Reviewers' comments:

Reviewer #1 (Remarks to the Author):

I would like to recommend this paper for publication after the following item is considered by the authors.

While the authors have taken great care to document their experiments, there are too many parameters set in a typical PE run to reproduce scientific findings using values quoted in the table. For example, table 4 shows what might be considered to be the most important sampler hyperparameters, but there are many more. For example, the dynasty sampler (when called through Bilby at least) has known issues when using reflective boundary conditions on the prior, and the number of auto-correlation times to use before accepting a point (nact in the code) can greatly impact the results. From my own experience, I've needed to run the code using nact values as high as 25 before getting converged results. My suggestion would be to make the input files used in these experiments publicly available either as part of the paper or hosted somewhere like gitlab, github, or bitbucket.

We thank the referee for their time and helpful comments while reviewing this manuscript. We have made the entirety of the code used to produce the results (and bilby posteriors) publicly available at the following GitHub repository (https://github.com/hagabbar/vitamin_c). We provide the input test data waveforms as well as the trained ML model on the Harvard Dataverse at the following publicly available link (https://doi.org/10.7910/DVN/DECSMV). Training/Test/Validation data may be produced exactly using the code in the linked GitHub. One can exactly reproduce all test sample posteriors/waveforms by using the command (python run_vitamin.py --gen_test True) from the linked GitHub repository.

Reviewer #2 (Remarks to the Author):

The revised draft has addressed my main initial concerns. In particular:
1) The introduction has been revised and I believe it is much more accessible now.
2) The details of the CVAE model have been substantially revised and the results now incorporate my suggestions and complementary modifications. I am pleased to see the new results. The model is now tailored to the problem it is trying to solve (by incorporating the correct support of the target density). The revised model has substantially strengthened the novelty factor of the paper.

We thank the referee for their time and helpful comments while reviewing this manuscript.

Reviewer #4 (Remarks to the Author):

Performing accurate Bayesian inference on gravitational-wave signals is a corner stone of gravitational-wave astronomy. Current methods (mcmc, nested sampling etc…) carry a notoriously high computational cost which will soon present a problem as the number of

interesting observations increases; as the complexity of signal models increases (as it generally tends to), and as detectors push down their low-frequency sensitivity so that the duration of observable signals increases.

The paper presents a novel approach to the problem of performing Bayesian inference using machine learning. The authors describe how a conditional variational autoencoder can be trained on known posterior/likelihoods, together with strain data, to quickly produce posterior densities given new strain data. Most impressive is the short time in which the network can produce results: around six orders of magnitude faster than the current state of the art. The authors demonstrate their method explicitly by performing inference on the parameters of synthetic binary black hole signals, and compare their results to those obtained using well established samplers, e.g., dynesty and emcee.

The paper is well written and clear. The authors have made efforts to compare their results to those obtained using well established methods, though I dispute some of their conclusions about the accuracy of their results. Most impressively, the authors have applied their machine learning method to an astrophysical interesting subset of the parameter space of binary black holes; 7 parameters, including the masses and sky location of the binary (2 parameters are marginalized over). This greatly expands on other work in the field which typically focuses on a small subset of parameters. I particularly enjoyed the clear description of the ML algorithm, which often is glossed over in other ML papers.

In compiling my review, I have considered whether the paper meets Nature's three criteria (https://www.nature.com/nature/for-authors/editorial-criteria-and-processes) for publication, namely that the results:

• report original scientific research (the main results and conclusions must not have been published or submitted elsewhere)
• are of outstanding scientific importance
• reach a conclusion of interest to an interdisciplinary readership.

The results are certainly original. The results reach a conclusion that is broadly interesting to the gravitational-wave astronomy community: they demonstrate that ML approaches *may* be promising for gravitational-wave observations in the future. However, I do not believe that the results constitute outstanding scientific importance. While I believe the work is undoubtably novel and an extremely impressive technical achievement, I have a number of issues with:

1 scope of the analysis
2. the accuracy of the results

We have taken on board the referee's concerns regarding these 2 key issues and are confident that we have now addressed them. The quality of results has significantly improved as have our methods of comparison. We now analyse the complete binary black hole parameter space and have increased the data sampling rate by a factor of 4.

