Versatile deep-learning: Monte Carlo methods for gravitational wave inference

Project No: 628

SUPA, School of Physics and Astronomy
University of Glasgow
Glasgow G12 8QQ, United Kingdom

Abstract. In the era of multi-messenger astronomy, deep-learning approaches to parameter estimation are gaining significant traction in an effort to speed up sky localisation for rapid electromagnetic follow-up. Of these approaches, many implement variational inference techniques to obtain drastic speed increases over traditional methods. Variational inference compromises on versatility in pursuit of speed by avoiding likelihood calculations. However, these likelihoods are key to many diagnostic frameworks and so the technique remains significantly inflexible in their absence. In this paper, we present a novel method of estimating these likelihoods using Monte Carlo approximations to increase the versatility of our variational inference model. We start by presenting our established `VItamin` model's capabilities of increasing parameter estimation sampling rate by more than 6 orders of magnitude compared to traditional methods. We then go on to present the results of our likelihood estimates and show that they are qualitatively self-consistent, followed by deriving an error relation to compare tolerances with other sampling algorithms. We finally present an application of this increased versatility by using an importance sampling framework to reweight our `VItamin` results relative to a `dynesty` target to improve our initial results by a process called likelihood reweighting.

1. Introduction

From the first hypothesis of gravitational waves in 1916 [1], to the first direct detection of a binary black hole merger by the advanced Laser Interferometer Gravitational wave Observatory (LIGO) [2] in 2015 [3], we continually find ourselves at the forefront of discovery when studying gravitational physics. As theorised in general relativity (GR), rotating massive objects release gravitational radiation in the form of gravitational waves, which cause ripples and deformations in the fabric of space-time. Signals from these massive objects can be divided into two general categories, continuous gravitational wave (CW) signals and transient gravitational wave (GW) signals. There is currently a high volume of work concerning CWs [4] and even significant effort to implement deep-learning (DL) techniques for imrproved detection [5,6]. However in this paper, we solely focus on GWs, in particular GWs arising from compact binary coalescences (CBCs). CBCs describe the violent coalescence of stellar-mass black holes and neutron stars in 3 stages of inspiral, merger and ringdown. As these stellar-mass binaries merge, they expel an incredible amount of gravitational radiation, resulting in GW signatures at a frequency-range that is detectable by a network of ground-based detectors on Earth [7–9]. Detectors are highly-sensitive square-kilometre laser interferometers that can detect these GW signals from their imprint on detector strain from space-time deformations. The exact morphology of the GW signal can be reconstructed from the incident strain pattern in the form of a GW waveform.

Table 1. The full PE parameter range presented alongside the subset of inferred parameters highlighted with prior boundary values.

parameter name	symbol	status	value	units
mass 1	m_1	inferred	35-80	$M_\odot$
mass 2	m_2	inferred	35-80	$M_\odot$
luminosity distance	d_L	inferred	1-3	Gpc
time of coalescence	t_0	inferred	0.65-0.85	s
inclination	Θ_{jn}	inferred	0-π	rad
polarisation	ψ	inferred	0-π	rad
right ascension	α	fixed	0	rad
declination	δ	fixed	0	rad
spins (x6)	-	fixed	0	-
phase at coalescence	ϕ_0	marginalised	0-2π	rad

Template waveforms are modelled using computational numerical relativity [10] and effective one-body formalisms [11] to solve GR equations. Creating high-quality template waveform banks and noise realisations are key to GW detection as LIGO software uses these template banks with additive noise injections to simulate real GW signals for matched filtering [12] analysis through specific `PyCBC` [13] and `GstLal` [14,15] detection pipelines. Having released data for their first two and a half observing runs (O1,O2,O3a) [16–18] and with the O3b data-release expected later this year, an increasing number of detected events promise to keep LIGO research at the forefront of new science.

1.1. Parameter Estimation

Once LIGO has detected a GW signal, the next task is to infer the parameters of the merger through a technique called parameter estimation (PE). CBCs are typically described by 15 parameters, as detailed in Table 1, with 8 intrinsic parameters for masses and spins and 7 extrinsic parameters to describe the location and orientation of the merger relative to the detector. In a similar framework to detection, these parameters can be inferred from their waveform signature, modelled against a bank of template waveforms. The optimum way to approach this inverse-problem is shown in [19] to be Bayesian inference, a technique built upon the mathematical framework of Bayesian statistics [20]. This statistical framework differs from classical statistics by approaching probability as a degree of belief relative to a combination of current data and prior knowledge, instead of current data in isolation. Bayesian statistics is based on Bayes' Theorem [21] which, applied to our inference problem, can be represented as

$$p(x|y) = \frac{\mathcal{L}(y|x)p(x)}{\mathcal{Z}}, \tag{1}$$

where x represents the inference parameters, y is the noisy GW waveform, $p(x|y)$ is the posterior probability, read as 'the probability of the parameters given the input data', $\mathcal{L}(y|x)$ is the likelihood, $p(x)$ is the prior knowledge of the parameters and $\mathcal{Z}$ is the Bayesian evidence. The Bayesian evidence can be evaluated from the likelihood and prior by

$$\mathcal{Z} = \int dx \mathcal{L}(y|x)p(x), \tag{2}$$

Table 2. Required run-times for traditional and DL sampling methods to produce benchmark $\mathcal{O}(10000)$ posterior samples. Run-times quoted for DL algorithms do not include training time.

sampler	algorithm	deep-learning	run-time (s)
`emcee` [23]	MCMC [24]	X	32070
`ptemcee` [25]	MCMC	X	24372
`dynesty` [26]	NS [27]	X	19400
`cpnest` [28]	NS	X	26202
`Nessai` [29]	NS	✓	9372
flows [30]	VI [31]	✓	2
`VItamin` [32]	VI	✓	1×10^{-1}

where $\mathcal{Z}$ is a very useful measurement to quantify the results of a Bayesian inference run [22]. However, in the framework of PE we are only interested in the shape of the posterior so $\mathcal{Z}$ becomes a normalisation constant which can be ignored. This means we can represent Bayes' Theorem with respect to GW PE as

$$p(x|y) \propto \mathcal{L}(y|x)p(x). \tag{3}$$

As the posterior describes the parameter space, the dimension of the posterior distribution matches the number of inferred parameters, and so, evaluation of this high-dimensional probability becomes infeasible analytically. Instead of trying to evaluate the probability surface, we stochastically sample from it to create an informed estimate of the posterior distribution. The main software used for LIGO PE is `Bilby` [33] and `LALInference` [34] which contain a selection of samplers operating on a Markov Chain Monte Carlo (MCMC) [23,25] or nested sampling (NS) algorithm [26,28]. NS [35] and MCMC [36–38] algorithms work by stochastically sampling from the prior distribution and assigning it a likelihood value by comparing the input data to a template waveform generated from the prior sample. Both methods then accept or reject this likelihood value according to their individual algorithmic methodology until the optimum posterior is estimated. An added benefit of NS samplers is their versatile ability to evaluate the Bayesian evidence for a sample run. These traditional sampling methods provide high quality posterior estimates but are very time-consuming due to repeated likelihood evaluates. The significant computational cost of these likelihood evaluations is demonstrated in Table 2, with NS and MCMC sampling run-times $\mathcal{O}(10)$ hours. Due to this time-consuming nature, much recent work has been focused on high-performance and high-throughput computing techniques to reduce this run-time and increase sampling speed by $\mathcal{O}$(1-2) orders of magnitude [39–45].

In the first LIGO observing run, there was a focus on accurate PE to enable new tests of relativistic astrophysical theories [46–48]. However, detecting the first binary neutron star CBC (GW170817) [49] in the second observing run, birthed a new field of multi-messenger astronomy [50], which requires fast PE for rapid sky localisation for follow-up observations of electromagnetic counterparts. This shift towards new science motivates the inception of new PE methods that allow for sub-second sampling. An ideal solution to this run-time bottleneck is the implementation of DL methods. There is a high volume of current work on implementing DL techniques into both the detection and PE pipelines [51]. A recent review [52] discusses the potential for DL implementations to transform every aspect of CBC detection, and places even further emphasis on the motivation for rapid PE. Concerning DL approaches, some works have been focused on enhancing traditional sampling methods [29] by reducing the effective number

of likelihood calculations during inference. Other implementations [30,32] look to completely replace traditional methods, circumventing the need for likelihood evaluations completely, using a technique called variational inference (VI) [31]. Once trained, these models boast sub-second run-times, as seen in Table 2.

