Simulation-based inference for stochastic gravitational wave background data analysis

James Alvey,[1,*] Uddipta Bhardwaj,[2,†] Valerie Domcke,[3,‡] Mauro Pieroni,[3,§] and Christoph Weniger[1,‖]

[1]*GRAPPA Institute, Institute for Theoretical Physics Amsterdam, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*

[2]*GRAPPA Institute, Anton Pannekoek Institute for Astronomy and Institute of High-Energy Physics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*

[3]*Theoretical Physics Department, CERN, 1211 Geneva 23, Switzerland*

(Received 5 October 2023; revised 6 December 2023; accepted 5 March 2024; published 8 April 2024)

The next generation of space- and ground-based facilities promise to reveal an entirely new picture of the gravitational wave sky: thousands of galactic and extragalactic binary signals, as well as stochastic gravitational wave backgrounds (SGWBs) of unresolved astrophysical and possibly cosmological signals. These will need to be disentangled to achieve the scientific goals of experiments such as LISA, Einstein Telescope, or Cosmic Explorer. We focus on one particular aspect of this challenge: reconstructing an SGWB from (mock) LISA data. We demonstrate that simulation-based inference (SBI)—specifically truncated marginal neural ratio estimation (TMNRE)—is a promising avenue to overcome some of the technical difficulties and compromises necessary when applying more traditional methods such as Monte Carlo Markov Chains (MCMC). To highlight this, we show that we can reproduce results from traditional methods both for a template-based and agnostic search for an SGWB. Moreover, as a demonstration of the rich potential of SBI, we consider the injection of a population of low signal-to-noise ratio supermassive black hole transient signals into the data. TMNRE can implicitly marginalize over this complicated parameter space, enabling us to directly and accurately reconstruct the stochastic (and instrumental noise) contributions. We publicly release our TMNRE implementation in the form of the code SAQQARA.

DOI: 10.1103/PhysRevD.109.083008

I. INTRODUCTION

With the detection of gravitational waves (GWs) by the LIGO-Virgo-KAGRA (LVK) collaboration [1], and more recently by pulsar timing arrays (PTAs) [2–4], GW astronomy has now entered the stage as a new player to explore our Universe. While existing GW observatories are noise dominated with transient signals that are relatively sparse and can be described by a few parameters, the situation will drastically change with the next generation of GW observatories. Both the space-based interferometer LISA [5] and the next-generation ground-based interferometers, such as the Einstein Telescope [6] or Cosmic Explorer [7], are expected to see thousands of binary systems as well as the stochastic gravitational wave background (SGWB) of unresolved signals. This "orchestra" of overlapping signals poses a severe data analysis challenge to successful parameter reconstruction and component separation [8].

Within this global analysis, the accurate reconstruction of an SGWB of cosmological origin is particularly challenging [9]—but also offers a unique window to probe particle physics at energy scales far beyond the reach of colliders [10]. Taking LISA as an example, the difficulties are quickly identified: we lack the possibility of carrying out cross-correlation (as in LVK or PTAs), and there are no perfect null-channels [11–13] or the possibility to "shield" GWs, so the instrumental noise cannot be measured independently of a possible SGWB signal. Combined with the variety and complexity of possible particle physics models that lead to SGWBs, the accurate reconstruction of signal and noise parameters quickly becomes a highly challenging data analysis task.

A range of recent work has taken up this challenge using traditional inference techniques such as Markov Chain Monte Carlo (MCMC). The goal of the so-called LISA "global fit" [8,14–16] is to simultaneously fit waveforms for different types of binaries and noise components. Focusing on LISA, and assuming that all individual above-threshold sources are removed, several approaches have demonstrated the possibility of achieving the simultaneous reconstruction of SGWB and noise. These rely on templates either for the noise [17–20], the SGWB

*j.b.g.alvey@uva.nl
†u.bhardwaj@uva.nl
‡valerie.domcke@cern.ch
§mauro.pieroni@cern.ch
‖c.weniger@uva.nl

signal [21,22], or both [13,23]. The overall challenge is the sheer dimensionality of the problem (in principle, there will likely be at least $\sim 10^5$ parameters in the full problem, see, e.g., Refs. [8,16]), and the high precision reconstruction required to extract an SGWB signal. In a very broad sense, the goal of our work is to argue that simulation-based inference (SBI) techniques may be a promising path toward mitigating these traditionally conflicting goals of precision and scale. Moreover, an SBI pipeline could offer an independent cross-check to validate the results obtained with traditional methods.

SBI techniques (see, e.g., Ref. [24] for a review) have recently undergone a significant up tick in popularity as we approach a new era of big data analysis challenges. In contrast to stochastic-sampling approaches such as MCMC or nested sampling, SBI algorithms look to solve the Bayesian inference problem of reconstructing the posterior distribution, $p(\boldsymbol{\theta}|\boldsymbol{x})$, without requiring an explicit expression for the likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$ (although there are additional, independent benefits including, e.g., amortization, scalability, and simulation efficiency). Instead, the likelihood distribution is sampled implicitly via some stochastic forward simulator that generates data $\boldsymbol{x}$ from parameters $\boldsymbol{\theta}$. This fundamentally shifts the focus from building a statistical model to developing a realistic computational forward model for the data, including e.g. all relevant instrumental and physical effects. SBI methods have now been shown to perform inference to the level of a full likelihood-based approach in several astrophysical and cosmological settings, including GW data analysis, see e.g. [25–34]. While several SBI algorithms exist, see Refs. [35–40], in this work, we will focus on the application of (truncated marginal) neural ratio estimation (TMNRE) [41], implemented within the code SWYFT [42].

Several crucial benefits suggest the TMNRE algorithm could be an ideal tool for LISA SGWB data analysis. First, the truncation aspect (which makes TMNRE a sequential SBI algorithm) allows us to effectively "zoom-in" on the relevant regions of parameter space for a given observation (see Refs. [27,42] for details on this procedure). In a variety of cases, such as for cosmic microwave background data [43], strong lensing image analysis [30], and GWs from compact binary coalescences [25,27], this makes TMNRE extremely simulation efficient compared to both traditional methods and nonsequential SBI algorithms. Indeed, Ref. [27] demonstrated that analysing LIGO-type binary black hole mergers with TMNRE requires 98% fewer waveform evaluations than the currently adopted nested sampling approach. Second, the TMNRE algorithm can target specific parts of the model while implicitly marginalizing over all other components [44]. We will highlight this property in our analysis and demonstrate that we can directly analyze only the noise and SGWB components, properly marginalized over additional transient sources. Third, realistic LISA data will contain numerous GW signals (binaries and SGWBs) and instrumental noise. The methodologies and pipelines for the forward modeling of GW waveforms, known individual noise components, and instrument response functions are, up to technical refinements, well established for LISA. Conversely, the complexity of the problem could prohibit an explicit, exact expression for these marginalized likelihoods, rendering the inverse problem of parameter estimation potentially very costly, since one would have to work with the full likelihood. The implicit likelihood benefits that are at the core of SBI can circumvent these difficulties.

Ultimately, our proposed use-case for this algorithm has the same spirit in mind as the "global fit" [16]: separating the multiple components in a LISA data stream. As such, if possible, it is useful to split these into distinct analysis "blocks" to combat the dimensionality issues and consistently pass the subsequent inference results around the full model. Here, we illustrate how one could do this for the block containing the SGWB and noise components, accounting for, e.g., the presence of transient sources. In this regard, our analysis should be seen as a first step, dealing with a setup that, in many ways, is simplified compared to the data analysis challenge that LISA will face. However, this proof of principle and verification against other methods is a key step to unlocking the potential of SBI for GW data analysis.

II. ANALYSIS SETUP AND DATA GENERATION

To explore the ability of SBI—and more specifically TMNRE—to address the challenges of SGWB analysis, we set up several case studies to analyse. These are broadly similar to those presented in Ref. [20] and cover both the recovery of a signal given a parametrized template, as well as the agnostic fitting of an unknown signal. We then extend the analysis to include transient signals for a mock population of supermassive black holes to investigate the implications at the level of parameter estimation. Given this, there are several technical components to setting up the analysis: a model for the instrument noise in LISA; characterization of the SGWB templates; explanation of the transient setup; and data generation.

Considering the instrument noise first, currently, the knowledge of the LISA noise comes from LISA Pathfinder (LPF) [45], which tested the purity of free-fall for the test masses (TM), and from on-ground experiments. A two-parameter noise model, specified in terms of low-frequency TM noise and high-frequency optical metrology system (OMS) noise, defines a reasonable approximation of the LISA noise [5,46]. Each noise component depends quadratically on a parameter (referred to as A and P, respectively, and with fiducial values $A = 3$, $P = 15$), which controls its amplitude. For more details on the LISA noise model and the measurements LISA will perform, see Sec. A

in the Appendices. Consistent with Refs. [13,18–20,46], in the present work, we make the somewhat strong assumption that the noise *shapes* are perfectly known while we allow for the amplitude to vary within a wide, uniform prior for the noise parameters (assumed to be positive) centered around the fiducial values. However, we stress that tackling realistic LISA data analysis will require more complex noise modeling, which we leave to future work.

As far as the SGWB signal itself is concerned, we consider two types of templates: a power law (PL) specified by a tilt (γ) and (log) amplitude (α); and a more agnostic form defined by a (log) amplitude in the first bin α_1 and a sequence of slopes γ_j for $j = 1, \ldots, N_{\rm bins}$, where $N_{\rm bins}$ is the number of equally spaced logarithmic bins that the template is split into. In this work, we consider $N_{\rm bins} \leq 10$, though we note that Ref. [47] recently demonstrated the possibility of scaling to a larger ($\sim$20) number of bins. More concretely, we specify the templates (written in terms of the GW energy density $\Omega_{\rm GW}h^2$) as

$$\textbf{power law}:\ \Omega_{\rm GW}(f)h^2 = 10^{\alpha}\left(\frac{f}{\sqrt{f_{\rm min}f_{\rm max}}}\right)^{\gamma}$$

$$\textbf{agnostic}:\ \Omega_{\rm GW}(f)h^2 = \sum_{i=1}^{N_{\rm bins}} 10^{\alpha_i}\left(\frac{f}{\sqrt{f_{{\rm min},i}f_{{\rm max},i}}}\right)^{\gamma_i} \times \Theta(f - f_{{\rm min},i})\Theta(f_{{\rm max},i} - f)$$

where $f_{\rm min} = 10^{-4}$ Hz, $f_{\rm max} = 5 \times 10^{-2}$ Hz, $f_{\rm min/max,i}$ are the boundaries of each of the bins, Θ denotes the Heaviside function, and α_i and γ_i are the amplitude and tilt in each bin. In the analysis, we will vary each of the corresponding parameters uniformly in prior ranges specified in Table I. Imposing continuity, this fixes all the α_i for $i \geq 2$.