Scope of the analysis

The authors demonstrate that the method can produce the 7-dimensional posterior density describing their synthetic black hole signals. In doing so, they marginalize over coalescence- and polarisation phase. My biggest issue is that their numerical experiments do not address the full problem of binary black hole parameter estimation: estimating the full set of 15 parameters describing the binary. The authors suggest that their method can be readily extended to the full set of parameters:

“We note that with the exception of requiring one-dimensional convolutional layers and an increase in the amount of training data to efficiently deal with a multi-detector analysis, the network complexity has not increased with the dimensionality of the physical parameter space nor with the sampling rate of the input data. We therefore do not anticipate that extending the parameter space to lower masses and including component spin parameters will be problematic.”

However I am unconvinced that the authors have demonstrated that including component spin parameters will not be problematic. I have two reasons for this:

Firstly, real gravitational-wave signals carry higher order modes beyond the leading-order modes which are the only ones present in their model, IMRPhenomPv2. It is the lack of modes beyond the dominant mode which allows one to explicitly marginalize over coalescence phase with IMRPhenomPv2. This cannot be done for models which include higher order modes and/or generic precessing spin effects. At the very least the authors would need to demonstrate that they are able to return accurate posteriors without explicitly marginalizing over phase.

We have shown in the results seen in the updated manuscript that our method is able to reconstruct Bayesian posteriors to within a reasonable degree of accuracy with respect to other sampler approaches up to a 14-dimensional parameter space (excluding phase). All bilby Bayesian samplers give the option to marginalise out the phase parameter and we have chosen to use this option primarily because it improves overall stability and convergence time for those samplers (see https://dcc.ligo.org/public/0102/T1300326/001/margphi.pdf and arxiv:1811.02042). We also note that in standard parameter estimation during an observation run, phase is not typically considered in inference since it’s measurement provides little intrinsic astrophysical value.

Whilst we compare against phase marginalised results on all other parameters, in this draft we have allowed our approach to also output samples from the phase but we do not compare this parameter to the existing samplers. Internal marginalisation over any parameter in the CVAE approach is as simple as excluding it as a predicted parameter but retaining it as a variable parameter in the training data, ensuring that it is sampled from the desired prior.

We agree with the referee on the issue of higher modes but we do not use higher mode waveforms in our analysis and only plan to include them in future work. We wish to stress that the phase marginalisation within the CVAE used in the previous draft was nothing more than omitting phase as an output and made no assumptions about the waveform approximant being used.