1.2. Deep-learning Approaches

DL is a specific branch of machine learning that uses deep neural networks (DNNs). The many layers of DL algorithms make them very efficient in deciphering relevant features of raw data in an unsupervised framework, as is required for GW inference. A specific class of DL algorithms that have proven to be very useful for GW PE are deep generative models. Deep generative modelling is a technique that trains DNNs to model distributions found in training data, making it particularly useful for approximate inference [53]. The rapid progress of applied deep generative modelling research has split the field into many different specialised architectures, each with individual strengths and weaknesses. A recent review [54] acknowledges this fragmentation of the research field and looks to provide a standard framework of its current status. To do this, they present direct comparisons between architectures for important aspects in the VI pipeline, such as training speed and sample quality. Looking at these comparisons, two architectures stood out to be particularly well-suited for our specific GW PE use case.

The first architecture, normalising flows, have been well-studied as a successful GW inference candidate [29,55]. Normalising flows work by mapping the input distribution to an auxiliary normal distribution, using an invertible linear transform. This change of basis allows easy sampling at high dimensions, due to the simple structure of the auxiliary distribution. However, the restrictions on the transform to be linear and invertible, as well as the high-dimensional auxiliary distribution, make normalising flows hard to train for the parameter spaces required for PE. For this reason, we have opted to use the second architecture recommended for use in GW inference, conditional variational autoencoders (CVAEs). CVAEs are comprised of two DNNs, the first encodes the input distribution into a lower dimensional latent space distribution, whilst the second decodes this latent representation to reconstruct the original data. The conditional aspect of the architecture comes from the input of data labels in the training phase, extending to a semi-supervised framework. We present our CVAE, designed particularly for VI of GW waveforms, aptly name `VItamin`. We follow on from the peer-reviewed `VItamin` paper [32] henceforth, paper 1.

1.3. `VItamin`*: User-friendly Inference*

Presented in paper 1, `VItamin` is a CVAE that is conditioned on input noisy GW waveform data (y), which seeks to generate a proposal posterior $r_\theta(x|y)$ to reconstruct the true posterior $p(x|y)$ seen during training. The constituent DNNs are trained by minimising a cross-entropy cost function

$$H(p, r_\theta) = -\int dx\, p(x|y) \log r_\theta(x|y), \tag{4}$$

which is inherently minimised when the proposal posterior tends to the true posterior. The proposal posterior is parameterised on the output of the encoder $r_{\theta_1}(z|y)$ and decoder $r_{\theta_2}(x|y,z)$ networks as follows

$$r_\theta(x|y) = \int dz\, r_{\theta_1}(z|y) r_{\theta_2}(x|y,z). \tag{5}$$

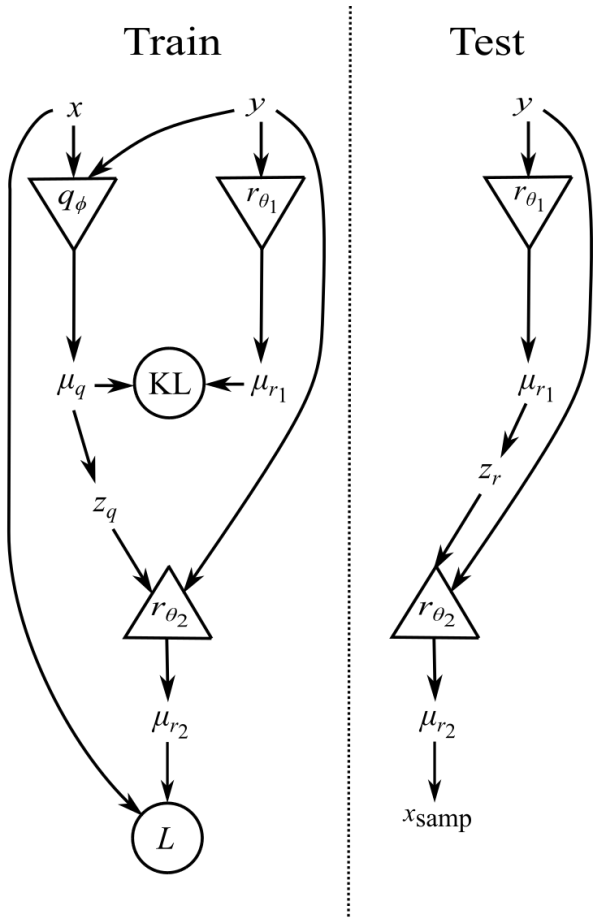


Figure 1. The structure of `VItamin` CVAE network. **Left:** Training procedure to minimise total cost function (H) from its constituent cost functions, KL-divergence (KL) and reconstruction loss (L) using Eq. 6. Noisy GW data (y) and true parameter values (x) are incident on recognition network (q_ϕ) whilst only y is incident on encoder network (r_{θ_1}). These two networks produce latent distributions (μ_{q,r_1}), a sample (z_q) is drawn from the encoder latent distribution to construct decoder network (r_{θ_2}) with corresponding distribution (μ_{r_2}) trained to reconstruct the true posterior $p(x|y)$. **Right:** `VItamin` structure in production, conditional only on y. A sample (z_r) is drawn from the encoder network latent distribution and the decoder network outputs proposal posterior samples (x_{samp}). Figure reproduced from paper 1 [32].

Here, the θ subscripts represent tunable training parameters and z represents samples drawn from the lower dimensional latent space distribution. Starting from its continuous form (Eq. 4), we can derive an approximate bound for the cross-entropy, parameterised further on a recognition network $q_\phi(z|x,y)$, based on an independet set of tunable parameters ϕ. The derivation (detailed in paper 1) uses a stochastic integral approximation to allow the cost function to be iterable over a training batch N_b. The final form of the cross-entropy cost function is given by the bound

$$H \lesssim \frac{1}{N_b}\sum_{n=1}^{N_\text{b}}\Big[\overbrace{-\log r_{\theta_2}(x_n|z_n,y_n)}^{L} + \overbrace{\text{KL}\left[q_\phi(z|x_n,y_n)||r_{\theta_1}(z|y_n)\right]}^{\text{KL}}\Big], \tag{6}$$

which is also represented graphically in Fig. 1. The total cost function is constructed as

a basic sum of 2 constituent cost functions. The Kullback–Leibler (KL)-divergence quantifies the convergence of r_{θ_1} and q_ϕ latent space distributions, minimised when both distributions are equal. The reconstruction cost L measures how well r_{θ_2} predicts the true parameters x. Concerning notation throughout this paper, inside the remit of model training and validation, $p(x|y)$ represents the true posterior formed from the training data x parameters. However, later when we start to compare with other algorithms, $p(x|y)$ represents the target posterior created using the `dynesty` sampling algorithm. Further details on our specific training procedure are presented in Section 2.1.

From Table 2, we can see the speed increase of `VItamin` compared to the traditional sampling methods and other DL frameworks. However, a consequence of this speed is limited network versatility. In this paper, we present a method to estimate the likelihoods of `VItamin` proposal posterior samples, using Monte Carlo approximations to increase the versatility of `VItamin`. This would align `VItamin` with the abilities of NS algorithms, with the end goal of estimating the Bayesian evidence, using Eq. 2. A particularly useful consequence of obtaining likelihood evaluations, is the ability to resample the proposal posterior using a technique called likelihood reweighting. A recent paper [56] showed this method to be successful in improving their approximate posterior results. In their paper, they used an approximate waveform generator [57] to construct a proposal posterior using the `Bilby` implementation of `cpnest`. They then reweighted these proposal samples according to their likelihood ratio with a more complete, but computationally expensive target waveform generator [58], using a process called sampling-importance resampling (SIR) [59]. They found that, through likelihood reweighting, they could effectively simulate drawing samples from the target posterior without ever having to run the sampling algorithm on the computationally expensive waveforms. Once we have estimated our proposal likelihoods, we look to use a similar SIR approach to improve the results of `VItamin` by reweighting our samples according to their likelihood ratio with the target posterior. We choose to use the `Bilby` implementation of `dynesty` for our target posterior due to its improved sampling speed over `cpnest`, as shown in Table 2.

The structure of the paper is as follows. In Section 2, we set out the framework for training our `VItamin` model on a reduced parameter space and describe the Monte Carlo methods used for likelihood evaluations and reweighting. In Section 3, we discuss the quality of these proposal likelihoods using qualitative self-consistency and quantitative reproducibility tests. We then go on to evaluate the suitability of the SIR method for our resampling test case, compared to the success of [56]. In Section 4, we contextualise our results in the wider field of rapid PE, by reiterating the significance of adding versatility to our `VItamin` model. We then finally set out a road map of related future work.