Beyond this, there are several assumptions that we make in our data generation. First, we work with a single time-delay interferometry (TDI) channel. Moreover, we model the time domain data $d(t)$ as a superposition of one (or more) signal component(s) $s_c(t)$, and detector noise $n(t)$ as $d(t) = n(t) + \sum_c s_c(t)$. We assume stationarity in the time-domain data,[1] which implies vanishing correlations between different frequencies in the Fourier domain. We consider mock data corresponding to an observation time T_d of 12 days.[2] For data compression, we divide it into

TABLE I. Prior choices for the case studies and analyses presented in this work for the (dimensionless) amplitudes, tilts, and instrumental noise parameters.

Parameter	Prior choice
(Log) Amplitudes, α, α_1	$\mathrm{U}(-20,-5)$
Tilts, γ, γ_i	$\mathrm{U}(-10,10)$
TM noise, A	$\mathrm{U}(0,6)$
OMS noise, P	$\mathrm{U}(0,30)$

$N_d = 100$ data segments of duration $T_s \equiv T_d/N_d$ each. In this scheme, the frequency resolution in each data segment is $\Delta f = 1/T_s = 10^{-4}$ Hz, and we denote with $\tilde{d}_s(f_{\rm k})$ the frequency-domain data for each segment s and (discrete) frequency $f_{\rm k}$. We also assume signal and noise to be Gaussian distributions with zero mean and variances based on their respective power spectral densities (PSDs). Under these assumptions, we generate N_d statistical realizations of the SGWB signal and noise. In the final analysis, we also investigate the implications of introducing a population of transient signals. To do so, we introduce a probability p that a given data segment contains a transient (this could easily be extended to include multiple transients). For each data segment, we inject a mock supermassive-black hole waveform with probability p.[3] We use the `IMRPhenomXAS` waveform template implemented in the `jax`-based `ripple` package [48] to generate the frequency-domain strains for this signal component. For the explicit choices of the population parameters see the Appendices.

Finally, there are several details that are relevant for constructing the comparison to the MCMC method. Specifically, following the approach introduced in [18,20], we define averaged data $\bar{D}_{\rm k} \equiv \sum_s \tilde{d}_s(f_{\rm k})\tilde{d}^*_s(f_{\rm k})/N_d$ and down-sample it through coarse-graining. This yields a new (binned) dataset $D_{\hat{k}}$, where $\hat{k}$ covers a sparser set of frequencies $f_{\hat{k}}$, and comes with weights $w_{\hat{k}}$ [18,20]. An unbiased (log-) likelihood for the compressed dataset can be built, e.g., as a mixture of a Gaussian and a log-normal component [20]. For the explicit expression see Sec. A in the Appendices. To sample the parameter space, we use the EMCEE sampler described in Ref. [49]. Given this setup, we can define the four benchmark analyses that we present the results for below:

C1: PL template and LISA noise.
C2: Agnostic template with 5 bins and LISA noise.
C3: Agnostic template with 10 bins and LISA noise.
C4: PL template, LISA noise, and additional transients.

[1]The assumption of stationarity does not apply to transient sources, or if so, only statistically.

[2]This corresponds to about 1/100th of the planned LISA observation time, however, it suffices to demonstrate our main points regarding the statistical flexibility and precision agreement of our algorithm. In addition, though, we explicitly tested the pipeline with $\Delta f = 10^{-5}$ Hz, or about 115 days of data split into 100 segments, and achieved similarly good agreement with MCMC. We leave the task of optimally scaling this up (e.g. via a similar coarse-graining scheme to Ref. [20]) to the full length of the LISA data—including importantly nonstationary noise components—for future work.

[3]Specifically we take $p = 0.05$, which on average would introduce 5 sources.

III. SIMULATION-BASED INFERENCE FRAMEWORK

The implementation (as presented in SAQQARA) of the TMNRE [42] algorithm in the context of SGWB recovery is one of the key results of this work. As such, we devote this short subsection to a description of some of the specific design choices relevant to SGWB analysis. We do not cover detailed explanations of the algorithm but instead refer the reader to, e.g., Ref. [42] for the initial presentation of TMNRE. In addition, we point the reader to Sec. 2 of Ref. [27] for a detailed description of the application of TMNRE to GWs from compact binary coalescence sources, outlining a lot of the logic we also follow in this work. Finally, the implementation of the autoregressive ratio estimation used to explore the parameter space in the final stage of inference is described in Ref. [44].

To understand where the design choices are made, it is useful to think of the ratio estimation step in TMNRE in two parts: compression, and ratio estimation. In particular, to implement the TMNRE algorithm, one must design a network architecture that can take in data $\boldsymbol{x}$ and parameters $\boldsymbol{\theta}$ and estimate the ratio $r(\boldsymbol{x},\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})/p(\boldsymbol{x})$, where $p(\boldsymbol{x})$ is the (Bayesian) model evidence. In practice, $\boldsymbol{x}$ is usually (or at least can be) a high-dimensional and complicated data structure, so we normally first define a compression network that compresses x to some lower-dimensional summary $s(\boldsymbol{x})$. This summary $s(\boldsymbol{x})$ is then combined with $\boldsymbol{\theta}$ and inputted into the standard ratio estimators implemented within SWYFT [41]. Importantly, both the summary and the ratio estimators are optimized simultaneously, resulting in an automatically learned summary statistic. As such, for various applications, the main novelty in the implementation lies in the design of the compression network that produces the summary $s(\boldsymbol{x})$. In the current context, we make use of the following structure. Starting with SGWB data $\boldsymbol{x}$ that is a time-averaged sequence of frequency bins, as in Fig. 1, we first perform a normalization step by taking the natural logarithm and then applying an online normalizing layer [41]. This is purely for performance reasons in the sense that the network optimization proceeds more robustly on normalized data. In the next stage, we do consider the structure of the data, however, and utilise an architecture similar to that described in Ref. [27]. In particular, we apply a set of 1-dimensional convolutional layers, organized into a UNet architecture [50]. The motivation for this is to allow both the local and global sharing of information across the various frequency bins as we look to learn the SGWB and noise templates. Scalability to more complex situations also guided this part of our architecture, allowing for direct application to either multiple channels or correlated noise without any modifications. Finally, we added a simple linear compression network to summarise this information into a lower dimensional vector that can then be combined with the parameters $\boldsymbol{\theta}$. The remainder of the network consists of the standard ratio estimators[4] implemented in SWYFT and described in Refs. [41,44], optimized on the standard binary cross-entropy loss relevant to neural ratio estimation [40,41,51,52]. As a reference, we provide the various numerical settings choices for the algorithm both in the SAQQARA release, as well as in the Appendices, where we also discuss the computational performance.

When designing this architecture, one motivation was to ensure its scalability to more complex situations. For example, with this implementation, we can directly apply this to a LISA data structure containing either multiple channels or correlated noise without any modifications. In addition, looking toward analyses carried out in the time-frequency domain, aside from possibly slight modifications to the compression network, the *entirety* of the rest of the pipeline can remain unchanged. This opens up the possibility of applying this algorithm directly to, e.g., the separation of several SGWB components, more complex noise models including nonstationary noise scenarios, or varying detector configurations, which is something we aim to do in future work.

IV. RESULTS AND DISCUSSION

All of the key results for this work are in the context of the case studies described above and are summarized in Figs. 1 and 2. In brief, they highlight two key points: first, SBI techniques reproduce the results from traditional sampling methods (MCMC in this case). This is true both in the case of a PL template, as well as a more agnostic fit. Secondly, when we introduce the additional complexity from transient sources, our SBI method still produces unbiased posterior distributions without any modification.

In more detail, we will first discuss the agreement with traditional methods. To do so, we take case studies *C1*, *C2*, and *C3*. The first of these is shown by the black solid and dashed contours in Fig. 1. These highlight very clearly the fact that we can accurately reproduce the unbiased and accurate posteriors obtained from a traditional likelihood-based analysis for a PL template. Furthermore, in Fig. 2, we illustrate our ability to constrain a more agnostic template (in this case with 10 signal bins, although the result with 5 bins is shown in the Appendices). The three panels show the relative constraining power of the analysis via the posterior draws (blue lines) compared to prior draws (black lines) for the total signal, as well as the separate SGWB signal and instrumental noise contribution. For comparison, the injected signal is shown in yellow across all panels. We see that we reproduce several desirable characteristics in this agnostic case, for example, the fact that we obtain tight constraints on the SGWB signal when it dominates over the

[4]In particular, we can either estimate the individual 1-dimensional marginals, higher dimensional marginals, or an autoregressive estimator. All of these are implemented within SAQQARA.

