boldsymbol{\mu}| \equiv |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \ll |\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2|$ and $\sigma_1 \approx \sigma_2$. In this case, the JS divergence reads

$$D_{\text{JS}}[\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)] \approx \frac{1}{4}\Delta\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\Delta\boldsymbol{\mu} + \frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}}\right) - \frac{d}{2}\log 2 \ . \tag{A1}$$

For $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \sigma^2\mathbb{I}$ this simplifies to

$$D_{\text{JS}} \approx \frac{(\Delta\boldsymbol{\mu})^2}{8\sigma^2} \ , \tag{A2}$$

and for $|\Delta\boldsymbol{\mu}| = 0$, $\boldsymbol{\Sigma}_1 = \sigma_1^2\mathbb{I}$, $\boldsymbol{\Sigma}_2 = \sigma_2^2\mathbb{I}$ to

$$D_{\text{JS}} \approx \frac{d}{2}\log\frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2} \ . \tag{A3}$$

Therefore, when only the mean of the distribution is misestimated, D_{JS} is a function of the distance $\Delta\boldsymbol{\mu}$ in units of σ (see Eq. (A2)). This is independent of the number of dimensions if only one parameter is misestimated whereas it is proportional to d if the misestimation occurs for multiple parameters. On the other hand, if the mean is properly estimated but the spread of the distribution is not, then the result is proportional to d whenever, on average, the spread is wrong by the same amount (see Eq. (A3)). We find that in less than 11 dimensions (the highest number of inferred parameters that we consider in this paper), the approximation holds up to $D_{\text{JS}} \approx 0.1$.

Appendix B: Hyperparameters and overhead of GPry

Table IV shows the values of the `GPry` settings adapted to this study, as motivated in Sec. III C. An in-depth explanation of the meaning of each setting can be found in `GPry`'s documentation[4]. In Fig. 7 we show a breakdown of the computation costs of the `GPry` runs into posterior evaluation time and overhead from the two computationally expensive steps of the Bayesian optimization loop.

Setting	Description	DWD	stBHB	SMBHB
`noise_level`	Expected level of numerical noise	0.1	4	2
`inf_threshold`	Cutoff in log-posterior of the SVM classifier	20σ	10σ	30σ
`fit_full_every`	Number of iterations between GPR hyperparameter optimizations	2	2	1
`n_restarts_optimizer`	Number of restarts per GPR hyperparameter optimization	$2d$	d	$4d$
`mc_every`	Number of iterations between NS runs to generate proposals	5	3	3
`n_points_per_acq`	Number of Kriging steps determining the proposal batch size	7	9	8
`trust_region_nstd`	Cutoff in log-posterior for defining the trust region	—	3σ	3σ
`trust_region_factor`	Enlargement factor of the trust region	—	2.5	2.5

TABLE IV. Non-default settings for `GPry`, as discussed in Sec. III C. In the parameter values, a number followed by d is multiplied by the dimensionality of the problem, whereas one followed by σ represents the difference between the log-posterior of the cutoff and that of the best training sample as the equivalent number of 1-dimensional standard deviations, i.e. 2σ represents the log-posterior difference from the top of the distribution of a d-dimensional Gaussian that leaves 95% of the mass above it.
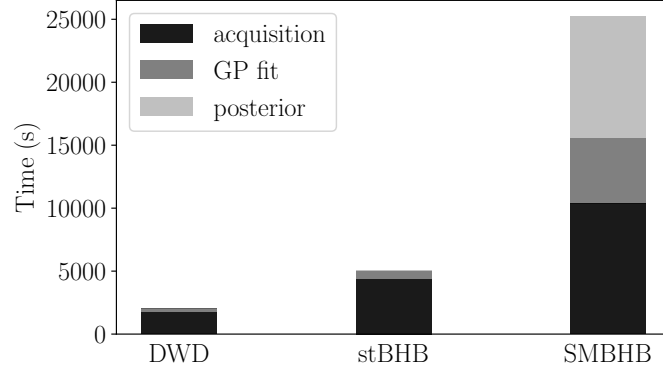
---