2. Methodology

2.1. Model Training

Our `VItamin` network was constructed as in Fig. 1, with the base network hyperparameters detailed in Table III of paper 1. Small modifications were made to this base structure, namely a reduction in latent space dimension to account for our lower dimensional sub-parameter space. We chose to infer a subset of 6 parameters in this test case (see Table 1) for two reasons. The first was to save time generating our target posterior using `dynesty`, the 10000 target samples were generated in $\mathcal{O}(3)$ hours compared to the benchmark 6 hours quoted in Table 2. The second was the strict requirement on all inferred parameters to be defined on uniform priors for likelihood calculations, as detailed in Section 2.2. Due to the sinusoidal prior on declination, we were unable to infer sky parameters so we decided on a single-detector PE run in the absence of sky-localisation capabilities. We then trained the model by minimising H in its iterable approximate form (Eq. 6). Figure 2 shows the progress of `VItamin` training with H decreasing with training iteration. The behaviour seen between 10^4 and 10^5 iterations is a consequence of

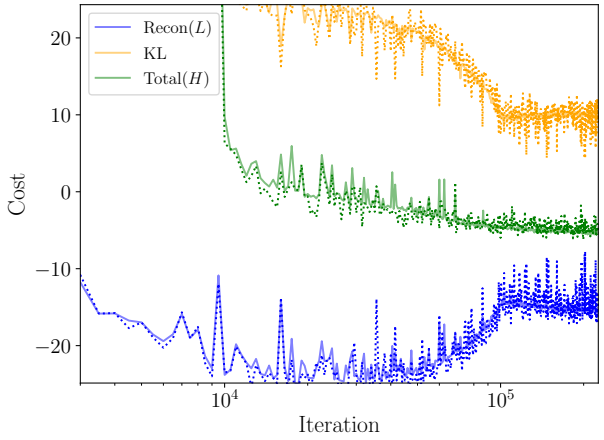


Figure 2. Cost as a function of training batch iteration. We show total cost (green) and its constituent functions KL-divergence (yellow) and reconstruction loss (blue), where total cost function is a simple sum of both constituents (Eq. 6). Solid lines correspond to training data cost and dotted lines represent validation cost. The convergence of training and validation cost indicates a lack of overfitting. The progress of this model was deliberately stopped prematurely as indicated by the total cost gradient never levelling off.

ramping L up to find the optimum gradient descent for H.

To quantify the progress of model training, it is useful to set aside some labelled data that the model does not see in the training batch called a validation set. This tests to see if the model is really improving with training iteration or has simply memorised the training data, giving a misnomer H decrease. This memorising of training data is called overfitting and occurs when the wrong configuration of training parameters is used. Most tunable training parameters, such batch size and number of epochs were kept consistent with paper 1 and the dotted lines on Fig. 2 show validation cost aligns with training cost, indicating there is no overfitting and the model is successfully training. The number of training iterations was initially set to 10^6, however training was stopped early to ensure the proposal posterior from `VItamin` $r_\theta(x|y)$ was sufficiently different from the `dynesty` target posterior $p(x|y)$ to allow resampling improvements to be more visible on plots, see Section 3.3. The fact that the H curve does not flatten in Fig. 2 indicates our model is not fully trained.

From this partially trained model, we generated 2 sets of `VItamin` samples. An initial set of 5000 was created, dictated by empirical findings that 5000 was the minimum number of samples required for smooth posterior plots. A second set of 100000 was then created to allow for sufficient resampling as detailed in Section 2.2. This set of 100000 proposal samples was compared to the 10000 target samples in Fig. 3. This partial corner plot shows the cross-section representation of the proposal and target posteriors, allowing meaningful visualisation of the complex multi-dimensional probability surfaces. Here, the black points represent the position of the true parameters from our noisy GW signal in parameter space, and the 2D contours represent the 50% and 90% credible intervals of the individual posteriors. The 100000 `VItamin` and 10000 `dynesty` samples are shown, with the order of magnitude discrepancy accounting for the increased smoothness of `VItamin` curves. For the uni-modal distributions, the proposal approximates the target well with significant convergence. For the multi-modal distributions of inclination and polarisation parameters, the proposal significantly diverges from target as a consequence of incomplete training. This creates regions of under-sampling in the 2D representations which we aim to rectify by resampling.

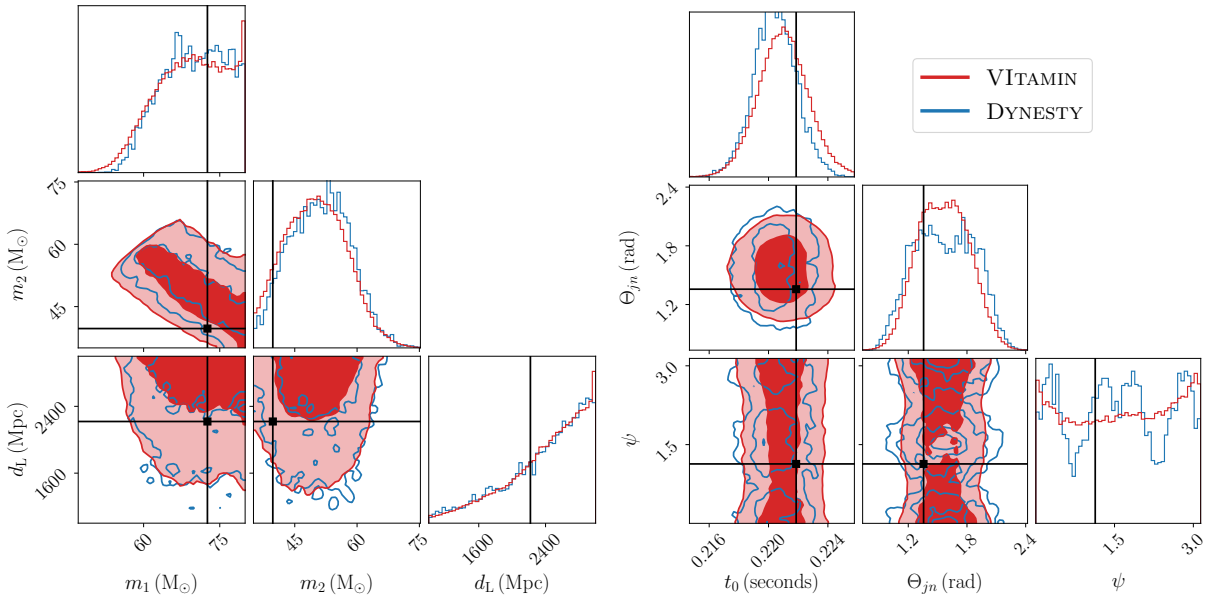


Figure 3. Partial corner plot for proposal and target posterior distributions. We show our 100000 `VItamin` proposal posterior (red) overlaid with our 10000 `dynesty` target posterior (blue) for the 6 inferred parameters identified in Table 1. True x parameters are shown in black with the 2D contours representing the 50% and 90% credible interval of the posteriors. Order of magnitude sample discrepancy accounts for increased smoothness of `VItamin` posterior with both overlapping for the uni-modal distributions. For the inclination and polarisation multi-modal target posterior, there exists regions of under-sampling from `VItamin`.

2.2. Likelihood Estimates

Monte Carlo methods provide a framework for estimating hard-to-compute expectation values of continuous functions by evaluating the convergence of stochastic samples drawn from these functions [59]. This stochastic integral evaluation was key to the derivation of the approximate cost function H (Eq. 6) from its continuous form $H(p, r_\theta)$ (Eq. 4), as detailed in paper 1. In its parametric form (Eq. 5), $r_\theta(x|y)$ is an integral that is very hard to solve analytically. By expressing it as an expectation value

$$r_\theta(x|y) = \mathbb{E}_{r_{\theta_1}(z|y)} r_{\theta_2}(x|y,z), \tag{7}$$

we can implement a Monte Carlo approximation to estimate the proposal posterior at a single point in parameter space x_i

$$r_\theta(x_i|y) \approx \frac{1}{N} \sum_{j=1}^{N} r_{\theta_{2,j}}(x_i|y,z_j)\big|_{z_j \sim r_{\theta_1}(z_j|y)}\,, \tag{8}$$

where x_i is a single output sample from our trained network. By iteratively drawing z_j samples from $r_{\theta_1}(z_j|y)$ to construct $r_{\theta_{2,j}}(x|y,z_j)$ and evaluating these decoder networks at x_i over a batch size N, we can directly estimate the proposal posterior for x_i. Due to our deliberate choice of parameters with uniform priors and, as PE does not require normalisation (Eq. 3), it is straight forward to show that this proposal posterior estimate can also be approached as a likelihood estimation