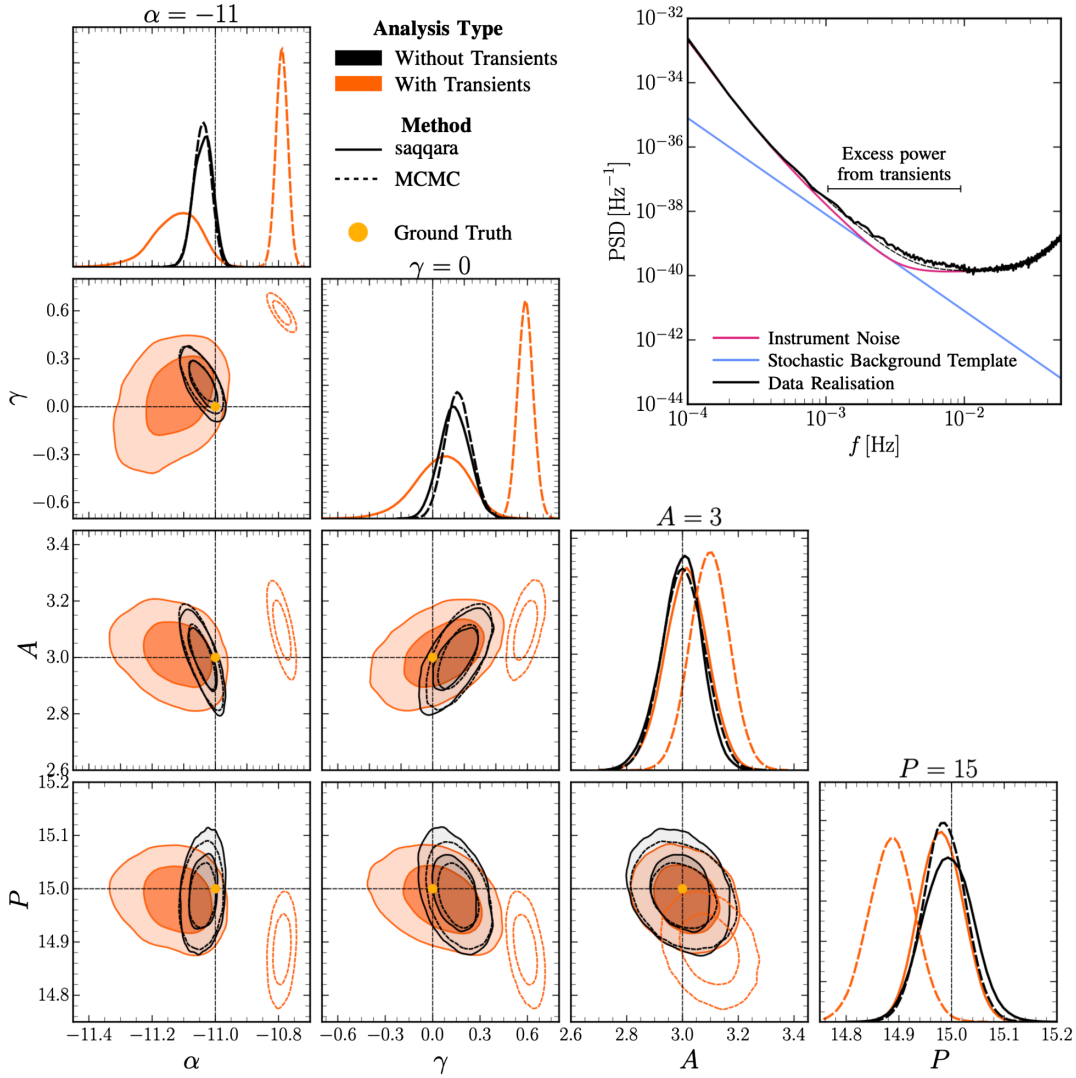

FIG. 1. Analysis results for the case studies *C1* and *C4*. Main Plot. Corner plot highlighting the two analysis results for case studies *C1*—which corresponds to the PL template without additional transients (shown in black solid and dashed contours)—and *C4*—where the transients are now present (shown in orange solid and dashed contours). The true injected values are highlighted by the dashed black horizontal and vertical lines, and by the yellow markers. Upper inset. Illustration of the explicit data realization (black line) for the case study *C4* along with the injected instrumental noise (pink line), stochastic background signal (blue line), and their sum (dashed black curve).

instrumental contribution, and wider constraints in the opposite case, i.e., outside the frequency band in which the instrument is most sensitive. The specific agreement with MCMC at the level of the posterior distribution over parameters for the five bin case is provided in the Appendices. Again, this highlights our precise agreement with traditional sampling-based methods.

The second key result is presented via the orange contours in Fig. 1. This is the result for the case study *C4* described above, where we introduce a population of supermassive black hole transients into the data. At the level of data realizations, we can easily understand the effect of this by looking at the upper inset in Fig. 1. In particular, the additional transients lead to a relative (and realization dependent) excess in the mid-frequency region compared to the noise templates alone. It is important to note that this excess is not distributed in the same way as the instrumental noise and SGWB contributions (which are distributed as colored zero mean Gaussian noise). This means that the impact of the transients cannot be simply accounted for in a likelihood-based approach unless each signal is individually analyzed and the full parameter space for transient signals is sampled within the MCMC. The cost of this is significant, however, since the relative dimensionality of the problem would then increase by a huge margin in line with the number of signals (hundreds) multiplied by the number of signal parameters (tens). If we look to avoid this and take the naive approach by analysing the data using the same likelihood model as in *C1–C3*, we obtain the dashed orange contours. Not unexpectedly, these are

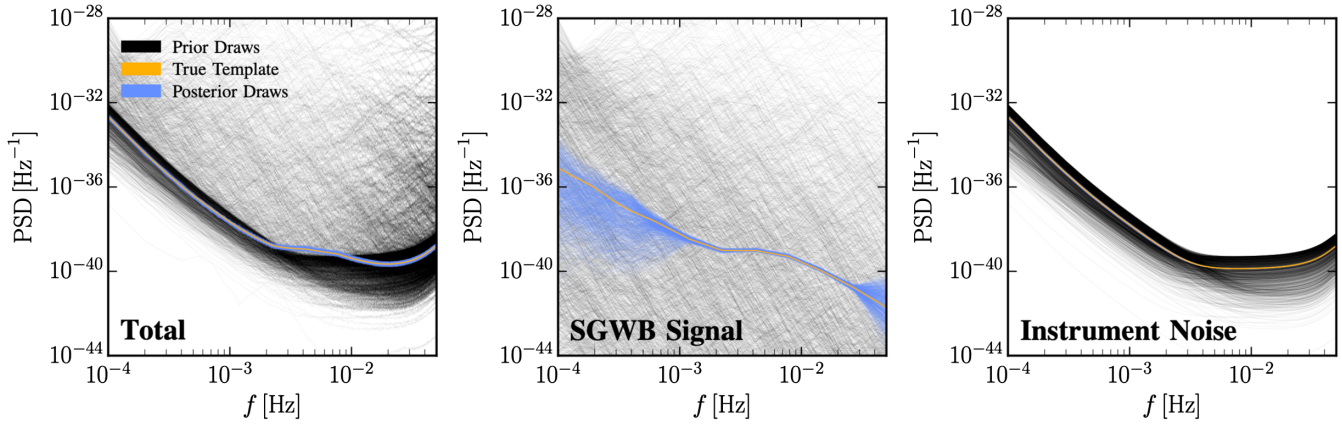

FIG. 2. Reconstruction of stochastic background and instrumental noise components for an agnostic template with ten signal bins. In all panels, draws from the initial Bayesian prior are shown in black, whereas the resulting posterior draws are shown in blue. The injected signal is shown in yellow. Left: reconstruction of the total power-spectral density (PSD). Middle: Reconstruction of the individual stochastic background component. Right: reconstruction of the instrumental noise contribution.

significantly biased compared to the true injection value shown by the yellow dot. We do emphasise that this is not necessarily an intrinsic limitation of the MCMC approach, since we knew that the likelihood was incorrect. What this result emphasises, however, is that if one wants to use traditional sampling methods, there is a crucial need for either (a) clean data with all transients removed, or (b) a higher-dimensional likelihood fitted to each transient, at the cost of significant sampling time.

At this point, we present the approach using our SBI algorithm, which looks to provide a compromise between these two options. Essentially, SBI allows us to implicitly marginalise over the complicated parameter space introduced by the transients by learning the effective marginalized likelihood-to-evidence ratio. This only depends on the SGWB and instrumental noise parameters and can be directly sampled. The results of this process are shown by the solid orange contours in Fig. 1. Crucially, we see that we are able to provide an unbiased and precise posterior that correctly accounts for the presence of the transients via a slight, but noticeable broadening of the SGWB parameter (α and γ) posteriors. Furthermore, we see that we are actually able to obtain identical constraints on the noise parameters as compared to the *C1* case. There are two reasons for this: firstly, in our setup, the transients mainly affect the mid-frequency region, so the noise parameters can be constrained to the same precision from the low- and high-frequency data, and secondly, we used the same data realization for the SGWB and noise components. In addition to this, we carry out standard posterior coverage tests [53] for this case study, which are provided in the Appendices. We find that for all parameters, our posteriors are extremely well calibrated, adding strength to the claim that we correctly marginalise over the additional transient components. This is a key result of this work and is the connection point to something resembling the LISA global fit challenge. In particular, this shows that SBI provides the possibility to directly analyse the SGWB or instrumental noise component of a LISA data stream without necessarily removing all transient artifacts. Instead, they could be simply included in the data generation, and implicitly marginalized over using the procedure presented in this work.

V. CONCLUSIONS AND OUTLOOK

This work is a first step in demonstrating the potential of SBI techniques in addressing the data analysis challenges of LISA. Specifically, we showed in the case studies *C1–C3* that posteriors computed with SBI match the ones obtained with traditional likelihood-based methods. The final study (*C4*) shows that these results are robust even in the presence of additional transients (when included in the training data), demonstrating the potential of directly estimating marginal posteriors, which poses a serious challenge in the MCMC approach. While we choose a specific class of transients, we anticipate this result to hold for different signals (or noise features), too. More investigations are required to test the robustness and limitations of the methodology. This characteristic of SBI opens up the intriguing possibility of integrating similar techniques in (or using them as independent cross-checks for some parts of) the LISA global fit pipeline.

Building on this, future work will need to address (i) the inclusion of multiple data channels (nominally the standard A, E, and T basis that is well-established in the literature), (ii) more realistic source modeling, providing either a full range of possible sources or a representative and realistic output of an initial analysis where sources are fully or partially removed, as well as robustness tests in view of limited source knowledge and (iii) instrumental noise that is less well-calibrated, nonstationary, correlated, and/or contains more relevant noise components. We stress that none of these strike us as fundamental obstacles within the

infrastructure presented here and that the simplifying assumptions taken in this work only served as a starting point in the development of these techniques. In addition, we envisage applications of SBI techniques to perform parameter estimations in other blocks of the LISA analysis problem, maintaining the ability to correctly marginalize over all other parameters. In this spirit, we hope that the release of our public code will trigger some of these developments.

Note added. In parallel to this work, neural posterior estimation (as opposed to neural ratio estimation used here) was investigated in Ref. [47], demonstrating in particular the feasibility of efficiently reconstructing signals using the full 3 years of LISA data and 3 data channels. They follow the same simplifying assumptions on the noise model as employed in this work.

Along with the results presented here, we also provide an extendable public code which can be found in [54]. `GitHub`: The SAQQARA simulation and inference library is available in [54] (`peregrine-gw/saqqara`). In addition, the TMNRE implementation SWYFT is available in [55] (`undark-lab/swyft`).

ACKNOWLEDGMENTS

We are grateful for fruitful discussions at the TH Institute on SGWB data analysis organized at CERN, in particular with A. Dimitriou, D. Figueroa, and B. Zaldivar. This work is part of the project CORTEX (NWA.1160.18.316) of the research programme Nederlandse Wetenschapsagenda-Onderzoek langs Routes door Consortia (NWA-ORC) which is (partly) financed by the Dutch Research Council (NWO). Additionally, C. W. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 864035). U. B. is supported through the CORTEX project of the NWA-ORC with Project No. NWA.1160.18.316 which is partly financed by the Dutch Research Council (NWO). J. A. is supported through the research program "The Hidden Universe of Weakly Interacting Particles" with Project No. 680.92.18.03 (NWO Vrije Programma), which is partly financed by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Dutch Research Council). J. A. acknowledges the hospitality of CERN TH, where this work was initiated. J. A. and M. P. acknowledge the hospitality of Imperial College London, which provided office space during some parts of this project.