[4] https://gpry.readthedocs.io

FIG. 7. Graph showing the dominant part of the overhead (acquisition and hyperparameter fits of the GPR) vs the total time spent on evaluating the log-posterior. Sub-dominant or non-necessary contributions to `GPry`'s overhead have been omitted such as determining convergence, checkpointing and the generation of the final MC sample, which typically add up to a few seconds. For the relatively fast to evaluate posteriors of the DWDs and stBHBs the runtime is dominated by `GPry`'s overhead which – due to the much lower number of posterior evaluations – still leads to a speedup over `nessai` in the case of the stBHBs and SMBHBs. For the slow SMBHB posterior, despite only roughly 1/3 of the time being spent on posterior evaluations, the large reduction in their number with respect to `nessai` still leads to a significant speedup (see Fig. 6b).

Appendix C: Corner plots

In this appendix, we show some corner plots for the sources studied in Sec. IV. In the upper part of each plot we furthermore show the sampling locations of `GPry`, omitting the samples that are far away from the mode (typically a few percent). The contours for `GPry` are obtained by sampling the surrogate model with an MCMC.
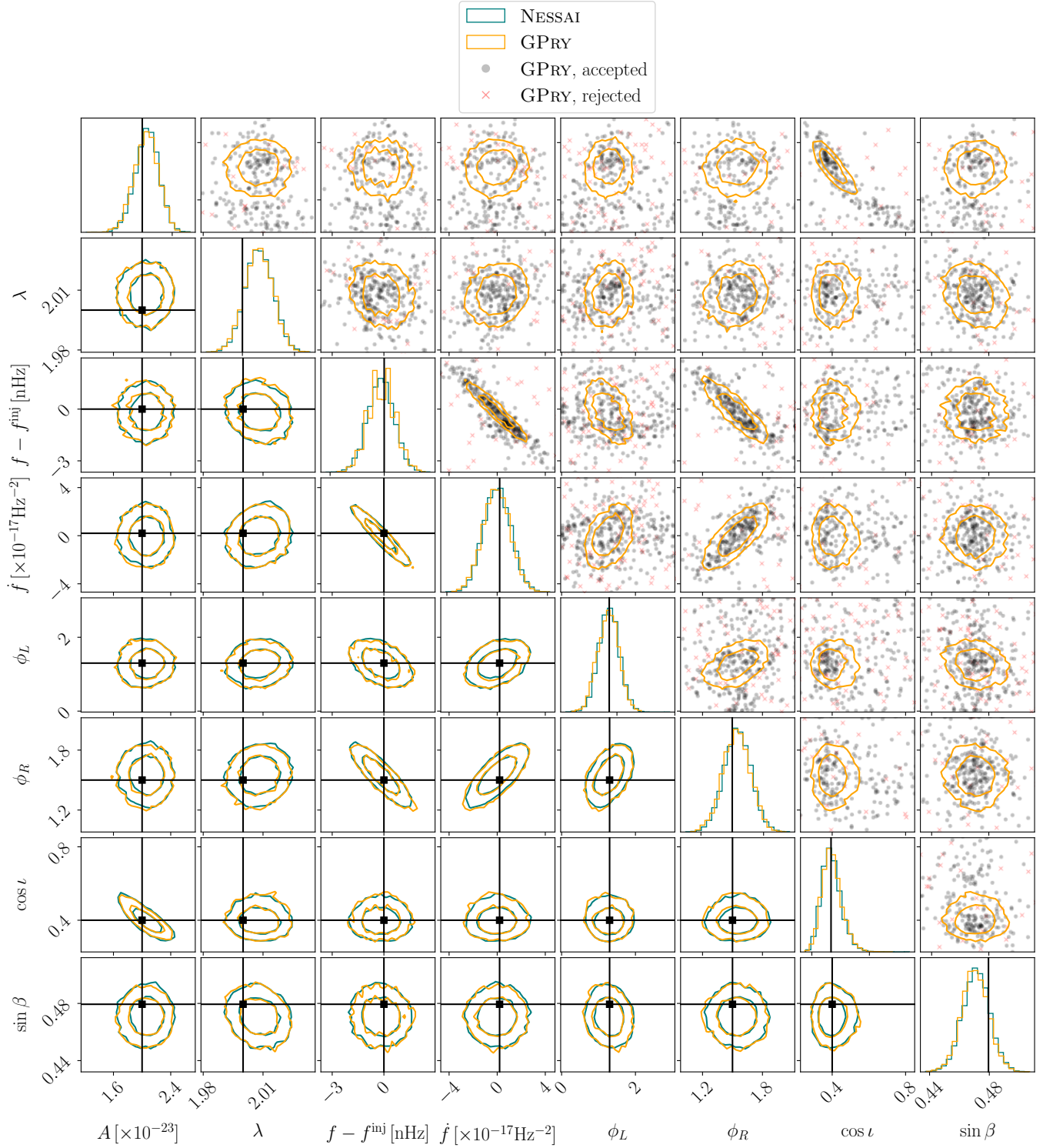
FIG. 8. Corner plot comparing `nessai` to `GPry` inference on the DWD source with the parameters specified in Table I for the run with the median JS divergence $D_{\rm JS} = 0.0043$ (see Fig. 3b for the distribution). The number of likelihood evaluations for `GPry` was ≈ 500 (shown in the upper triangle, missing a few percent that would fall outside the ranges of the plot), and for `nessai` it was ≈ 