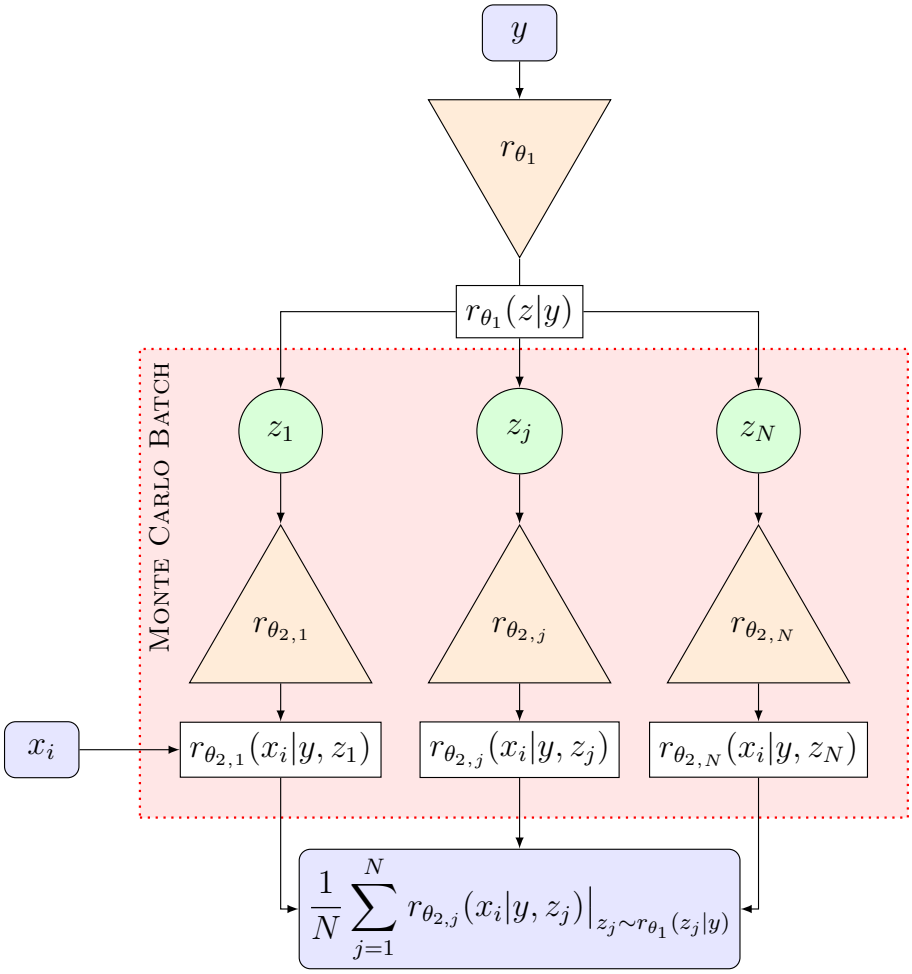


Figure 4. Flow diagram for estimation of proposal likelihood $\mathcal{L}_\theta(y|x_i)$ of a single proposal posterior sample (x_i) conditional on noisy GW data (y). Encoder network $r_{\theta_1}(z|y)$ is constructed as described in Section 1.3 and a batch of latent space samples $(z_{1:N})$ defined by Monte Carlo batch size N are drawn to construct an array of parallel decoder networks $r_{\theta_{2,1:N}}(x|y,z_{1:N})$. Decoder networks are then evaluated in parallel at x_i and combined using a Monte Carlo approximation (Eq. 8) to output a single proposal likelihood estimate. N is limited on hardware so can repeat Monte Carlo batch (red square) R times in series to obtain an effective batch size $z_{\text{eff}} = NR$.

$$r_\theta(x_i|y) \sim \mathcal{L}_\theta(y|x_i), \tag{9}$$

where $\mathcal{L}_\theta(y|x_i)$ is the proposal likelihood of a single vitamin sample x_i.

A practical approach to this Monte Carlo approximation is shown in Fig. 4. A batch of z_N samples are drawn from r_{θ_1} and ran through the `VItamin` network in parallel to reconstruct the likelihood of a single sample. We want N to be as large as possible to outweigh the stochastic nature of Monte Carlo sampling, however this is limited by hardware. For reference, our `Nvidia Tesla V100` GPU allowed a maximum $N = 65000$. A way to increase this further would be to run the Monte Carlo batch section (highlighted as a red box in Fig. 4) R times in series to create an effective batch size $z_{\text{eff}} = NR$. In practice, the r_{θ_2} networks are constructed

using `tensorflow.v1` (`tf.1`) [60] and consist of separate complex distributions for the different constituent parameters of x_i, which need to be evaluated separately and combined later. To allow this combination to be a summation rather than a more costly product, the natural log of the proposal likelihood is evaluated as a proposal log-likelihood.

Using these methods, we evaluate log-likelihoods for our set of 5000 and 100000 `VItamin` samples. However, as a consequence of our version of `VItamin` being written in discontinued `tf.1` at the time of work, there was a known memory leak issue in the code which limited our ability to maximise z_{eff}. For the 5000 sample run, we were able to load the GPU with $N = 65000$ and $R = 4$ repetitions giving $z_{\text{eff}} = 260000$, however we were only able to run the 100000 samples with $z_{\text{eff}} = 25000$. The consequences of this order of magnitude discrepancy and further software limitations are dealt with in Section 3.2. For both sets of proposal samples, the target log-likelihoods were obtained using `Bilby`, ensuring the same noisy GW data and waveform generator was used as the initial `VItamin` proposal sample run.

2.3. Likelihood Reweighting

Importance sampling is a well-established method for drawing from a target distribution using only proposal samples and relative likelihoods [61]. SIR is a Monte Carlo approximation of this technique which allows reweighting of the proposal distribution to add more importance to under-sampled regions compared to the target posterior, as discussed for the multi-modal distributions in Fig. 3. This is done by constructing a normalised weight function as a ratio of target and proposal likelihoods

$$w(y|x) \equiv \frac{\mathcal{L}(y|x)}{\mathcal{L}_\theta(y|x)}. \tag{10}$$

Following the methodology used in [56], we start with the PE version of Bayes' Theorem for the target posterior (Eq. 3), then multiply by unity with respect to the proposal likelihoods. Extracting the weight function from the right-hand side leaves only the proposal posterior

$$\begin{aligned} p(x|y) &\propto \frac{\mathcal{L}_\theta(y|x)}{\mathcal{L}_\theta(y|x)}\mathcal{L}(y|x)p(x) \\ &\propto w(y|x)\overbrace{\mathcal{L}_\theta(y|x)p(x)}^{r_\theta(x|y)}, \end{aligned} \tag{11}$$

which provides a way of drawing from the target posterior by weighted resampling of the proposal posterior. In [56], a resampling function is derived to dictate the resampling ratio, that is, how many weighted samples to extract from the proposal posterior. However, [59] suggests a simpler benchmark resampling ratio of 20 gives sufficient results. As a proof-of-concept, we resample according to this simpler benchmark to extract 5000 weighted samples from our 100000 proposal samples.

3. Results

3.1. Self-consistency

The first criterion of our proposal log-likelihoods is that they endure self-consistency. That is, higher log-likelihood values must correspond with samples in the centre of the proposal posterior and lower log-likelihood values must correspond with samples located in the tails of the distribution. This is an essential qualitative test to diagnose any internal methodological errors before comparing with target distributions. In Fig. 5, we present another partial corner

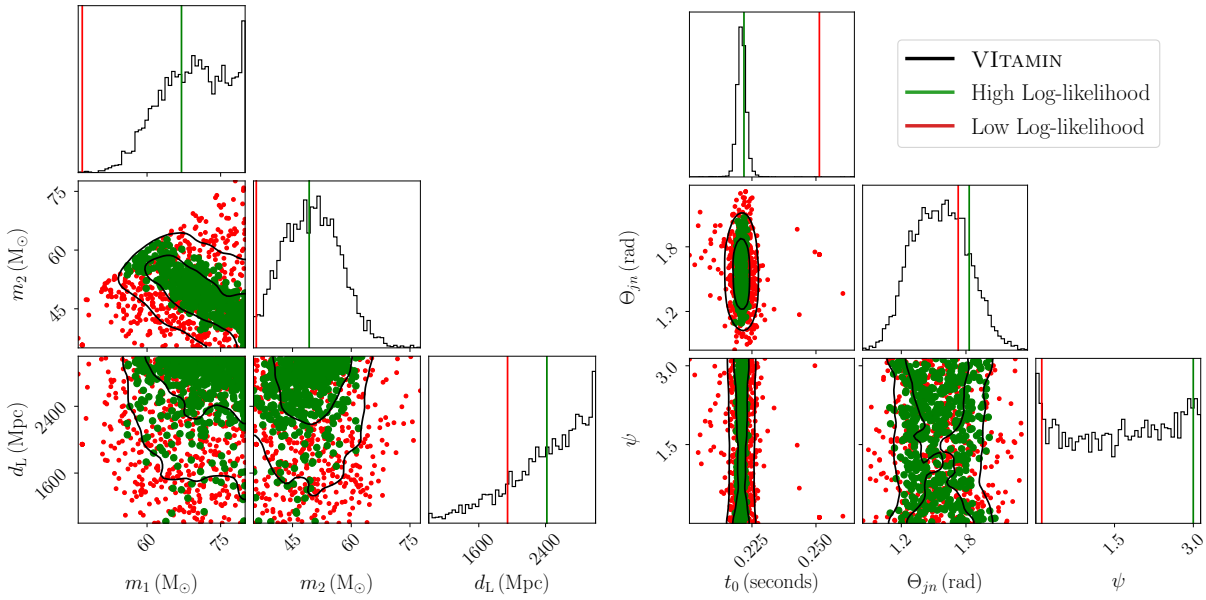


Figure 5. Partial corner plot of proposal posterior and log-likelihood estimates. We present our 5000 `VItamin` proposal posterior samples (black) with 2D contours representing the 50% and 90% confidence intervals. Samples with the highest 500 (green) and lowest 500 (red) log-likelihood estimates (calculated using Eq. 8) are plotted over the 2D contours. For the 1D distributions, the single highest and lowest log-likelihoods are plotted. Self-consistency is seen explicitly in the uni-modal distributions, whereas the multi-modal distributions only exhibit clear self-consistency in the 2D plots.