APPENDIX A: LISA MEASUREMENTS

This section summarizes the formalism to characterize signal and noise in LISA. Following Ref. [13], we start with a brief description of single link GW measurements, proceed by discussing the noise PSDs, and conclude by introducing time delay interferometry (TDI) variables. As discussed in the main body of this work, the data $d(t)$ contains a linear superposition of signal and noise, and has a Fourier transform[5] which reads

$$\tilde{d}(f) = \int_{-T/2}^{T/2} \mathrm{e}^{2\pi i f t} d(t) \mathrm{d}t, \tag{A1}$$

where T is the observation time. Assuming these components are uncorrelated—which implies we can treat them independently—and stationary—which implies vanishing correlations between different frequencies—we obtain

$$\begin{aligned}\langle \tilde{d}(f)\tilde{d}^*(f')\rangle &= \frac{1}{2}\delta(f-f')[S^{\mathrm{N}}(f)+S^{\mathrm{GW}}(f)] \\ &= \frac{1}{2}\delta(f-f')\left[S^{\mathrm{N}}(f)+\sum_{\lambda}\mathcal{R}_{\lambda}(f)P_h^{\lambda}(f)\right],\end{aligned} \tag{A2}$$

where $S^{\mathrm{N}}(f)$, $S^{\mathrm{GW}}(f)$ are the noise and signal PSDs. These are real, positive, and even functions of f. In the second equality, we have further expanded $S^{\mathrm{GW}}(f)$ in terms of $\mathcal{R}_{\lambda}$, the sky-averaged LISA response function, which projects the GW PSD P_h^{λ} (with λ running over the two GW polarizations), onto the data. For this purpose, we have assumed P_h^{λ} to be homogeneous, isotropic, and diagonal in the GW polarization basis (for details, see, e.g. Refs. [13,20,56,57]). For completeness, we recall that, given the intensity $I(f) \equiv \sum_{\lambda} P_h^{\lambda}/2$, the SGWB energy density reads:

$$\Omega_{\mathrm{GW}}h^2 \equiv \frac{4\pi^2}{3(H_0/h)^2} f^3 I(f), \tag{A3}$$

where $H_0 \approx 3.24 \times 10^{-18} h$ Hz is the Hubble constant today and its dimensionless value [58] is $h = 0.6766 \pm 0.0042$. Since planar interferometers like LISA are not sensitive to chirality,[6] $\mathcal{R}_{\mathrm{L}} = \mathcal{R}_{\mathrm{R}}$ (with L, R, denoting the two GW helicities), and thus, the data only depends on $I(f)$.

For the transient signals, we generate a database of 2 million waveforms evaluated on the same frequency grid

[5]We are assuming $d(t)$ is a continuous function of time, which is not strictly correct since, in reality, data sampling occurs at a finite rate (and typically is impacted further by some downsampling procedure). On the other hand, since these effects will only affect the high-frequency end of the spectrum ($f \gtrsim 1$ Hz), which we do not include in the analysis, we can safely ignore this effect.

[6]Due to symmetry, left and right-hand polarized GW waves coming from opposite directions would induce the same effect in a planar interferometer. Possible ways to break this degeneracy and measure chirality include, e.g., correlating the signal measured by non-coplanar interferometers [59–61] or using the dipole induced by the motion of the detector with respect to the SGWB frame [57,62,63].

as in the main analyses.[7] Each waveform has parameters sampled uniformly from the priors: $M_c \in [8,9] \times 10^5 M_\odot$ (chirp mass), $\eta \in [0.25, 1]$ (mass ratio), $\chi_1, \chi_2 \in [-1, 1]$ (dimensionless spins), $d_L \in [5, 10] \times 10^4$ Mpc (luminosity distance), $t_c = 0$ s (time of coalescence), and $\phi_c = 0$ (phase). We then rescale the strain amplitude of the transient by 10^{-3} to make them behave like a population of sub-threshold sources. While this setup is not fully realistic, it provides a case study to show that these injections are sufficient to bias the MCMC approach.

1. Single link signal response and noise PSDs

Following the notation of Ref. [13], the fractional frequency shift $\eta_{ij}^{\rm GW}(t)$ induced by a GW perturbing the path of a photon released at time $t - L_{ij}$ from an emitter located at $\vec{x}_i$, to a receiver $\vec{x}_j$ at time t, (with $L_{ij} \equiv |\vec{x}_i - \vec{x}_j|$), reads

$$\eta_{ij}^{\rm GW}(t) = i \int_{-\infty}^{\infty} {\rm d}f \frac{f}{f_{ij}} {\rm e}^{2\pi i f (t - L_{ij})} \times \int {\rm d}\Omega_{\hat{k}} {\rm e}^{-2\pi i f \hat{k}\cdot\vec{x}_i} \sum_\lambda \xi_{ij}^\lambda(f,\hat{k}) \tilde{h}_\lambda(f,\hat{k}), \quad \text{(A4)}$$

where we have expanded the GW in plane waves, with $\vec{k}$ being the GW momentum, $\Omega_{\hat{k}}$ denoting the solid angle, $\tilde{h}_\lambda(f,\hat{k})$ being the coefficients of the expansion, and λ running over the two GW polarizations. We have also introduced the characteristic frequencies $f_{ij} \equiv (2\pi L_{ij})^{-1}$ and

$$\xi_{ij}^\lambda(f,\hat{k}) = {\rm e}^{\pi i f L_{ij}(1 - \hat{k}\cdot\hat{l}_{ij})} {\rm sinc}(\pi f L_{ij}(1 + \hat{k}\cdot\hat{l}_{ij})) \frac{\hat{l}_{ij}^a \hat{l}_{ij}^b}{2} e_{ab}^\lambda(\hat{k}), \quad \text{(A5)}$$

where $\hat{l}_{ij} = (\vec{x}_j - \vec{x}_i)/|\vec{x}_j - \vec{x}_i|$ is a unit vector pointing from i to j and $e_{ab}^\lambda(\hat{k})$ are the GW polarization tensors. Let us proceed by assuming that fluctuations in the arm lengths are negligible and the LISA configuration is perfectly equilateral at all times,[8] so that $L_{ij} = L$ and $f_{ij} = f_*$. By substituting the statistical properties for a homogeneous, isotropic, and nonchiral SGWB GW signal

$$\langle \tilde{h}_\lambda(f,\hat{k}) \tilde{h}^*_{\lambda'}(f',\hat{k}') \rangle = \delta(f - f') \delta(\hat{k} - \hat{k}') \delta_{\lambda\lambda'} \frac{P_h^{\lambda\lambda'}(f)}{16\pi}$$
$$\langle \tilde{h}_\lambda(f,\hat{k}) \tilde{h}_{\lambda'}(f',\hat{k}') \rangle = 0, \quad \text{(A6)}$$

where $P_h^{\lambda\lambda'}(f)$ denotes the one-sided GW PSD, and comparing with Eq. (A2), we express the signal PSD as

$$S_{ij,mn}^{\rm GW}(f) \equiv \sum_\lambda \mathcal{R}_{ij,mn}^\lambda P_h^{\lambda\lambda}(f) \equiv \left(\frac{f}{f_*}\right)^2 \sum_\lambda P_h^{\lambda\lambda}(f) \times \int \frac{{\rm d}\Omega_{\hat{k}}}{4\pi} {\rm e}^{-2\pi i f \hat{k}\cdot(\vec{x}_i - \vec{x}_m)} \xi_{ij}^\lambda(f,\hat{k}) \xi_{mn}^\lambda(f,\hat{k})^*. \quad \text{(A7)}$$

Here, $\mathcal{R}_{ij,mn}^\lambda$ are the (polarization-dependent) single link response functions.

As far as noise is concerned, we assume two contributions dominate the noise budget: test mass (TM) noise, typically dominating at low-frequency, and optical metrology system (OMS) noise, typically dominating at high frequencies. These are the two preeminent secondary noises that remain unsuppressed at the end of the TDI procedure[9] (see next section). TM and OMS noise contribute to the single link measurement as

$$\eta_{ij}^{\rm N}(t) = n_{ij}^{\rm OMS}(t) + D_{ij} n_{ji}^{\rm TM}(t) + n_{ij}^{\rm TM}(t), \quad \text{(A8)}$$

where D_{ij} denotes the time-delay operator that, under the assumptions of the present work (static and equilateral LISA constellation), acts on any function of time $x(t)$ as $D_{ij}x(t) = x(t - L)$, which, in the frequency domain, reduces to a phase shift represented by a multiplicative $\exp\{-2\pi i f L\}$ factor. To express the noise contribution to Eq. (A2), we should then proceed by computing the noise PSD $S^{\rm N}(f)$. For this purpose, we assume the individual noise terms to be uncorrelated and zero mean so that

$$\langle \tilde{n}_{ij}^{\rm TM}(f) \tilde{n}_{lm}^{\rm TM*}(f') \rangle = \frac{\delta_{ij,lm}}{2} S^{\rm TM}(f) \delta(f - f'),$$
$$\langle \tilde{n}_{ij}^{\rm OMS}(f) \tilde{n}_{lm}^{\rm OMS*}(f') \rangle = \frac{\delta_{ij,lm}}{2} S^{\rm OMS}(f) \delta(f - f'), \quad \text{(A9)}$$

This in turn assumes that all the TM and OMS components are equal. Furthermore, following Ref. [5], the PSDs are given by:

[7] Assuming each merger to occur within a data segment, the observation time defines, beyond the frequency resolution, the minimal frequency at which the system emits in that segment. In practice, this would correspond to cutting each of the transients at some minimal frequency, such that the time for the evolution from this frequency to the mergers is $\leq T_c$. For the sake of simplicity, we ignore this effect, which, in some sense, is only an artifact of the short observation period used in this analysis.

[8] In reality, several effects will contribute to breaking this perfectly symmetric configuration, leading to unequal and time-varying arm lengths. A more accurate description of the system should account for these modifications, which, e.g., will break the orthogonality of the usual AET Michelson TDI variables (see next section). For a discussion of some of these effects and their data analysis implications, see e.g. Ref. [13].

[9] A more realistic noise model would also have to account for subdominant contributions, e.g., the tilt-to-length noise [64–69], due to angular jitter in the readout system).

$$S^{\mathrm{TM}}(f) = A^2 \times 10^{-30} \times \left[1 + \left(\frac{0.4\ \mathrm{mHz}}{f}\right)^2\right] \times \left[1 + \left(\frac{f}{8\ \mathrm{mHz}}\right)^4\right] \times \left(\frac{1}{2\pi f c}\right)^2 \times (\mathrm{m}^2/\mathrm{s}^3),$$
$$S^{\mathrm{OMS}}(f) = P^2 \times 10^{-24} \times \left[1 + \left(\frac{2 \times 10^{-3}\ \mathrm{Hz}}{f}\right)^4\right] \times \left(\frac{2\pi f}{c}\right)^2 \times (\mathrm{m}^2/\mathrm{Hz}). \quad \text{(A10)}$$

In these expressions, the amplitudes of the TM and OMS noise PSDs are controlled by the dimensionless A and P parameters, respectively. To reproduce the noise level specified in [70], we set the fiducial values for these parameters to be $A = 3$, $P = 15$.

We stress that this is an overly simplified model of noise expected in LISA. For example, note that LPF's in-flight noise differs in level and shape from predictions, mostly at frequencies below 10^{-3} Hz. Moreover, we have not included any non-stationarity, such as glitches or drifts in the instrument noise. These effects will almost certainly be present in realistic data and will need to be taken into account in the development of data analysis pipelines. For longer observation times, possible anisotropies in the SGWB would also need to be taken into account.