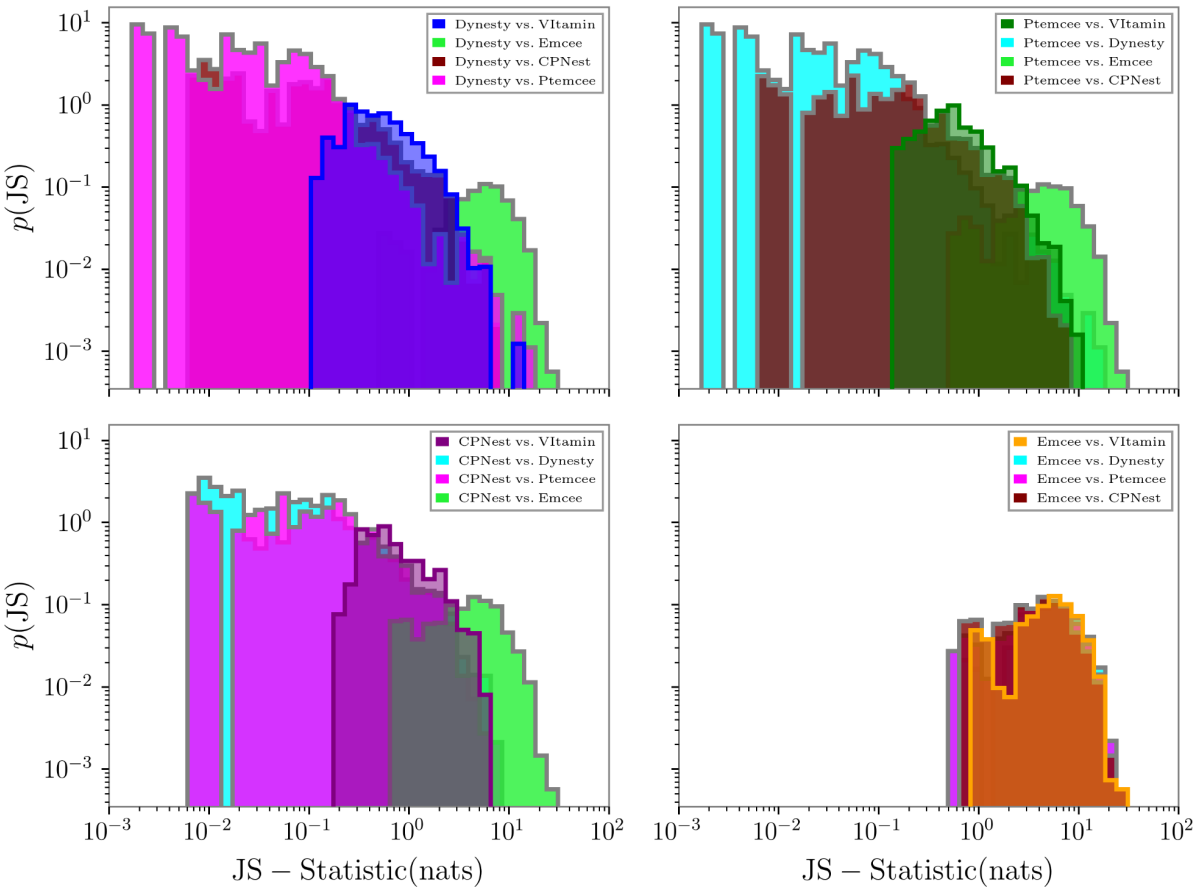
Secondly, I cannot conclude that the ML method is necessarily robust if they were to include precessing spin parameters. This is because there may be correlations between precessing spin parameters, phase at coalescence and other parameters, e.g., component masses can correlate with spin-tilt angles; mass ratio correlates with aligned-spin components. I believe that tackling this full 15D parameter estimation problem (explicitly sampling phase at coalescence) would be necessary to warrant outstanding scientific importance.

We have included all 6 spin parameters in our updated analysis and have shown in our results that we are able to accurately predict precessing spin parameters - in fact, for the majority of the additional 6 parameters we arguably perform our best inference. The trained network is clearly able to model the complex correlations between the spin and other parameters. As mentioned earlier, whilst we do not provide direct comparisons with regards to phase, we do explicitly sample from the phase parameter.

Accuracy of the results

The authors provide three quantitative measures of the accuracy of their results: they compare posteriors from VItamin to dynesty and ptemcee, perform a PP test (as is standard in internal LIGO peer review), and compute KL-statistics between VItamin and the other samplers. The PP tests suggest that 1D posterior densities are robust, which is nice to see. However, the PP test does not address whether the full N-dimensional PDF is itself unbiased, and I have concerns that they are in fact biased in a way that is not tested by the PP test or the KL test.

Although we agree with the referee that PP tests do not necessarily address whether the full N-dimensional PDF is itself unbiased, the KL divergence figure of merit (applied to the 14 dimensional space excluding phase) does directly address this issue - we have attached the updated version of this. We would add that such high dimensional comparisons have not been performed between the existing samplers in the literature and instead single dimensional comparisons have been the norm. We have also now adopted this standard and have replaced the KL distribution (Fig. 5) with the more accurate and trustworthy JS divergence distribution for each parameter separately. We would also like to clarify that the KL divergence previously calculated for Fig. 5 was incorrectly labelled and is actually the JS-divergence.