plot of our 5000 proposal posterior samples with the same 50% and 90% credible intervals in the 2D contours. We show the positions in parameter space corresponding to the highest 500 (green) and lowest 500 (red) proposal log-likelihood estimates. For the 1D distributions, the positions corresponding to the single highest and lowest log-likelihood values are plotted. For all 6 inferred parameters, there is a clear trend of higher log-likelihoods being located exclusively within the 90% credible contour, whereas there is a tendency for the lower log-likelihood values to exist in the tails of the distributions. In the 1D remit, the uni-modal distributions show a similarly successful trend, with the highest log-likelihood value corresponding to a denser region in the histogram. However for the multi-modal parameters, where the model is under-trained, the quality is not as obvious. This said, Fig. 5 qualitatively shows that the Monte Carlo methods in Section 2.2 produce self-consistent proposal log-likelihoods.

3.2. Reproducibility

The second requirement of our proposal log-likelihood estimates is that they must be reproducible. That is, they must be deterministic on sample location, ensuring independent runs on the same location in parameter space all return the same value (as is the case for our `dynesty` target log-likelihood method). The stochastic nature of drawing z_j from $r_{\theta_1}(z|y)$ in Eq. 8 sets an upper-limit on the deterministic ability of our proposal method, which contrasts with the perfect determinism of our target evaluations. However, we can approach this proposal upper-limit in practice by increasing z_{eff} to outweigh stochastic influence when obtaining the mean. The relationship between reproducibility and z_{eff} is demonstrated in Fig. 6. To produce these error profiles, we selected a single x_i sample and ran 5000 independent parallel runs of

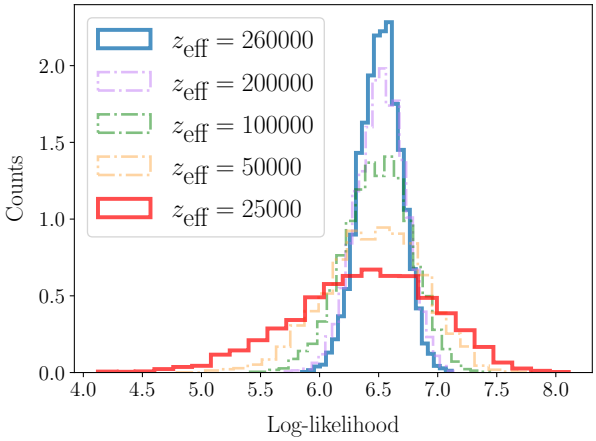


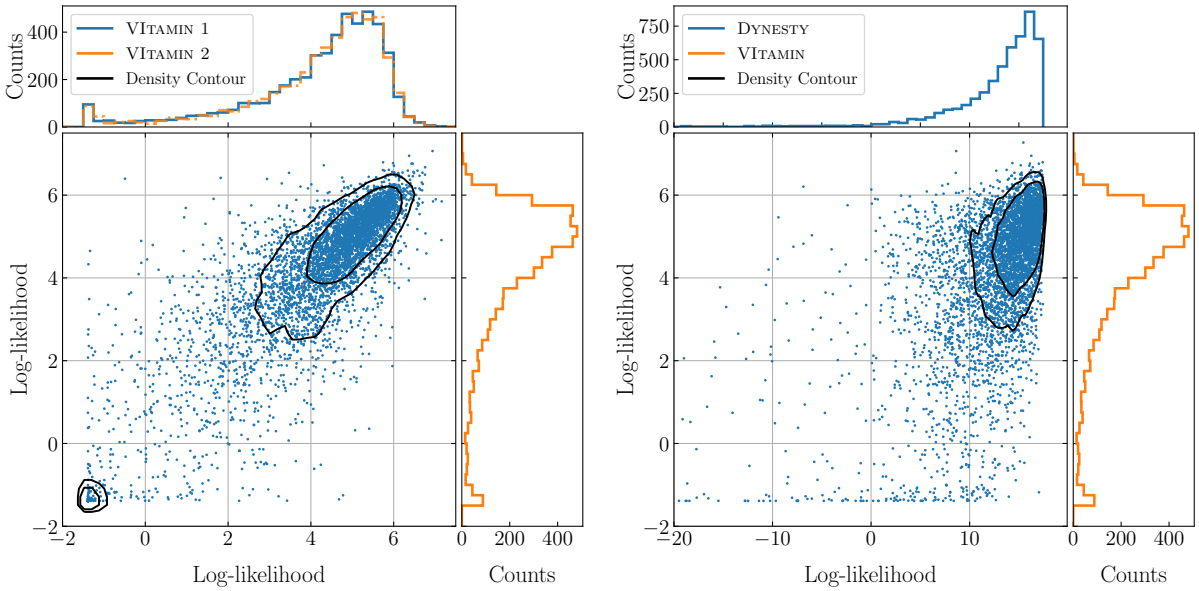
Figure 6. Single sample log-likelihood error profiles for varying Monte Carlo batch sizes (z_{eff}). We present the relationship between proposal log-likelihood reproducibility and z_{eff}. Each profile represents a set of 5000 independent log-likelihood estimates for a single point in parameter space. Modelled as normal distributions, standard deviation decreases with increased z_{eff} (Eq. 12). The highlighted z_{eff} limits represent the batch sizes used for our 5000 and 100000 proposal posterior sample runs.

our proposal log-likelihood method for varying z_{eff} values. We can see that for all quoted z_{eff} values, there is a strong agreement in mean log-likelihood value. By fitting normal distributions to these profiles we obtain the relation

$$\sigma \propto \frac{1}{\sqrt{z_{\text{eff}}}}, \tag{12}$$

as expected from basic statistics of independent samples, where σ is the standard deviation of the normal distribution fitted to the error profile. The reasoning behind the upper-limit on z_{eff} in Fig. 6 was previously explained in Section 2.2 as being due to a combination of hardware and software limitations. However, the lower-limit of $z_{\text{eff}} = 25000$ is also linked to a software limitation of `tf.1`. When evaluating the decoder network at a location in the tails of the proposal posterior (a red sample in Fig. 5), `tf.1` sets a hard minimum limit on this value, preventing meaningful analysis of lower log-likelihood values by truncating the density profile into a minimum bin. This binning of lower log-likelihoods is more pronounced for low z_{eff}, causing any $z_{\text{eff}} < 25000$ error profile (as in Fig. 6) to be severely negatively-skewed towards this lower-limit. Whilst this binning is still present in higher z_{eff}, it is outweighed by the density of higher log-likelihood estimates. This binning prevents a more quantitative error analysis as, in its absence, we could estimate the error from Fig. 6 as $\sigma/\sqrt{5000}$. We could then determine the z_{eff} required to produce a proposal log-likelihood error of ± 0.1 to align with `Bilby` benchmark standards, providing a quantitative upper-limit on proposal reproducibility. This upper-limit forms one component of scatter in the data comparing proposal and target log-likelihoods. The second component of scatter originates from the discrepancy between proposal and target posteriors.