2. Projection on the TDI variables and likelihood

TDI [71–77] is a data processing technique consisting of combining several interferometric measurements, typically performed at different times, that will be used in LISA to suppress the primary noise sources (mostly laser noise, a white noise contribution several orders of magnitude larger than the required noise level). While several TDI variables can achieve the target noise suppression [12,78–80], in this work, we focus on the most commonly used variables, the Michelson XYZ variables, and their orthogonal combinations, typically referred to as AET TDI variables. Moreover, several generations of TDI variables exist, which achieve noise cancellation in scenarios with increasing complication and realism. In the present work, we employ first-generation TDI variables, which can suppress primary noises for a constellation with unequal (but constant) arm lengths. This choice is sufficient, given that we restrict ourselves to the case of a maximally symmetric, i.e., equilateral and equal noise levels, configuration. In this framework, the X TDI variable is defined as:

$$\mathrm{X} = (1 - D_{13}D_{31})(\eta_{12} + D_{12}\eta_{21}) + (D_{12}D_{21} - 1)(\eta_{13} + D_{13}\eta_{31}), \quad \text{(A11)}$$

and the Y and Z variables correspond to cyclic permutations of the three satellites. Effectively, working in the XYZ TDI basis consists of considering three interferometers that share their arms, leading to correlated measurements. For this reason, it is customary to introduce the AET basis, which, under the assumptions made throughout this work, can be shown to be orthogonal. For explicit expressions of the noise PSDs in the AET basis see, e.g., Refs. [13,20].

We recall that for simplicity, in the analysis carried out in this work, we used the X TDI variable only, the generalization to all channels and to higher generation TDI variables is left for future work. In particular, exploiting the different response functions of these channels will enable some discrimination between signal and noise [20], although degeneracies remain when the instrument is modeled more realistically. Under the assumptions of this work, a generalization to three TDI channels is straightforward, however relaxing some assumptions on the symmetry of the configuration would require more elaborate treatment [13].

As discussed in Ref. [20], the data obtained following the procedure discussed in the main text presents mild non-Gaussianity. Thus, the (log-)likelihood should include some skewness corrections to model this component and avoid a systematic bias in the results. It is known [81–84] that an appropriate (log-)likelihood has form:

$$\ln \mathcal{L}(\vec{\theta}|D_k) = \frac{1}{3}\ln \mathcal{L}_{\mathrm{G}}(\vec{\theta}|D_k) + \frac{2}{3}\ln \mathcal{L}_{\mathrm{LN}}(\vec{\theta}|D_k), \quad \text{(A12)}$$

where $\vec{\theta} = \{\vec{\theta}_s, \vec{\theta}_n\}$ are the parameters (with $\vec{\theta}_s$, $\vec{\theta}_n$ being the signal and noise parameters, respectively), and $\ln \mathcal{L}_{\mathrm{G}}$, $\ln \mathcal{L}_{\mathrm{LN}}$ are a Gaussian and log-normal likelihood:

$$\ln \mathcal{L}_{\mathrm{G}}(\vec{\theta}|D_k) = -\frac{N_d}{2}\sum_k w_k \left[1 - D_k/\mathcal{D}_k(\vec{\theta})\right]^2,$$
$$\ln \mathcal{L}_{\mathrm{LN}}(\vec{\theta}|D_k) = -\frac{N_d}{2}\sum_k w_k \ln^2\left[\mathcal{D}_k(\vec{\theta})/D_k\right]. \quad \text{(A13)}$$

Here, $\mathcal{D}_k(\boldsymbol{\theta})$ denotes the theoretical model for the data $\mathcal{D}_k(\boldsymbol{\theta}) = \Omega_{\mathrm{GW}}(f_k, \boldsymbol{\theta}) + \Omega_{\mathrm{n}}(f_k, \boldsymbol{\theta})$, with $\Omega_{\mathrm{GW}}(f_k, \boldsymbol{\theta})$ and $\Omega_{\mathrm{n}}(f_k, \boldsymbol{\theta})$, being the signal and noise model, respectively.

APPENDIX B: TECHNICAL DETAILS, COVERAGE TESTS, AND PERFORMANCE

In this section, we report some technical details concerning the implementation of our technique, discuss its computational performance, and present coverage test results for the case study *C4*, which includes the additional transient sources. Starting with technical details regarding the implementation of SAQQARA, several numerical settings should be chosen for the general structure of the algorithm, as well as the network architecture. Each of the options is explained in detail within the configuration files in the SAQQARA repository, however, here we detail the choices made to produce the results in this work. First, in each round of inference, we use 500,000 simulations. For training the network, we train for a maximum of 50 epochs,

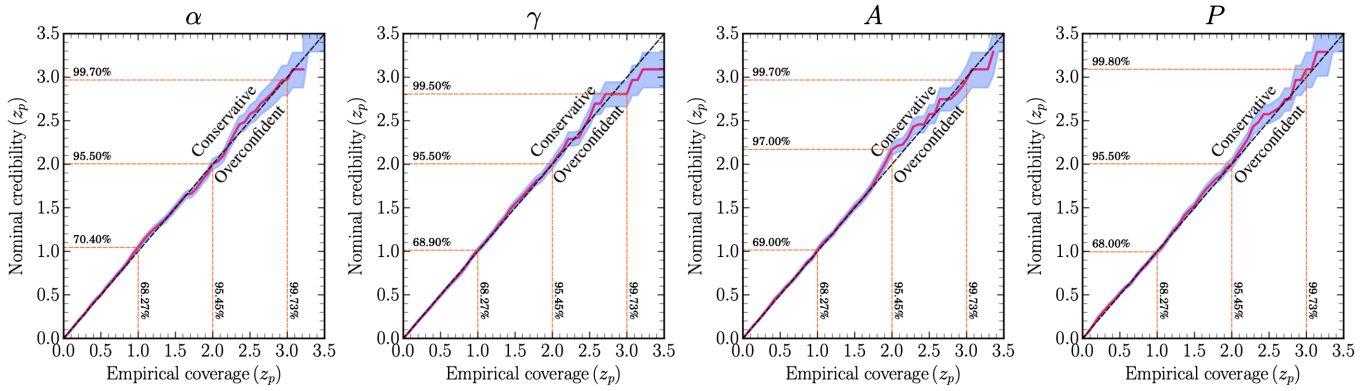

FIG. 3. Coverage test results for the analysis of the power law template in the presence of additional transient signals (case study *C4*). From left to right, the panels show the coverage statistics for each of the four parameters in the model (α, γ, A, P). For all panels, the empirical coverage is shown on the horizontal axis, and the nominal coverage is shown on the vertical axis. The 68% confidence interval on the coverage is shown by the blue shaded region, and the central value is shown by the pink line.

and set the patience before early stopping is triggered (due to a non-decreasing validation loss) to 7 epochs. Note that after training, we then reset the network to the state with the lowest validation loss. We take training and validation batch sizes of 512, splitting the simulation dataset in the ratio 0.9∶0.1, and start with an initial learning rate of 2×10^{-5}. In terms of inference, for truncating the 1-dimensional priors, we take $\alpha = 10^{-5}$ (in the sense of Refs. [27,42]). For the sampler, we follow the notation in Ref. [44] and take $\epsilon = 10^{-3}$, $\log \mathcal{L}_{\max} = 500$, $n_{\text{batch}} = 10$, $n_{\text{samples/slice}} = 20$, and $n_{\text{steps}} = 4$. As mentioned, these are detailed both in the SAQQARA repository, as well as in Ref. [44]. Finally, for the MCMC runs, we take $n_{\text{burn}} = 500$, $n_{\text{steps}} = 1000$, $r_{\text{conv}} = 10^{-3}$, and a maximum of 50 iterations.

As mentioned in the main text, coverage tests are an important consistency check of our SBI pipeline, especially in scenarios where a comparison to a traditional method like MCMC is not possible (either computationally, or statistically). As a brief review, coverage tests (see e.g. Ref. [53] for a more complete discussion in the context of SBI) implement the idea that in Bayesian parameter estimation, repeated inference over different noise/statistical realizations of the same signal should result in posteriors that shift relative to the true value. The rationale behind this sort of expected coverage test is that—simply as a result of statistical fluctuations—the $x\%$ credible interval for the posterior should contain the simulation-truth value $x\%$ of the time. To carry out this test for the case study in this work, we generated 1000 additional test simulations generated from the truncated prior. We then perform inference on each observation. In each case we can note how often the injected value was contained inside the $x\%$ confidence interval and construct a cumulative distribution. A well-calibrated posterior distribution will be a totally diagonal line when the expected coverage is plotted against the empirical findings. The results for case study *C4* are shown in Fig. 3 for each model parameter. We see that in every case, we obtain extremely well calibrated coverage for our posterior estimates. This strongly supports our claim in the main text of recovering unbiased posteriors for noise and SGWB parameters even in the presence of transient signals.

The final relevant discussion point concerns the computational performance of our SBI algorithm. In terms of computational complexity, there are a number of steps to generating posteriors; simulation/data generation (which is fully parallelized within SAQQARA); network training/likelihood-to-evidence estimation; and inference. With the setup considered here, we perform the inference in two steps. The first of these learns the individual marginal posteriors for the SGWB and noise parameters. This step is fully amortized (in the sense that once the training is complete, the inference is almost immediate on any signal) and allows us to efficiently "zoom in" (or truncate) to the prior region most relevant to the given observation, see Refs. [27,42] for more details on this process. Then, in the second step, we use the techniques developed in Ref. [44] to estimate the full joint posterior, which we explore with a `pytorch`-based sampling technique.[10] In terms of timing (on a 20 CPU-core resource with a single *NVIDIA GeForce GT 73* graphics card), the 500,000 simulations we use in each step took around 15 minutes to generate, and the subsequent network training took an additional 25 minutes (which does not need to be repeated). For the 1-dimensional ratio estimators, the inference is then essentially instantaneous, only requiring a single network evaluation. For the higher-dimensional marginals, the sampling adds a slight overhead (around 7 minutes in e.g. case study *C1*).

[10]We could also have estimated, e.g., 1- or 2-dimensional marginals for any/all parameters, depending on the specific inference needs.

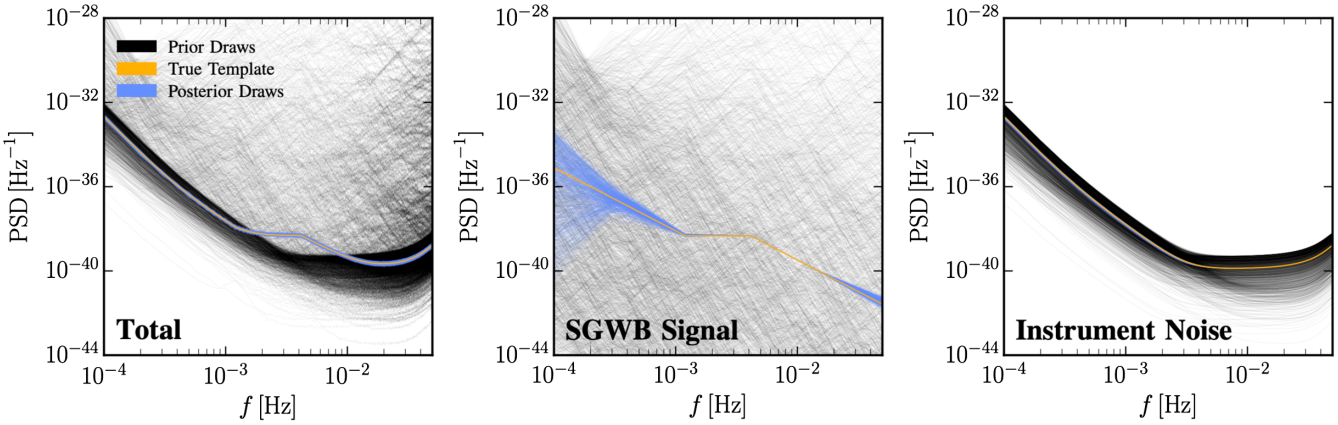

FIG. 4. Reconstruction of SGWB and instrumental noise components for an agnostic template with five signal bins. In all panels, draws from the initial Bayesian prior are shown in black, whereas the resulting posterior draws are shown in blue. The injected signal is shown in yellow. Left: reconstruction of the total power-spectral density. Middle panel. Reconstruction of the individual SGWB component. Right: reconstruction of the instrumental noise contribution.