136500. The 2d contour levels show the 68% and 95% CL constraints. On the upper triangular we show the locations where `GPry` has evaluated the true posterior distribution. The gray dots represent accepted samples (samples that are used to train the GPR), while the red crosses are rejected (used to train the SVM classifier).
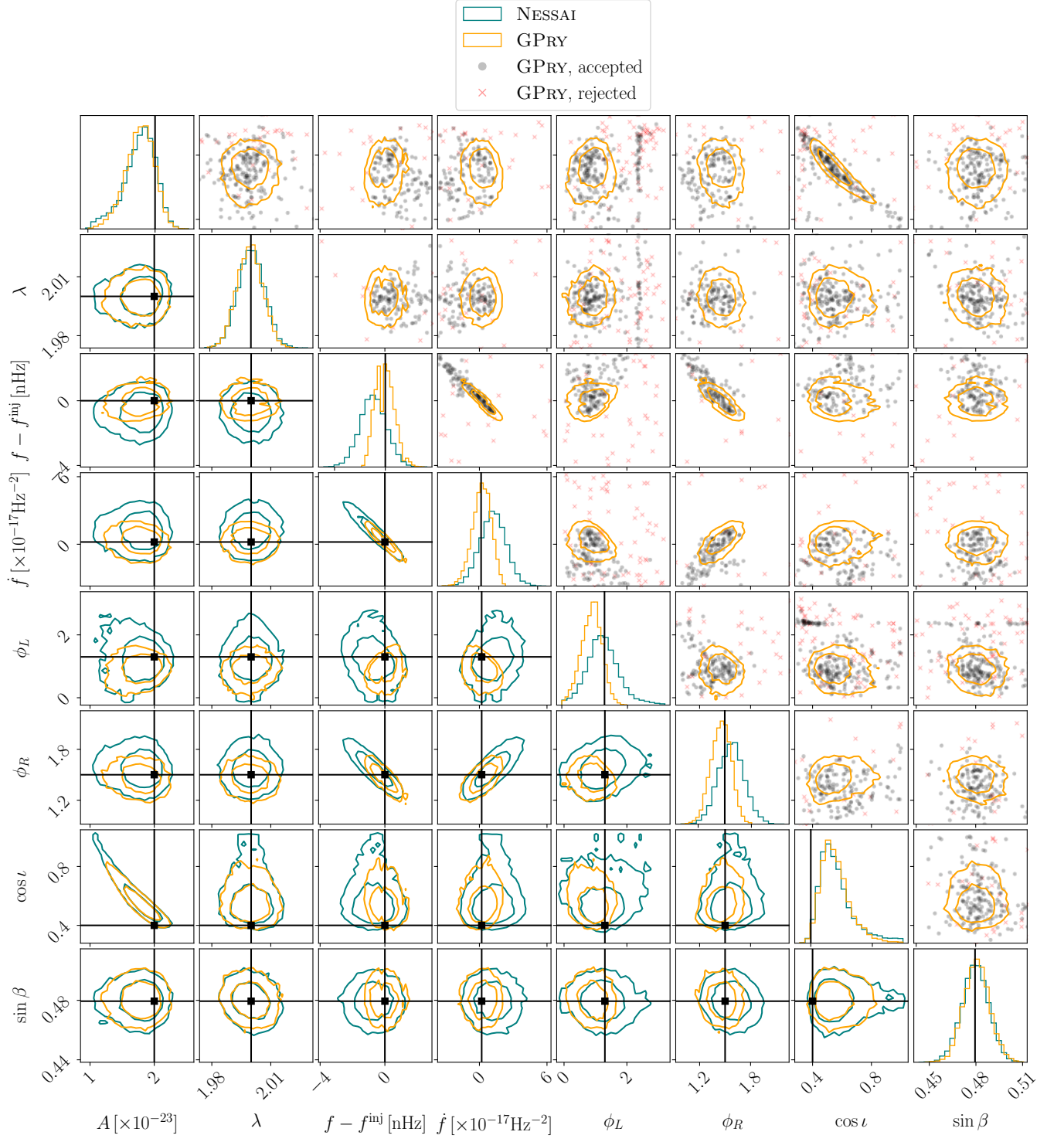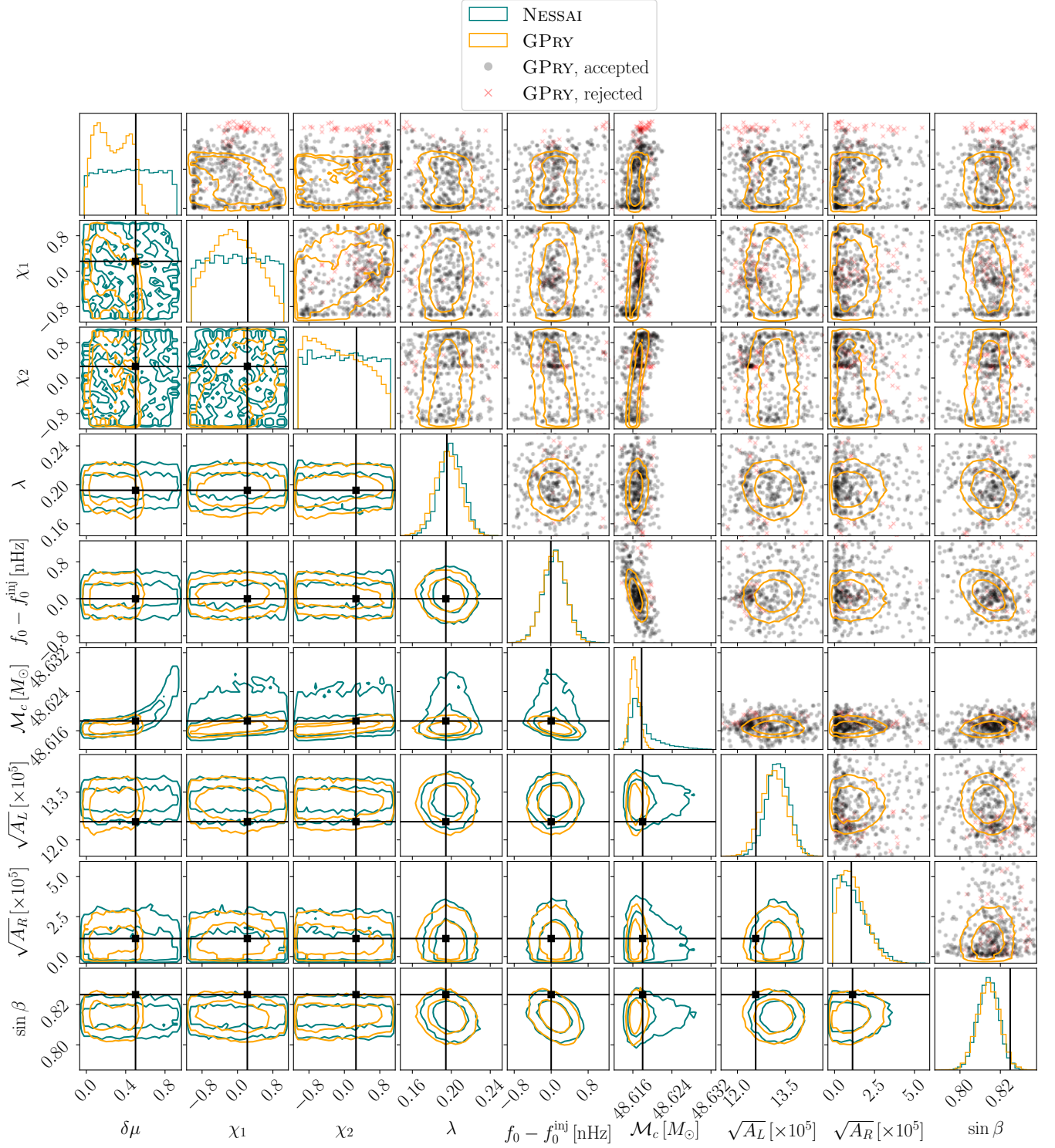
FIG. 9. Same as Fig. 8, comparing `nessai` to `GPry` inference on the DWD source with the parameters specified in Table I for the run with the highest