Further analysis is done in Fig. 7(a) which presents the scatter in data of two independent proposal log-likelihood runs on the set of 5000 proposal posterior samples at $z_{\text{eff}} = 260000$. By keeping the posterior consistent between both runs, we isolate the reproducibility scatter. The binning is identified in the top plot, as the tightly correlated profiles of both proposal runs



(a) Scatter profile of two independent proposal runs

(b) Scatter profile of target and proposal runs

Figure 7. Scatter plots and profiles of independent log-likelihood runs. For our set of 5000 proposal posterior samples we ran two independent proposal log-likelihood estimates and one target log-likelihood estimate. Density contours (black) indicate the 50% and 68% confidence intervals of the scatters. Proposal runs truncate at low log-likelihood values whereas the target run log-likelihoods decrease naturally into the tails of distribution which accounts for inconsistent log-likelihood ranges between run methods.

truncate at the same lower limit, despite maximising z_{eff}. These two runs are then presented as a scatter plot with a density contour of the 50% and 68% (1-σ) credible intervals plotted in black. For comparison, a perfectly deterministic approach would result in all scatter points lying on a line of unit gradient. Although fairly symmetrical about this line, the density contours show the detrimental effect of this binning, with a high density of low log-likelihoods flattened to the same minimum value. Figure 7(b) visualises the resultant of both sources of data scatter by comparing independent runs on proposal and target posteriors. The top plot shows the target log-likelihood profile with no evidence of binning as log-likelihood evaluations continue far into the natural tails of the distribution. The inconsistent truncation between runs skews the resultant scatter plot which loses its symmetry about the line of unit gradient, with a hard lower-limit in the proposal axis. These plots demonstrate the limitations on further analysis from software binning, which need to be addressed before a more quantitative measure of proposal reproducibility can be implemented.

3.3. Importance Resampling

Having carried out likelihood reweighting on our 100000 proposal samples with $z_{\text{eff}} = 25000$ in accordance with the methodology set out in Section 2.3, we present our findings in Fig. 8. We start with the same data as in Fig. 3 and overlay our resampled posterior in green. The first notable improvement is that the resampled posterior converges much better with the target posterior for the multi-modal parameters, where the proposal posterior is under-sampled. Looking at the 2D contours, the resampled posterior matches better with the target posterior in every case. In the 1D remit, all but distance show improvement, where the resampled distance

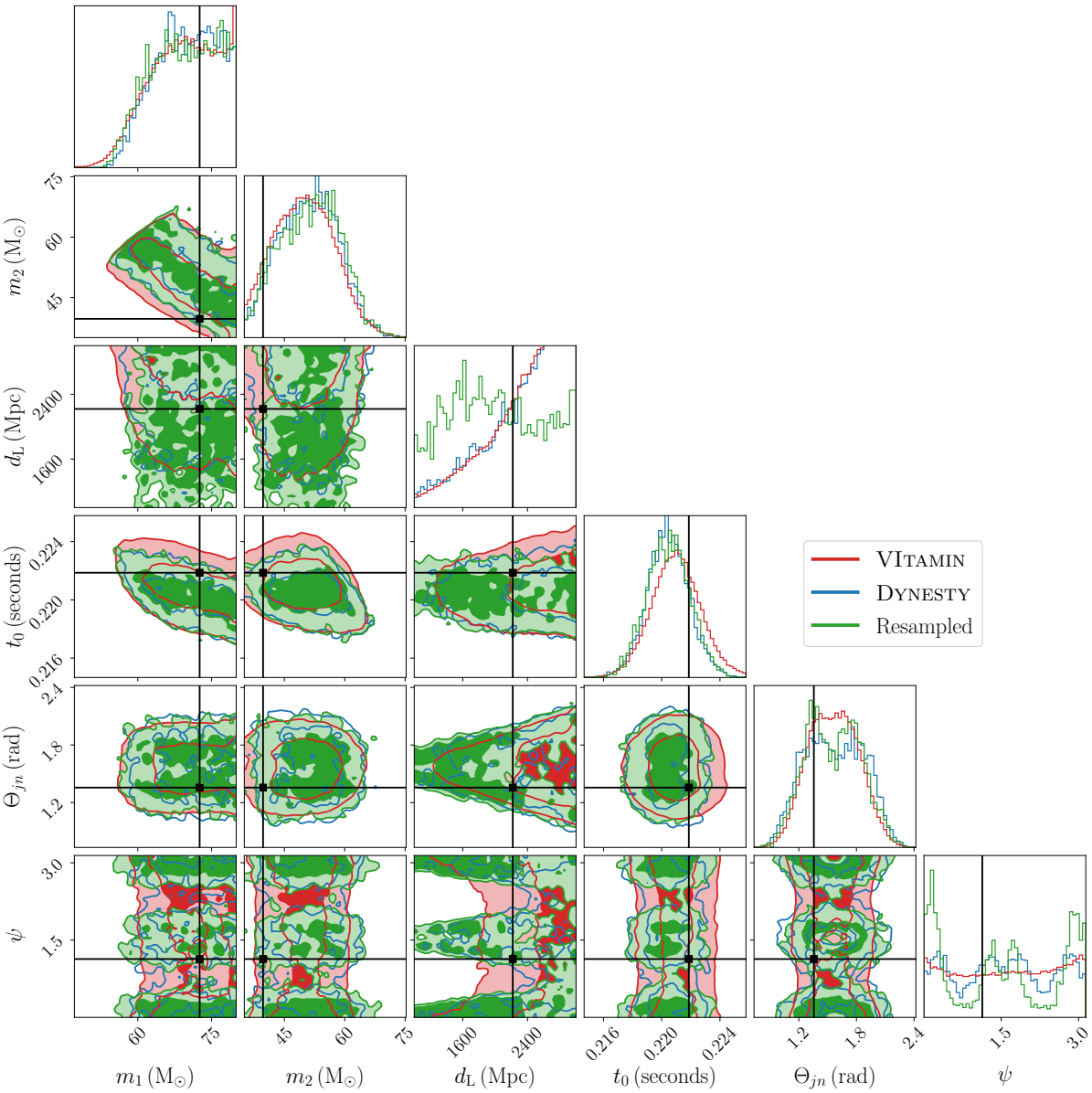


Figure 8. Full corner plot comparing the 100000 sample proposal (red), 10000 sample target (blue) and 5000 sample resampled (green) posterior distributions for the 6 inferred parameters. True parameter values are shown in black and contours represent the 50% and 90% confidence intervals in the 2D distributions. Resampled posteriors show better sampling from multi-modal parameters compared with proposal with every 2D distribution showing resampled to be better aligned with target. Resampled distance posterior is, however, uninformative despite original proposal being well-fitted.

posterior is very different to the target despite the proposal posterior being a good match. In testing, it was found that this uninformative distance posterior could be recreated using a subset of proposal posterior samples according to the 500 lowest log-likelihood locations from Fig. 5. From these findings, this distance discrepancy is likely a consequence of $z_{\text{eff}} = 25000$ limitations (Fig. 6), rather than a methodological failure. As a proof-of-concept, Fig. 8 shows great potential for likelihood reweighting as, in general, the resampled posterior matches the target better than the proposal posterior.

4. Conclusions

In this paper, we have discussed the recent focus on rapid PE in preparation for multi-messenger astronomy. We have shown this pursuit of speed to be a driving motivation for increased DL implementations in PE pipelines and have commented on the suitability of deep generative models for GW inference. We introduced our CVAE model `VItamin` (as defined in paper 1 [32]) which boasts an inference speed-up of more than 6 orders of magnitude (Table 2) compared to traditional sampling algorithms by using VI to circumvent costly likelihood calculations. The trade-off between speed and versatility between VI and NS methods was discussed with an emphasis on the importance of being able to estimate likelihoods during inference to evaluate the Bayesian evidence (Eq. 2) and carry out useful auxiliary techniques, such as likelihood reweighting. We discussed the success of a recent paper [56] in using SIR methods to reweight the proposal posterior from an approximate waveform and highlighted the compatibility of their method with our likelihood reweighting task.

In Section 2, we created a `VItamin` model for our sub-parameter space. In Fig. 3, we presented the end state of this model's incomplete training phase as a benchmark proposal posterior $r_\theta(x|y)$ to improve upon, using our `dynesty` target $p(x|y)$. We then used Monte Carlo methods to estimate $r_\theta(x|y)$ at a single position in parameter space (Eq. 8) and were able to equate this to a proposal likelihood, as a consequence our selective choice in parameters with flat priors (Eq. 9). We then described the SIR framework required to reweight our proposal samples using the proposal likelihood estimates, following the methodology of [56] and the fixed resample ratio of 20 suggested in [59].

In Section 3, our proposal likelihoods passed their qualitative self-consistency test (Fig. 5) which is a mandatory internal check needed before comparing with other algorithms. We showed that increasing the Monte Carlo batch size (z_{eff}) increased the reproducibility of our method (Fig. 6), but attempts at more quantitative error analysis were prevented by software limitations originating in the antiquated `tf.1` framework that `VItamin` was written in at the time of work. These software limitations were two-fold. First, an unavoidable memory leak placed an upper-limit on z_{eff}, capping further efforts to suitably minimise error. Secondly, inherent binning of low likelihood values in the code made in-depth profile analyses of proposal likelihood runs impossible and limited the ability to compare values with the target method due to inconsistent truncation (Fig. 7). A priority of future work is to re-code our Monte Carlo algorithm (Fig. 4) in the updated `keras.tensorflow` framework of `VItamin` which we predict to fix both software limitations. Once fixed, a complete error analysis can be applied to our proposal method to derive the minimum z_{eff} required to produce a log-likelihood error of ± 1 which is comparable to the quoted tolerance in other sampling algorithms.