My first concern is with the benchmark results themselves. There are clear differences between dynesty and ptemcee in Figure 2. To understand VItamin's results I would first need to see consistency between benchmark results. The authors state:

"It is also evident that whilst we refer to the Bilby sampler results as benchmark cases, different existing samplers do not perfectly agree with each other. For each of our 256 test cases we see equivalent levels of disparity between pairs of benchmark samplers and between any benchmark sampler and our CVAE results"

However, I am not convinced that this should be the case. A large amount of effort has been undertaken within the LIGO/Virgo collaboration to converge on sampler settings with bilby, which accurately match results obtained using the code LALInference (bilby's predecessor). The comparisons (summarised in the bilby GWTC-1 paper https://arxiv.org/abs/2006.00714) demonstrate that both 1D and 2D credible intervals match well between the samplers. There is a clear disparity between the 2D credible intervals of ptemcee and dynesty which makes picking a benchmark set of results confusing. I suspect that the real issue here is that ptemcee is simply not converged fully. I am basing this on the fact that the sampler settings for dynesty in Table IV are very aggressive; much more so than for production analyses in the LVC. Certainly it is the

case with, e.g., code reviews, that converging on good settings is challenging, and so far this has only been achieved with dynesty.

Having discussed with experts in the Bayesian sampler community, it is evident that Bayesian samplers are certainly not guaranteed to converge to the same results. Full convergence in many cases may require much fine tuning over many iterations for each individual run. Although we do not fine tune Bayesian benchmark samplers for each sampler and each individual test case, we do use settings which have been recommended to us by bilby developers and outside experts for each respective Bayesian sampler. Both the Dynesty and CPNest samplers have a tolerance threshold (change in the log evidence from one proposal to the next) which guarantees a certain level of convergence. We use the recommended tolerance level of 0.1 for both nested samplers. For the MCMC samplers, Emcee performs poorly, but is known to have difficulties with convergence within the community. There are a handful of Ptemcee test cases (~5 of the 250) which show some minor indication of incomplete convergence, but after careful review we have determined that a lengthier burn-in period does not significantly improve the resulting posteriors.

We appreciate the referee's use of the bilby GWTC-1 paper but the few comparison posterior plots shown in that paper refer to comparisons between bilby and lalinference, not between the different bilby samplers. In those cases the specific bilby sampler is Dynesty which, in parallel to those studies, has had it's settings fine-tuned such that it is the only trusted (reviewed) successor to lalinference in the LIGO-Virgo-Kagra collaboration. We also note that the bilby GWTC-1 paper does not compare those samplers using more than 2D credible intervals. In our previous draft we computed the JS divergence between approaches over a 14-dimensional parameter space, so it is not necessarily a fair comparison to make. As can now be seen even more clearly in our updated results, our mass and distance posteriors are directly comparable in quality to those in the referenced paper. We have also re-run all the Bilby samplers using settings suggested from experts in the field. Even with optimal parameter settings, it is not guaranteed that Bayesian samplers will all agree with each other perfectly, nor to what precision an "ideal" agreement should be.

When we asked one of the lead authors of the bilby GWTC-1 paper (also a lead bilby developer) if the other bilby samplers were as trustworthy as the Dynesty implementation, their response was "*No, we do not trust the others! The off the shelf samplers do not work "out of the box" to the precision required and a lot of fine-tuning is needed.*"

In order to make comparisons between VItamin and the benchmark, I will assume for the moment that dynesty is more robust, given that dynesty has been adopted as the flagship sampler of bilby and is used in production LIGO/Virgo analyses. Then there is a clear problem with the VItamin results: while marginalized 1D posterior samples may look fine, some of these samples occur in pairs, triples, quadruples etc… which are statistically *extremely unlikely*. For instance, the 2D VItamin posteriors for (distance, time), (distance, ra), (distance, dec), (ra, dec) have tails in their 95% CI which have effectively zero probability in the dynesty PDF. This bias

does not show up in the PP test, because the PP test does not look for bias in 2D and higher-D PDFs.

Although the referee is correct in stating that such multi-dimensional biases will not necessarily show up in a traditional PP test, our JS divergence plot should be able to accurately reflect such biases. From the JS divergence plot, it can clearly be seen that we are generally consistent with other sampler approaches using hyperparameter settings provided to us by experts (and in some cases the sampler authors).