APPENDIX C: FULL RESULTS FOR AGNOSTIC TEMPLATE FITS

In addition to the results presented in the main text, here we show additional results for the case studies *C2* and *C3*. In particular, in Fig. 4, we illustrate the reconstruction of the stochastic background and instrumental contributions in the context of the agnostic template with ten signal bins. Furthermore, in Fig. 5, we present the analysis at the level of parameter constraints for the agnostic template fit with five signal bins. We also show the comparison with the MCMC approach, which we see agrees extremely well with our SBI approach. Finally, in Fig. 6, we show the corresponding analysis for the ten bin agnostic fit.

APPENDIX D: REVIEW OF SIMULATION-BASED INFERENCE AND TMNRE

In this section we provide a brief review of simulation-based inference. Specifically, we cover the various classes of SBI methods before focusing on the technicalities of TMNRE.

1. Classes of SBI algorithm

All SBI algorithms are designed to answer the same question: *how can we do robust Bayesian inference given an implicit representation of the likelihood through a generative model?* In other words, suppose we are only given a "simulator" that takes model parameters $\theta = (\theta_1, \theta_2, \ldots)$ and stochastically generates data x, then the key idea is that running this simulator is equivalent to sampling from the likelihood $x \sim p(x|\theta)$. This is the origin of the term "likelihood-free inference", although this has now been replaced by the more appropriate "simulation-based" or "implicit likelihood" description [24]. It is worth noting that many of the recent advances in SBI have been facilitated by corresponding developments in machine learning. This has opened up the opportunity to use SBI methods to analyse high-dimensional and complex data structures such as images, time series, point clouds etc. On a more historical note, all of these algorithms move beyond the paradigm of approximate Bayesian computation (ABC) [85], which requires the choice of a hand-crafted summary statistic and distance measure to quantify the similarity between two sets of data.

There are a number of classes of SBI algorithms that vary in terms of how they estimate the relevant quantities in Bayes' theorem. In each case, the general goal is to obtain the posterior $p(\theta|x) = p(x|\theta)p(\theta)/p(x)$, where $p(\theta)$ is the (Bayesian) prior and $p(x)$ is the (Bayesian) evidence. In particular, they can be broadly classified as follows:

(i) *Neural posterior estimation (NPE).* These methods aim to directly estimate the posterior $p(\theta|x)$, typically utilizing neural network structures such as normalizing flows, which are manifestly normalized probability densities and easy to sample from [35]. This has been used successfully in a number of contexts including compact binary GW data analysis [28,29,86].

(ii) *Neural likelihood estimation (NLE).* In contrast to NPE, likelihood-estimation techniques construct an approximation to the (simulated) data likelihood $p(x|\theta)$ [35,38,39]. This can then be sampled using traditional stochastic sampling techniques such as MCMC or nested sampling.

(iii) *Neural ratio estimation (NRE).* The third class of methods is neural ratio estimation [40,42,51,52,87,88], which, contrary to the two mentioned above, approaches the Bayesian inference problem by constructing an estimate of the likelihood-to-evidence ratio $r(x, \vartheta) = p(x|\vartheta)/p(x) = p(\vartheta|x)/p(\vartheta)$, where ϑ is some collection

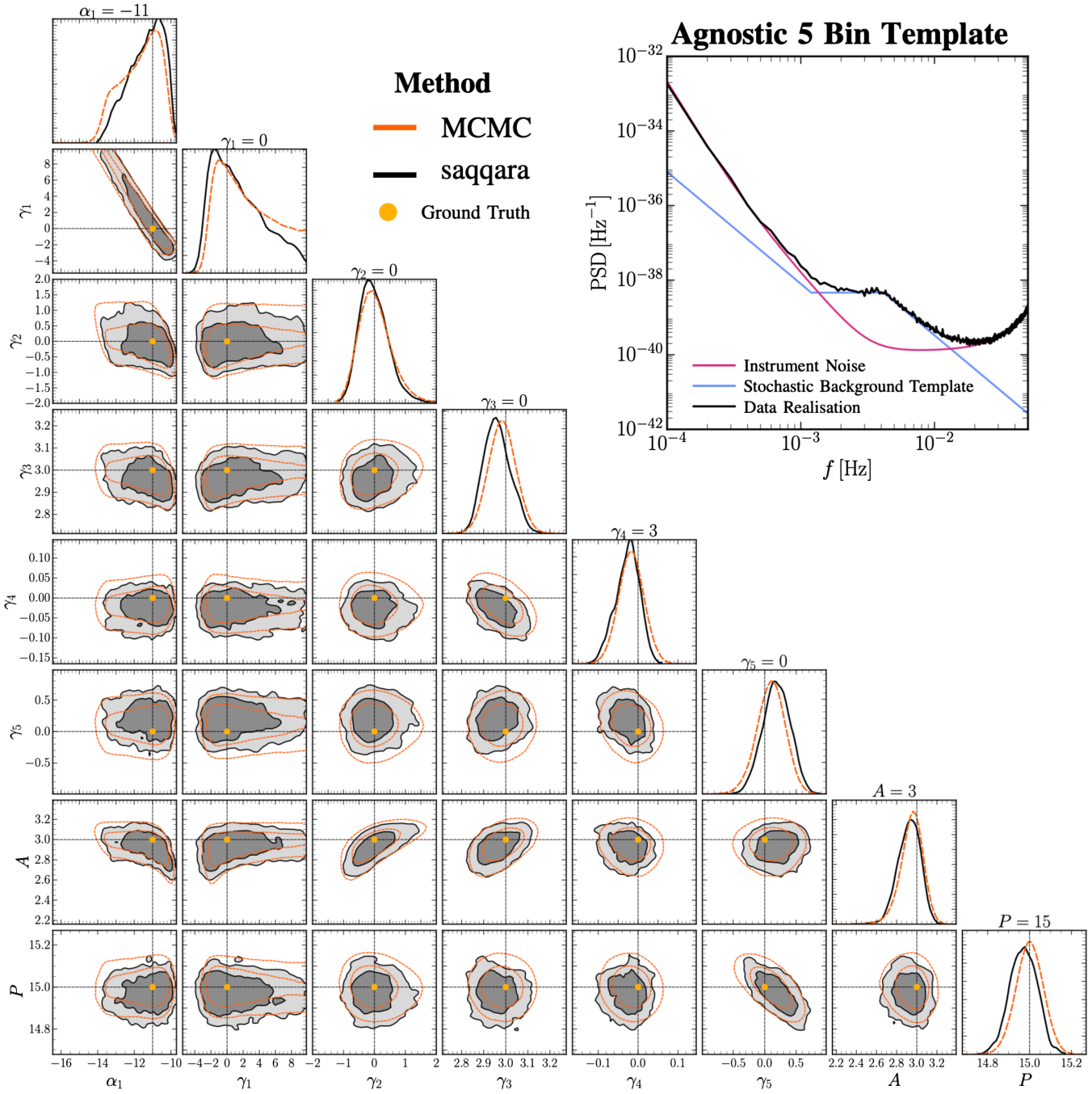

FIG. 5. Parameter constraints for agnostic template fit with five signal bins. Main plot. Corner plot highlighting the relative agreement between the MCMC (shown by orange dashed lines) and SBI (shown in solid black) approaches for all parameters in the agnostic template fit with five signal bins. The true injected values are highlighted by the dashed black horizontal and vertical lines, and by the yellow markers. Upper inset. Illustration of the explicit data realization (black line) for the case study *C2* along with the injected instrumental noise (pink line) and stochastic background signal (blue line).