In Fig. 8, we present our final reweighted results which align better with our target posterior, as desired. This plot shows the specific capabilities of resampling to improve multi-modal posteriors, whose topology was previously missed. The resampled distance posterior became uninformative but we were able to replicate this discrepancy in a fixed low log-likelihood framework. Considering that these log-likelihoods were capped at $z_{\text{eff}} = 25000$, due to the aforementioned software limitations, we are confident that these results successfully show the significant potential of Monte Carlo methods in increasing the versatility of DL for GW inference.

Taking these findings from proof-of-concept to production status requires a significant overhaul in methodology and flexibility. Firstly, we need to expand our posterior dimension to encompass the entire 15 parameter space. This means working with non-flat priors, and so requiring a more involved approach than Eq. 9 in converting from proposal posterior to likelihood. Secondly, the resampling speed must be dramatically increased to be useful in rapid PE, as computing our 100000 proposal likelihoods required a run-time of 6 hours. A straightforward first step towards this is to decrease the number of original proposal posterior samples from 100000. However, if we desire a meaningful number of reweighted samples, we must update our primitive fixed sampling ratio approach. An improved SIR framework is presented in [62] that claims to significantly reduce the required sampling ratio by minimising resampling bias. Expanding the parameter space and reducing computing time, lend themselves to be the intuitive next steps in this process, with potential for this method to help accelerate the practical realisation of multi-messenger astronomy.

Acknowledgements
I would like to thank my supervisor Chris Messenger for his support and guidance throughout this project. I would also like to thank Hunter Gabbard for his insight and patience explaining and modifying `VItamin` for my use case. I am also indebted to Michael Williams and Daniel Williams for their continued technical support. Finally, I would like to express my gratitude toward the entire University of Glasgow IGR-data analysis group for making me feel very welcome over the past year.

References

[1] A. Einstein, *Sitzungsber. K. Preuss. Akad. Wiss*, vol. 1, no. 688, 1916.

[2] J. Aasi, B. P. Abbott, R. Abbott, T. Abbott, M. R. Abernathy, K. Ackley, C. Adams, T. Adams, P. Addesso, and et al., "Advanced ligo," *Classical and Quantum Gravity*, vol. 32, no. 7, p. 074 001, Mar. 2015, ISSN: 1361-6382. DOI: 10.1088/0264-9381/32/7/074001. [Online]. Available: http://dx.doi.org/10.1088/0264-9381/32/7/074001.

[3] B. P. Abbott, R. Abbott, T. Abbott, M. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. Adhikari, *et al.*, "Observation of gravitational waves from a binary black hole merger," *Physical review letters*, vol. 116, no. 6, p. 061 102, 2016.

[4] J. Bayley, G. Woan, and C. Messenger, "Soap: A generalised application of the viterbi algorithm to searches for continuous gravitational-wave signals," *arXiv preprint arXiv:1903.12614*, 2019.

[5] J. Bayley, C. Messenger, and G. Woan, "Robust machine learning algorithm to search for continuous gravitational waves," *Physical Review D*, vol. 102, no. 8, p. 083 024, 2020.

[6] C. Dreissigacker, R. Sharma, C. Messenger, R. Zhao, and R. Prix, "Deep-learning continuous gravitational waves," *Physical Review D*, vol. 100, no. 4, p. 044 009, 2019.

[7] F. Acernese, M. Agathos, K. Agatsuma, D. Aisa, N. Allemandou, A. Allocca, J. Amarni, P. Astone, G. Balestri, G. Ballardin, *et al.*, "Advanced virgo: A second-generation interferometric gravitational wave detector," *Classical and Quantum Gravity*, vol. 32, no. 2, p. 024 001, 2014.

[8] "Kagra: 2.5 generation interferometric gravitational wave detector," *Nature Astronomy*, vol. 3, no. 1, pp. 35–40, Jan. 2019, ISSN: 2397-3366. DOI: 10.1038/s41550-018-0658-y. [Online]. Available: http://dx.doi.org/10.1038/s41550-018-0658-y.

[9] C. S. UNNIKRISHNAN, "Indigo and ligo-india: Scope and plans for gravitational wave research and precision metrology in india," *International Journal of Modern Physics D*, vol. 22, no. 01, p. 1 341 010, Jan. 2013, ISSN: 1793-6594. DOI: 10.1142/s0218271813410101. [Online]. Available: http://dx.doi.org/10.1142/S0218271813410101.

[10] F. Pretorius, "Evolution of binary black-hole spacetimes," *Phys. Rev. Lett.*, vol. 95, p. 121 101, 12 Sep. 2005. DOI: 10.1103/PhysRevLett.95.121101. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.95.121101.

[11] A. Buonanno and T. Damour, "Effective one-body approach to general relativistic two-body dynamics," *Phys. Rev. D*, vol. 59, p. 084 006, 8 Mar. 1999. DOI: 10.1103/PhysRevD.59.084006. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.59.084006.

[12] B. J. Owen and B. S. Sathyaprakash, "Matched filtering of gravitational waves from inspiraling compact binaries: Computational cost and template placement," *Phys. Rev. D*, vol. 60, p. 022 002, 2 Jun. 1999. DOI: 10.1103/PhysRevD.60.022002. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.60.022002.

[13] S. A. Usman, A. H. Nitz, I. W. Harry, C. M. Biwer, D. A. Brown, M. Cabero, C. D. Capano, T. D. Canton, T. Dent, S. Fairhurst, M. S. Kehl, D. Keppel, B. Krishnan, A. Lenon, A. Lundgren, A. B. Nielsen, L. P. Pekowsky, H. P. Pfeiffer, P. R. Saulson, M. West, and J. L. Willis, "The PyCBC search for gravitational waves from compact binary coalescence," *Classical and Quantum Gravity*, vol. 33, no. 21, p. 215 004, Oct. 2016. DOI: 10.1088/0264-9381/33/21/215004.

[14] C. Messick, K. Blackburn, P. Brady, P. Brockill, K. Cannon, R. Cariou, S. Caudill, S. J. Chamberlin, J. D. E. Creighton, R. Everett, and et al., "Analysis framework for the prompt discovery of compact binary mergers in gravitational-wave data," *Physical Review D*, vol. 95, no. 4, Feb. 2017, ISSN: 2470-0029. DOI: 10.1103/physrevd.95.042001. [Online]. Available: http://dx.doi.org/10.1103/PhysRevD.95.042001.

[15] S. Sachdev, S. Caudill, H. Fong, R. K. L. Lo, C. Messick, D. Mukherjee, R. Magee, L. Tsukada, K. Blackburn, P. Brady, P. Brockill, K. Cannon, S. J. Chamberlin, D. Chatterjee, J. D. E. Creighton, P. Godwin, A. Gupta, C. Hanna, S. Kapadia, R. N. Lang, T. G. F. Li, D. Meacher, A. Pace, S. Privitera, L. Sadeghian, L. Wade, M. Wade, A. Weinstein, and S. L. Xiao, *The gstlal search analysis methods for compact binary mergers in advanced ligo's second and advanced virgo's first observing runs*, 2019. arXiv: 1901.08580 [gr-qc].

[16] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, *et al.*, "Binary black hole mergers in the first advanced ligo observing run," *Phys. Rev. X*, vol. 6, p. 041 015, 4 Oct. 2016. DOI: 10.1103/PhysRevX.6.041015. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.6.041015.

[17] B. P. Abbott, R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, R. X. Adhikari, V. B. Adya, C. Affeldt, M. Agathos, K. Agatsuma, N. Aggarwal, O. D. Aguiar, L. Aiello, A. Ain, P. Ajith, G. Allen, *et al.*, "Gwtc-1: A gravitational-wave transient catalog of compact binary mergers observed by ligo and virgo during the first and second observing runs," *Phys. Rev. X*, vol. 9, p. 031 040, 3 Sep. 2019. DOI: 10.1103/PhysRevX.9.031040. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.9.031040.

[18] R. Abbott, T. Abbott, S. Abraham, F. Acernese, K. Ackley, A. Adams, C. Adams, V. Adya, C. Affeldt, M. Agathos, K. Agatsuma, N. Aggarwal, O. Aguiar, L. Aiello, A. Ain, P. Ajith, S. Akcay, *et al.*, *Gwtc-2: Compact binary coalescences observed by ligo and virgo during the first half of the third observing run*, Oct. 2020.