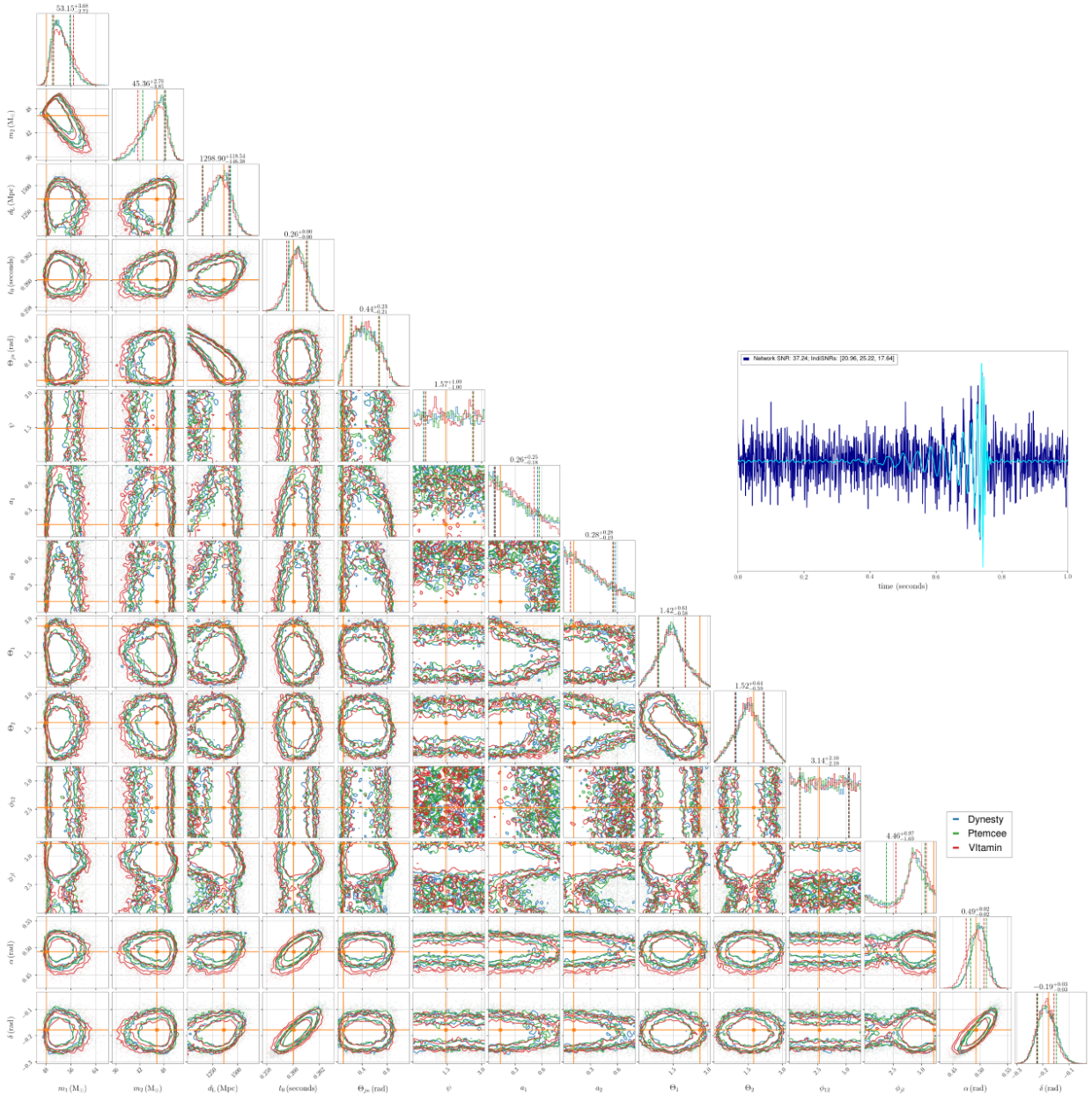
We do not understand what the referee is referring to in terms of samples occurring in pairs, triples, quadruples, etc. We see no evidence of this in Fig. 2 or in the data. We do accept that the joint posteriors in the previous draft did not match perfectly - our point was to show that we matched at the same level as other bilby samplers. We believe that our updated results provide a better representation of the improved quality of the Vitamin results.