JS divergence $D_{\text{JS}} = 0.12$ (see Fig. 3b for the distribution). The number of likelihood evaluations for `GPry` was ≈ 350 (shown in the upper triangle), and for `nessai` it was ≈ 146000.
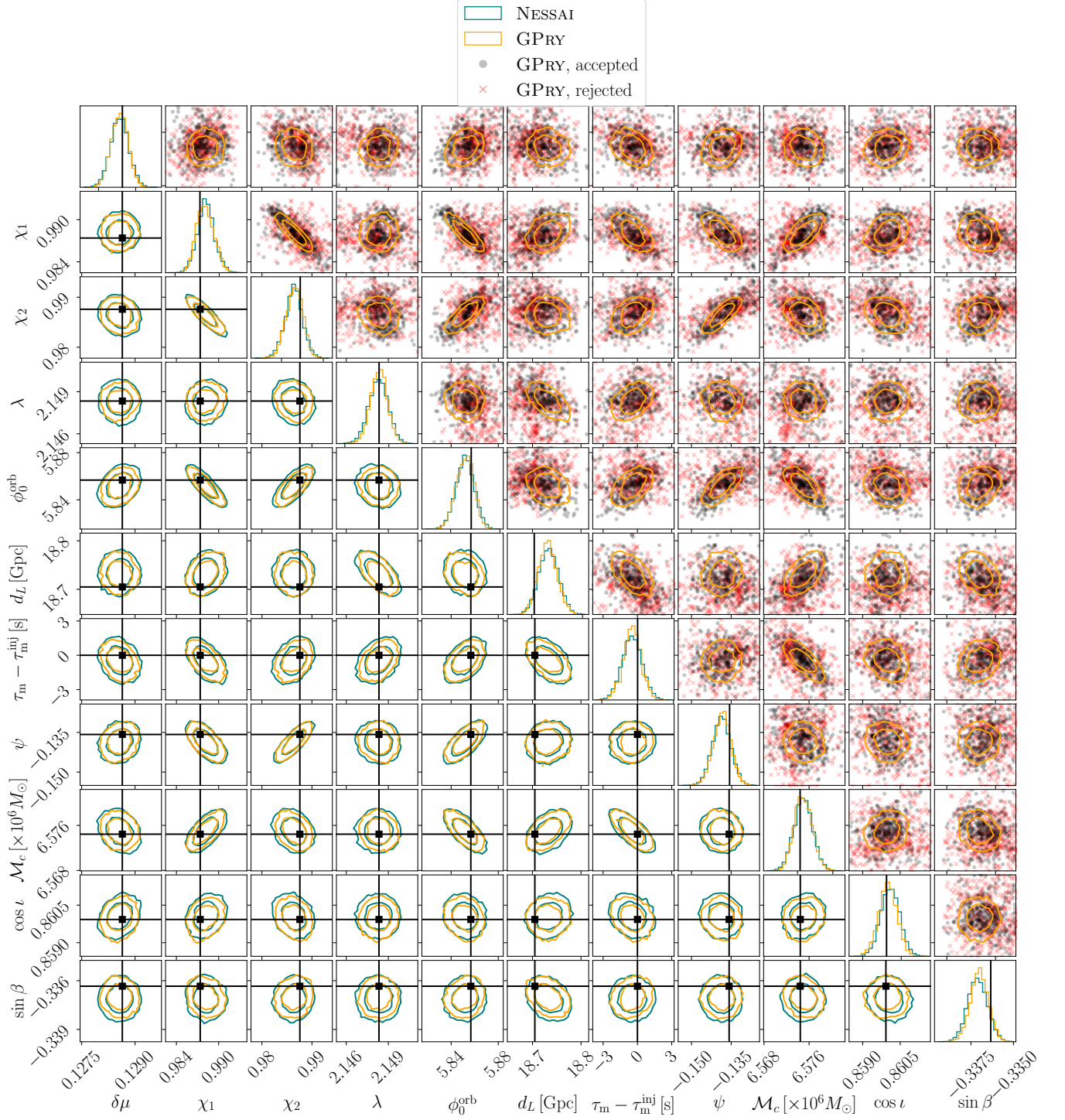
FIG. 10. Same as Fig. 8, comparing `nessai` to `GPry` inference on the stBHB source with the parameters specified in Table II for the run with the median JS divergence $D_{\mathrm{JS}} = 0.19$ (see Fig. 4b for the distribution). The number of likelihood evaluations for `GPry` was ≈ 450, and for `nessai` it was ≈ 23484500.