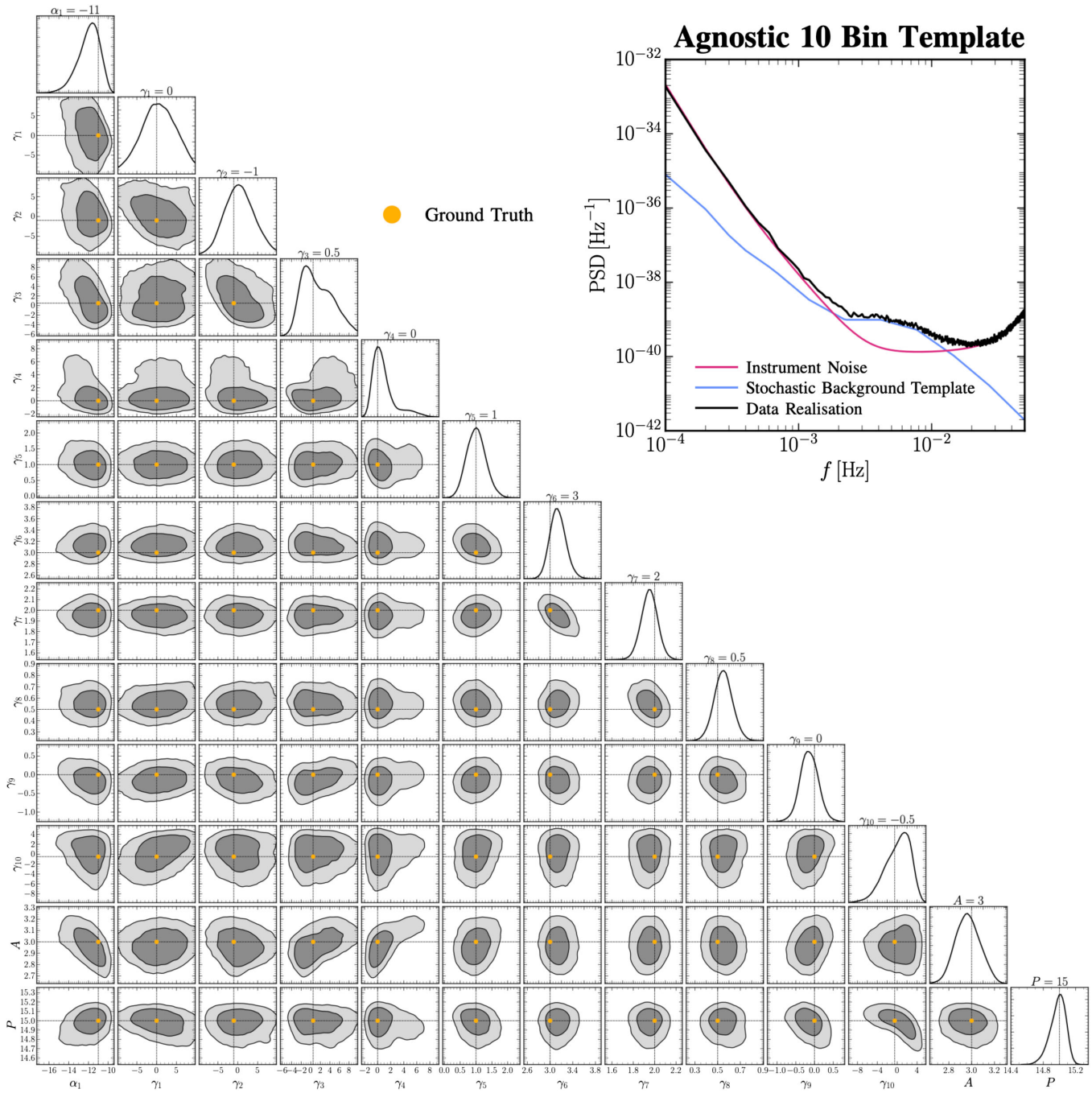

FIG. 6. Parameter constraints for agnostic template fit with five signal bins. Main plot. Corner plot highlighting the parameter constraints using the SBI approaches for all parameters in the agnostic template fit with ten signal bins. The true injected values are highlighted by the dashed black horizontal and vertical lines, and by the yellow markers. Upper inset. Illustration of the explicit data realization (black line) for the case study *C3* along with the injected instrumental noise (pink line) and stochastic background signal (blue line).

of parameters in θ, or derived parameters. The equality here follows simply as a result of probability rules and therefore if one can construct $r(x, \vartheta)$, then it is possible to directly access the posterior by re-weighting prior samples. This work is based on a specific implementation of NRE, truncated marginal neural ratio estimation (TMNRE) [41], which we discuss in detail below.

2. Truncated marginal neural ratio estimation (TMNRE)

In the main text, we discussed the reasons why we believe that, from the big-picture perspective, TMNRE is a suitable algorithm for the application of SBI to SGWB analysis. Broadly, these were focussed on two key properties of the algorithm: the truncation/sequential aspect (the "T"), and the ability to directly estimate marginal posteriors (the "M"). We argued that we expected these to lead to significant simulation efficiency and statistical flexibility when applied to the LISA data analysis challenge. In this section, we briefly review the technical aspects of our TMNRE algorithm. It is also worth noting that TMNRE has been successfully applied in a number of scenarios beyond SGWB and LISA data analysis, including CMB and 21 cm cosmology, stellar streams, strong lensing image analysis, point sources, and GWs from compact binaries [25–27,30,31,43,89–91].

Estimating the ratio $r(x, \vartheta)$. The first point to clarify about the TMNRE algorithm is how it estimates $r(x, \vartheta)$. In particular, we note here that ϑ need not be the full set of model parameters θ. In what follows, it could be a single parameter $\vartheta = \theta_i$, a set of parameters $\vartheta = (\theta_i, \theta_j, \ldots)$, or some derived model parameter. This directly encodes the marginal aspect of the algorithm and defines what we mean in the main text with "implicitly marginalizing" over, e.g., the parameters of transient signals. With this clarified, the estimation of the ratio proceeds as follows: first, we note that we can reexpress $r(x, \vartheta) = p(x|\vartheta)/p(x) = p(x, \vartheta)/p(x)p(\vartheta)$. In other words, $r(x, \vartheta)$ is the ratio between a *joint* sample $x, \vartheta \sim p(x, \vartheta) = p(x|\vartheta)p(\vartheta)$ and a *marginal* sample $x, \vartheta \sim p(x)p(\vartheta)$. It is worth noting that for any choice of ϑ, it is trivial to get either set of samples. Joint samples $x, \vartheta \sim p(x|\vartheta)p(\vartheta)$, properly marginalized over the variation in the other parameters, are obtained by running the simulator on prior samples from the full model $\theta \sim p(\theta)$. Parts of the parameter space that are not of interest can be simply discarded or masked. Similarly, to generate marginal samples, one can take a pair (x, θ) from a simulation run, and resample $\theta \sim p(\theta)$. Again irrelevant parameters can be discarded.

The second step for ratio estimation is to construct a binary classification task between joint and marginal samples. Specifically, the goal is to find a classifier $d_\phi(x, \vartheta)$ with some trainable parameters ϕ (almost always in the form of a neural network and its weights) that optimally outputs, e.g., $d_\phi(x, \vartheta) = 0$ if x, ϑ is a joint sample and $d_\phi(x, \vartheta) = 1$ if it is drawn marginally. Said differently, we can rephrase the task of performing parameter inference as a classification (marginal vs. joint samples) problem, which is extremely well suited to modern supervised machine learning techniques. In practice, the mapping of ratio estimation onto the binary classification task defined above is realized using the standard TMNRE (binary cross-entropy) loss function [41],

$$\mathcal{L}[f_\phi] = -\int \mathrm{d}x\mathrm{d}\vartheta \big[p(x, \vartheta) \ln\big(\sigma(f_\phi(x, \vartheta))\big) + p(x)p(\vartheta) \ln\big(1 - \sigma(f_\phi(x, \vartheta))\big)\big], \tag{D1}$$

where $d_\phi(x, \vartheta) = \sigma(f_\phi(x, \vartheta))$ and $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. The main motivation to justify such choice for the loss function is that it can be shown analytically (i.e., taking a functional derivative with respect to f_ϕ) that the *optimal* classifier $f^\star_\phi(x, \vartheta)$ is precisely $f^\star_\phi(x, \vartheta) = \ln r(x, \vartheta)$.

The above review explains the general reason that TMNRE works as a parameter inference algorithm, and also highlights how it can be used to implicitly marginalise over any part of the model. To further emphasise this point, TMNRE does not require any hand-crafted or analytic method for marginalization, just the ability to sample from a simulator, and mask/ignore the parameters that you wish to marginalize over.

Truncation. Before we end this section by discussing the relevant algorithm design choices in the context of SGWB analysis, we will briefly explain how truncation, or "zooming-in" is achieved. To illustrate this process, imagine that ϑ is just a single parameter $\vartheta = \theta_i$ and that we have a learned estimator $\hat{r}(x, \theta_i)$ of the ratio $r(x, \theta_i)$. Then, to obtain posterior estimates for some observation x_0, we can sample from the prior $\theta_i \sim p(\theta_i)$ and reweight by the ratio $r(x_0, \theta_i) = p(\theta_i|x_0)/p(\theta_i)$. In this process, if θ_i is a well-measured parameter, then there will be regions where $r(x_0, \theta_i)$ is very small. The general intuition behind the truncation part of TMNRE is to sequentially exclude these regions from the initial prior range (without changing the shape of the prior), and re-simulate new data targeting the region of interest. This approach has been shown to be extremely simulation efficient compared to either traditional joint inference or amortized techniques [25,27,43] to analyze data for individual GW observations. See Refs. [41,44] for a technical description of how this is achieved in single- and multiparameter setups, as well as the various precision settings that are required to achieve efficient but conservative truncation.

3. Design choice for SGWB analysis

Within this general framework, there are a number of design choices that are relevant to the application of

TMNRE to SGWB analysis. The most obvious of these is the forward simulation model, which we discuss in the main text. The second set of choices are the TMNRE settings which are also discussed above. The final concrete design specifications that are required concern the design of the network architecture for the classifier $f_\phi(x, \vartheta)$. As we mention in the main text, for the application of TMNRE, this network typically splits into two components: a compression network $\tilde{s}(x)$ and a ratio estimator $\tilde{r}(s, \vartheta)$, which are combined to get $f_\phi(x, \vartheta) = \tilde{r}(\tilde{s}(x), \vartheta)$. Importantly, both the compression network $\tilde{s}$ and $\tilde{r}$ are trained simultaneously. This ensures that no hand-crafted compression statistics are required, but rather, they are learnt directly. In our implementation, we use a relatively standard form for the ratio estimator $\tilde{r}(s, \vartheta)$, see e.g. Refs. [41,44], and therefore we spend the rest of the section explaining the choices for the compression network $\tilde{s}(x)$.

Ultimately the compression network architecture that we chose for $\tilde{s}(x)$ is motivated by the data that we are trying to analyse for extracting the SGWB. In particular, in terms of structure, the data x consists of noise variances $\langle \bar{d}^\star(f_i)\bar{d}(f_i)\rangle$ across a sequence of frequency bins f_i, in one or more channels. We know that broadly, the signal we are looking for in this work is characterized by a set of parameters $\theta_{\rm SGWB}$ that define a template that spans the various bins. Beyond this, we know that there are/could be additional contributions from instrumental noise or transient sources that may induce excesses to the signal, or correlated, non-Gaussian statistics across frequency bins. This physical intuition motivated the main component of the compression network in SAQQARA, which is essentially a 1-dimensional version of a `unet` architecture, similar to that described in Ref. [27]. In simple terms, the `unet` consists of two parts: a part that shrinks the initial data down, and a second part that rescales it back up. In the downscaling part, the goal is to gradually downsample (using a sequence of convolutional and max-pooling layers) the data and extract sequentially more fine-grained features. This creates an information bottleneck at the bottom of the "U" structure which encourages the network to extract the most important features from the data. In the decompression step, the `unet` attempts to rebuild the data, with various segments identified and classified. The final technical addition to the network architecture is the existence of "skip" connections, which allow for the higher-level features learned in the downsampling steps to be used in the corresponding reconstruction. In the context of LISA data, we can imagine this architecture first extracting the relevant frequency bins for a given SGWB signal, before focussing on the fine-grained details that are controlled by e.g. the template, and then reconstructing the signal. These are very general statements of course, but one of our aims with the SAQQARA (see github repository) pipeline is to be agnostic to the classes of excess on top of an SGWB signal.