[19] A. C. Searle, P. J. Sutton, and M. Tinto, "Bayesian detection of unmodeled bursts of gravitational waves," *Classical and Quantum Gravity*, vol. 26, no. 15, 155017, p. 155 017, Aug. 2009. DOI: 10.1088/0264-9381/26/15/155017. arXiv: 0809.2809 [gr-qc].

[20] T. Bayes and n. Price, "Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s," *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370–418, 1763. DOI: 10.1098/rstl.1763.0053.

[21] *Bayes' theorem*, Jan. 2009. [Online]. Available: https://www.oxfordreference.com/view/10.1093/acref/9780199233991.001.0001/acref-9780199233991-e-3561.

[22] D. Sivia and J. Skilling, *Data analysis: a Bayesian tutorial*. OUP Oxford, 2006.

[23] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, "Emcee: The mcmc hammer," *PASP*, vol. 125, pp. 306–312, 2013. DOI: 10.1086/670067. eprint: 1202.3665.

[24] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," 1970.

[25] W. Vousden, W. M. Farr, and I. Mandel, "Dynamic temperature selection for parallel-tempering in Markov chain Monte Carlo simulations," 2015. DOI: 10.1093/mnras/stv2422. eprint: arXiv:1501.05823.

[26] J. S. Speagle, *dynesty: A Dynamic Nested Sampling Package for Estimating Bayesian Posteriors and Evidences*, 2019. [Online]. Available: https://arxiv.org/abs/1904.02180v1.

[27] J. Skilling, "Nested sampling for general bayesian computation," *Bayesian Anal.*, vol. 1, no. 4, pp. 833–859, Dec. 2006. DOI: 10.1214/06-BA127. [Online]. Available: https://doi.org/10.1214/06-BA127.

[28] W. D. Pozzo and J. Veitch, *Cpnest*, 2019. [Online]. Available: DOI:10.5281/zenodo.835874.

[29] M. J. Williams, J. Veitch, and C. Messenger, "Nested sampling with normalising flows for gravitational-wave inference," *arXiv preprint arXiv:2102.11056*, 2021.

[30] S. R. Green and J. Gair, *Complete parameter inference for gw150914 using deep learning*, 2020. arXiv: 2008.03312 [astro-ph.IM].

[31] F. Tonolini, A. Lyons, P. Caramazza, D. Faccio, and R. Murray-Smith, *Variational inference for computational imaging inverse problems*, To appear in JMLR, 2019. eprint: arXiv:1904.06264.

[32] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, "Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy," *arXiv preprint arXiv:1909.06296*, 2019.

[33] G. Ashton, M. Huebner, P. D. Lasky, Colm Talbot, K. Ackley, Sylvia Biscoveanu, Q. Chu, A. Divarkala, P. J. Easter, Boris Goncharov, Francisco Hernandez Vivanco, J. Harms, M. E. Lower, Grant D. Meadors, D. Melchor, E. Payne, M. D. Pitkin, J. Powell, N. Sarin, Rory J. E. Smith, and E. Thrane, "Bilby: A user-friendly Bayesian inference library for gravitational-wave astronomy," *Astrophys. J. Supp.*, vol. 241, p. 27, 2019.

[34] J. Veitch *et al.*, "Parameter estimation for compact binaries with ground-based gravitational-wave observations using the lalinference software library," *Phys. Rev. D*, vol. 91, p. 042 003, 2015.

[35] J. Skilling, "Nested Sampling," *AIP Conf. Proc.*, vol. 735, no. 1, pp. 395–405, 2004. DOI: 10.1063/1.1835238. [Online]. Available: http://aip.scitation.org/doi/abs/10.1063/1.1835238%20https://projecteuclid.org/euclid.ba/1340370944.

[36] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953. DOI: 10.1063/1.1699114. [Online]. Available: https://doi.org/10.1063/1.1699114.

[37] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970, ISSN: 00063444.

[38] D. W. Hogg and D. Foreman-Mackey, "Data analysis recipes: Using Markov Chain Monte Carlo," *Astrophys. J. Supp.*, vol. 236, p. 11, 2018.

[39] R. Smith, S. E. Field, K. Blackburn, C.-J. Haster, M. Pürrer, V. Raymond, and P. Schmidt, "Fast and accurate inference on gravitational waves from precessing compact binaries," *Phys. Rev. D*, vol. 94, p. 044 031, 2016.

[40] M. Pürrer, "Frequency-domain reduced order models for gravitational waves from aligned-spin compact binaries," *Class. Quant. Grav.*, vol. 31, p. 195 010, 2014.

[41] C. Talbot, R. Smith, E. Thrane, and G. B. Poole, "Parallelized Inference for Gravitational-Wave Astronomy," 2019, arxiv/1904.02863.

[42] D. Wysocki, R. O'Shaughnessy, Y.-L. L. Fang, and J. Lange, "Accelerating parameter inference with graphics processing units," 2019, arxiv/1902.04934.

[43] R. Smith, S. E. Field, K. Blackburn, C.-J. Haster, M. Pürrer, V. Raymond, and P. Schmidt, "Fast and accurate inference on gravitational waves from precessing compact binaries," *Physical Review D*, vol. 94, no. 4, 044031, p. 044 031, Aug. 2016. DOI: 10.1103/PhysRevD.94.044031. arXiv: 1604.08253 [gr-qc].

[44] D. Wysocki, R. O'Shaughnessy, J. Lange, and Y.-L. L. Fang, "Accelerating parameter inference with graphics processing units," *Physical Review D*, vol. 99, no. 8, 084026, p. 084 026, Apr. 2019. DOI: 10.1103/PhysRevD.99.084026. arXiv: 1902.04934 [astro-ph.IM].

[45] C. Talbot, R. Smith, E. Thrane, and G. B. Poole, "Parallelized inference for gravitational-wave astronomy," *Physical Review D*, vol. 100, no. 4, 043030, p. 043 030, Aug. 2019. DOI: 10.1103/PhysRevD.100.043030. arXiv: 1904.02863 [astro-ph.IM].

[46] B. P. Abbott *et al.*, "A gravitational-wave standard siren measurement of the Hubble constant," *Nature*, vol. 551, p. 85, 2017.

[47] ——, "Tests of General Relativity with GW150914," *Phys. Rev. Lett.*, vol. 116, p. 221 101, 2016.

[48] B. F. Schutz, "Determining the Hubble constant from gravitational wave observations," *Nature*, vol. 323, p. 310, 1986.

[49] B. P. Abbott *et al.*, " Properties of the binary neutron star merger GW170817," *Phys. Rev. X*, vol. 9, p. 011 001, 2019.

[50] ——, "Multi-messenger observations of a binary neutron star merger," *Astrophys. J. Lett.*, vol. 848, p. L12, 2017.

[51] H. Gabbard, M. Williams, F. Hayes, and C. Messenger, "Matching matched filtering with deep networks for gravitational-wave astronomy," *Phys. Rev. Lett.*, vol. 120, p. 141 103, 14 Apr. 2018. DOI: 10.1103/PhysRevLett.120.141103. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.120.141103.

[52] E. Cuoco, J. Powell, M. Cavaglià, K. Ackley, M. Bejger, C. Chatterjee, M. Coughlin, S. Coughlin, P. Easter, R. Essick, *et al.*, "Enhancing gravitational-wave science with machine learning," *arXiv preprint arXiv:2005.03745*, 2020.

[53] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, *Handling incomplete heterogeneous data using VAEs*, 2018. eprint: arXiv:1807.03653.

[54] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models," *arXiv preprint arXiv:2103.04922*, 2021.

[55] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *arXiv preprint arXiv:1912.02762*, 2019.

[56] E. Payne, C. Talbot, and E. Thrane, "Higher order gravitational-wave modes with likelihood reweighting," *Physical Review D*, vol. 100, no. 12, p. 123 017, 2019.

[57] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé, "Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era," *Phys. Rev. D*, vol. 93, p. 044 007, 2016.

[58] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, L. E. Kidder, and H. P. Pfeiffer, *Phys. Rev. D*, vol. 99, p. 064 045, 6 2019.

[59] A. S. C. Brian, "The resampling weights in sampling-importance resampling algorithm," Ph.D. dissertation, The Chinese University of Hong Kong, 2006.

[60] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[61] K. Heine, "Unified framework for sampling/importance resampling algorithms," in *2005 7th International Conference on Information Fusion*, vol. 2, 2005, 6 pp.-. DOI: 10.1109/ICIF.2005.1592027.

[62] Ø. Skare, E. Bølviken, and L. Holden, "Improved sampling-importance resampling and reduced bias importance sampling," *Scandinavian Journal of Statistics*, vol. 30, no. 4, pp. 719–737, 2003, ISSN: 03036898, 14679469. [Online]. Available: http://www.jstor.org/stable/4616798.