FIG. 11. Same as Fig. 8, comparing `nessai` to `GPry` inference on the SMBHB source with the parameters specified in Table III for the run with the median JS divergence $D_{\mathrm{JS}} = 0.048$ (see Fig. 5b for the distribution). The number of likelihood evaluations for `GPry` was ≈ 850, and for `nessai` it was ≈ 207000 with 500 live points (used for the D_{JS} calculation and the PP plot), and ≈ 567000 for the high resolution run (2000 live points) whose contours are shown.
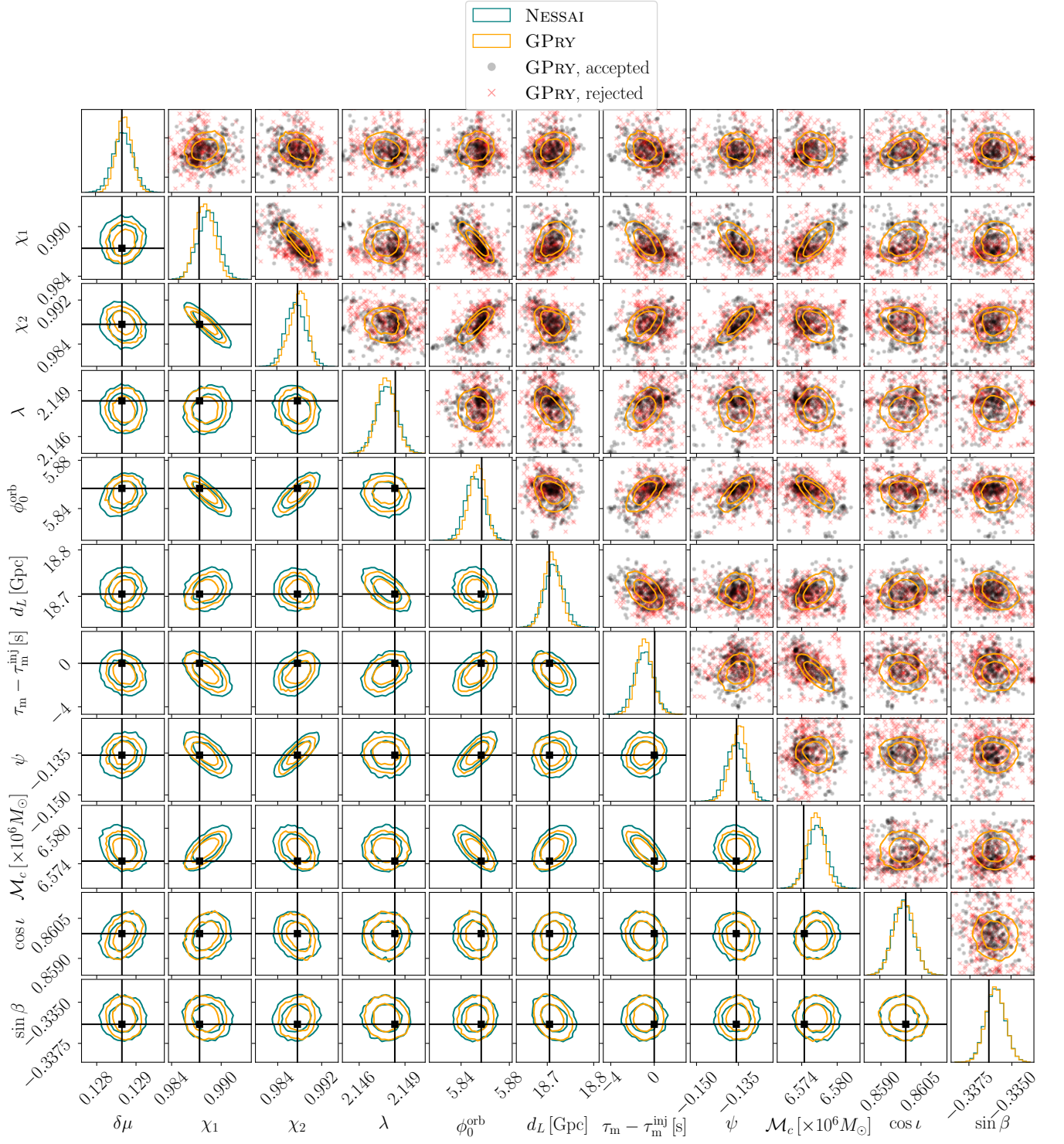
FIG. 12. Same as Fig. 8, comparing `nessai` to `GPry` inference on the SMBHB source with the parameters specified in Table III for the run with the highest JS divergence $D_{\mathrm{JS}} = 0.078$ (see Fig. 5b for the distribution). The number of likelihood evaluations for `GPry` was ≈ 650, and for `nessai` it was ≈ 198500 at low resolution (500 live points), ≈ 364000 at high resolution (2000 live points), whose contours are shown.