[1] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 061102 (2016).
[2] G. Agazie *et al.* (NANOGrav Collaboration), Astrophys. J. Lett. **951**, L8 (2023).
[3] J. Antoniadis *et al.*, Astron. Astrophys. **678**, A50 (2023).
[4] H. Xu *et al.*, Res. Astron. Astrophys. **23**, 075024 (2023).
[5] P. Amaro-Seoane *et al.* (LISA Collaboration), arXiv:1702.00786.
[6] M. Punturo *et al.*, Classical Quantum Gravity **27**, 194002 (2010).
[7] D. Reitze *et al.*, Bull. Am. Astron. Soc. **51**, 035 (2019); arXiv:1907.04833.
[8] N. J. Cornish and J. Crowder, Phys. Rev. D **72**, 043005 (2005).
[9] J. D. Romano and N. J. Cornish, Living Rev. Relativity **20**, 2 (2017).
[10] C. Caprini and D. G. Figueroa, Classical Quantum Gravity **35**, 163001 (2018).
[11] M. R. Adams and N. J. Cornish, Phys. Rev. D **82**, 022002 (2010).
[12] M. Muratore, D. Vetrugno, S. Vitale, and O. Hartwig, Phys. Rev. D **105**, 023009 (2022).
[13] O. Hartwig, M. Lilley, M. Muratore, and M. Pieroni, Phys. Rev. D **107**, 123531 (2023).
[14] M. Vallisneri, Classical Quantum Gravity **26**, 094024 (2009).
[15] S. Babak *et al.* (Mock LISA Data Challenge Task Force Collaboration), Classical Quantum Gravity **27**, 084009 (2010).
[16] T. B. Littenberg and N. J. Cornish, Phys. Rev. D **107**, 063004 (2023).
[17] N. Karnesis, M. Lilley, and A. Petiteau, Classical Quantum Gravity **37**, 215017 (2020).
[18] C. Caprini, D. G. Figueroa, R. Flauger, G. Nardini, M. Peloso, M. Pieroni, A. Ricciardone, and G. Tasinato, J. Cosmol. Astropart. Phys. 11 (2019) 017.
[19] M. Pieroni and E. Barausse, J. Cosmol. Astropart. Phys. 07 (2020) 021; 09 (2020) E01.
[20] R. Flauger, N. Karnesis, G. Nardini, M. Pieroni, A. Ricciardone, and J. Torrado, J. Cosmol. Astropart. Phys. 01 (2021) 059.
[21] Q. Baghi, N. Karnesis, J.-B. Bayle, M. Besançon, and H. Inchauspé, J. Cosmol. Astropart. Phys. 04 (2023) 066.
[22] M. Muratore, J. Gair, and L. Speri, Phys. Rev. D **109**, 042001 (2024).
[23] G. Boileau, N. Christensen, R. Meyer, and N. J. Cornish, Phys. Rev. D **103**, 103529 (2021).

[24] K. Cranmer, J. Brehmer, and G. Louppe, Proc. Natl. Acad. Sci. U.S.A. **117**, 30055 (2020).
[25] J. Alvey, U. Bhardwaj, S. Nissanke, and C. Weniger, arXiv:2308.06318.
[26] J. Alvey, M. Gerdes, and C. Weniger, Mon. Not. R. Astron. Soc. **525**, 3662 (2023).
[27] U. Bhardwaj, J. Alvey, B. K. Miller, S. Nissanke, and C. Weniger, Phys. Rev. D **108**, 042004 (2023).
[28] 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).
[29] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, Phys. Rev. Lett. **127**, 241103 (2021).
[30] N. A. Montel, A. Coogan, C. Correa, K. Karchev, and C. Weniger, Mon. Not. R. Astron. Soc. **518**, 2746 (2022).
[31] N. Anau Montel and C. Weniger, in *36th Conference on Neural Information Processing Systems* (2022); arXiv:2211.04291.
[32] A. Dimitriou, C. Weniger, and C. A. Correa, arXiv:2206.11312.
[33] K. Karchev, R. Trotta, and C. Weniger, Mon. Not. R. Astron. Soc. **520**, 1056 (2022).
[34] T. L. Makinen, T. Charnock, J. Alsing, and B. D. Wandelt, J. Cosmol. Astropart. Phys. 11 (2021) 049.
[35] G. Papamakarios and I. Murray, arXiv:1605.06376.
[36] A. Tejero-Cantero, J. Boelts, M. Deistler, J.-M. Lueckmann, C. Durkan, P. J. Gonçalves, D. S. Greenberg, and J. H. Macke, J. Open Source Softwaare **5**, 2505 (2020).
[37] J. Zeghal, F. Lanusse, A. Boucaud, B. Remy, and E. Aubourg, arXiv:2207.05636.
[38] J. Alsing, T. Charnock, S. Feeney, and B. Wandelt, Mon. Not. R. Astron. Soc. **488**, 4440 (2019).
[39] K. Lin, M. von Wietersheim-Kramsta, B. Joachimi, and S. Feeney, Mon. Not. R. Astron. Soc. **524**, 6167 (2023).
[40] F. Rozet and G. Louppe, arXiv:2110.00449.
[41] B. K. Miller, A. Cole, C. Weniger, F. Nattino, O. Ku, and M. W. Grootes, J. Open Source Softwaare **7**, 4205 (2022).
[42] B. K. Miller, A. Cole, P. Forré, G. Louppe, and C. Weniger, in *35th Conference on Neural Information Processing Systems* (2021); arXiv:2107.01214.
[43] A. Cole, B. K. Miller, S. J. Witte, M. X. Cai, M. W. Grootes, F. Nattino, and C. Weniger, J. Cosmol. Astropart. Phys. 09 (2022) 004.
[44] N. Anau Montel, J. Alvey, and C. Weniger, arXiv:2308.08597.
[45] E. Castelli, LISA Pathfinder noise performance results: disturbances in the sub-mHz frequency band and projection to LISA, Ph.D. thesis, Trento U., 2020.
[46] S. Babak, A. Petiteau, and M. Hewitson, arXiv:2108.01167.
[47] A. Dimitriou, D. G. Figueroa, and B. Zaldivar, arXiv:2309.08430.
[48] T. D. P. Edwards, K. W. K. Wong, K. K. H. Lam, A. Coogan, D. Foreman-Mackey, M. Isi, and A. Zimmerman, arXiv:2302.05329.
[49] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, Publ. Astron. Soc. Pac. **125**, 306 (2013).
[50] O. Ronneberger, P. Fischer, and T. Brox, arXiv:1505.04597.
[51] C. Durkan, I. Murray, and G. Papamakarios, arXiv:2002.03712.
[52] J. Hermans, V. Begy, and G. Louppe, arXiv:1903.04057.
[53] J. Hermans, A. Delaunoy, F. Rozet, A. Wehenkel, V. Begy, and G. Louppe, arXiv:2110.06581.
[54] https://github.com/peregrine-gw/saqqara.
[55] https://github.com/undark-lab/swyft.
[56] N. Bartolo, V. Domcke, D. G. Figueroa, J. García-Bellido, M. Peloso, M. Pieroni, A. Ricciardone, M. Sakellariadou, L. Sorbo, and G. Tasinato, J. Cosmol. Astropart. Phys. 11 (2018) 034.
[57] V. Domcke, J. Garcia-Bellido, M. Peloso, M. Pieroni, A. Ricciardone, L. Sorbo, and G. Tasinato, J. Cosmol. Astropart. Phys. 05 (2020) 028.
[58] N. Aghanim *et al.* (Planck Collaboration), Astron. Astrophys. **641**, A6 (2020); **652**, C4(E) (2021).
[59] N. Seto and A. Taruya, Phys. Rev. Lett. **99**, 121101 (2007).
[60] S. G. Crowder, R. Namba, V. Mandic, S. Mukohyama, and M. Peloso, Phys. Lett. B **726**, 66 (2013).
[61] G. Orlando, M. Pieroni, and A. Ricciardone, J. Cosmol. Astropart. Phys. 03 (2021) 069.
[62] N. Seto, Phys. Rev. Lett. **97**, 151101 (2006).
[63] N. Seto, Phys. Rev. D **75**, 061302 (2007).
[64] M.-S. Hartig, S. Schuster, G. Heinzel, and G. Wanner, J. Opt. **25**, 055601 (2022).
[65] M.-S. Hartig, S. Schuster, and G. Wanner, J. Opt. **24**, 065601 (2022).
[66] S. Paczkowski, R. Giusteri, M. Hewitson, N. Karnesis, E. D. Fitzsimons, G. Wanner, and G. Heinzel, Phys. Rev. D **106**, 042005 (2022).
[67] D. George, J. Sanjuan, P. Fulda, and G. Mueller, Phys. Rev. D **107**, 022005 (2023).
[68] M.-S. Hartig and G. Wanner, Phys. Rev. D **108**, 022008 (2023).
[69] M. Armano *et al.*, Phys. Rev. D **108**, 102003 (2023).
[70] LISA Science Study Team, LISA Science Requirements Document, ESA, Technical Report No. ESA-L3-EST-SCI-RS-001, 2018, https://lisa.nasa.gov/documentsReference.html.
[71] J. W. Armstrong, F. B. Estabrook, and M. Tinto, Astrophys. J. **527**, 814 (1999).
[72] M. Tinto and J. W. Armstrong, Phys. Rev. D **59**, 102003 (1999).
[73] F. B. Estabrook, M. Tinto, and J. W. Armstrong, Phys. Rev. D **62**, 042002 (2000).
[74] M. Tinto and S. V. Dhurandhar, Living Rev. Relativity **24**, 1 (2021).
[75] T. A. Prince, M. Tinto, S. L. Larson, and J. W. Armstrong, Phys. Rev. D **66**, 122002 (2002).
[76] D. A. Shaddock, M. Tinto, F. B. Estabrook, and J. W. Armstrong, Phys. Rev. D **68**, 061303 (2003).
[77] M. Tinto, F. B. Estabrook, and J. W. Armstrong, Phys. Rev. D **69**, 082001 (2004).
[78] D. A. Shaddock, Phys. Rev. D **69**, 022001 (2004).
[79] M. Vallisneri, Phys. Rev. D **72**, 042003 (2005); **76**, 109903 (E) (2007).
[80] M. Muratore, D. Vetrugno, and S. Vitale, Classical Quantum Gravity **37**, 185019 (2020).
[81] J. R. Bond, A. H. Jaffe, and L. E. Knox, Astrophys. J. **533**, 19 (2000).
[82] J. L. Sievers *et al.*, Astrophys. J. **591**, 599 (2003).

[83] L. Verde *et al.* (WMAP Collaboration), Astrophys. J. Suppl. Ser. **148**, 195 (2003).
[84] S. Hamimeche and A. Lewis, Phys. Rev. D **77**, 103013 (2008).
[85] M. Sunnåker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz, PLoS Comput. Biol. **9**, 1 (2013).
[86] J. Wildberger, M. Dax, S. R. Green, J. Gair, M. Pürrer, J. H. Macke, A. Buonanno, and B. Schölkopf, Phys. Rev. D **107**, 084046 (2023).
[87] A. Delaunoy, J. Hermans, F. Rozet, A. Wehenkel, and G. Louppe, arXiv:2208.13624.
[88] B. K. Miller, C. Weniger, and P. Forré, arXiv:2210.06170.
[89] A. Saxena, A. Cole, S. Gazagnes, P. D. Meerburg, C. Weniger, and S. J. Witte, Mon. Not. R. Astron. Soc. **525**, 6097 (2023).
[90] S. Gagnon-Hartman, J. Ruan, and D. Haggard, Mon. Not. R. Astron. Soc. **520**, 1 (2023).
[91] A. Coogan, N. Anau Montel, K. Karchev, M. W. Grootes, F. Nattino, and C. Weniger, Mon. Not. R. Astron. Soc. **527**, 66 (2024).