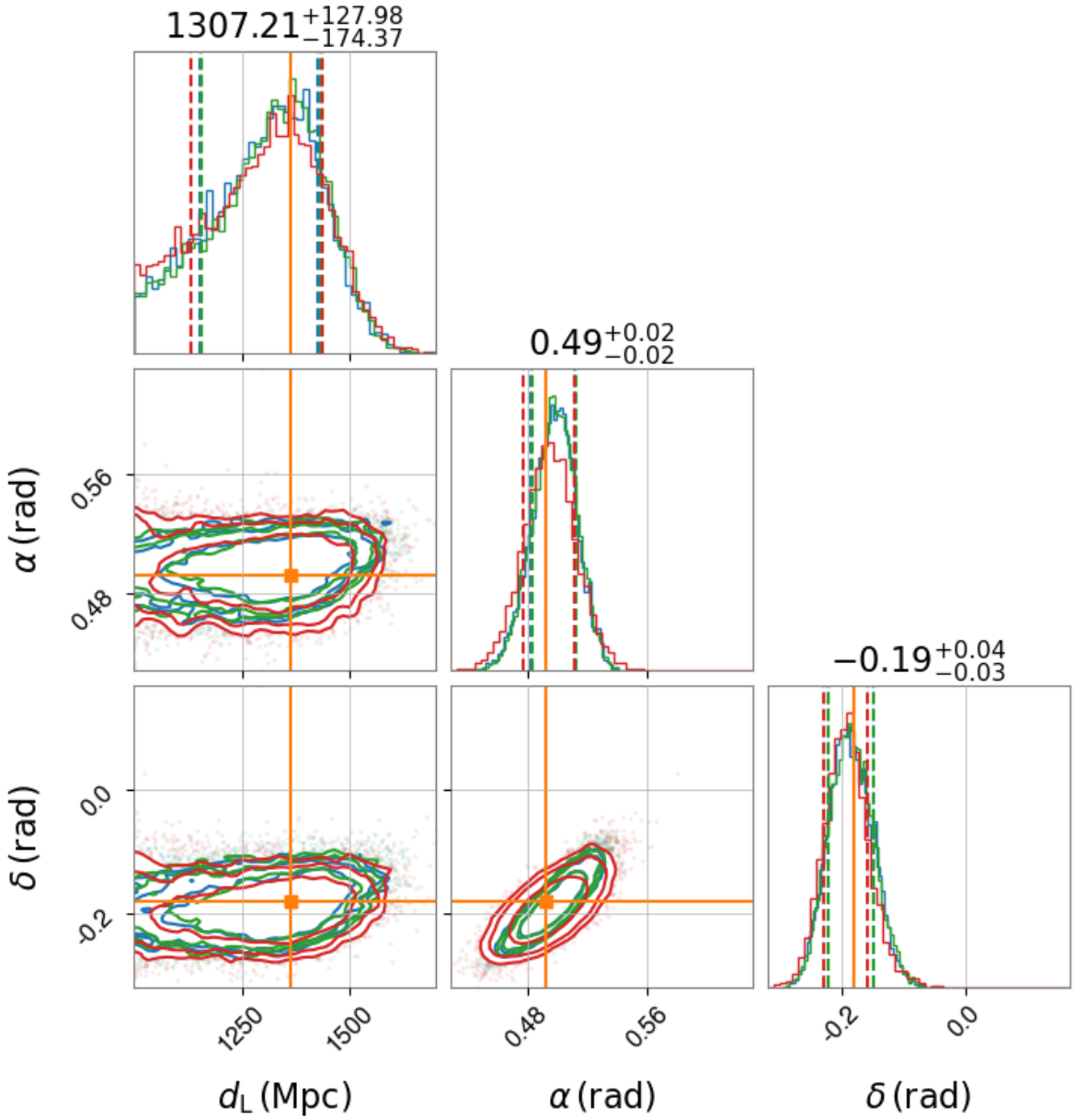
As an example of how 1D PDFs can hide bias in 2D PDFs I have attached a corner plot showing that with a software injection, the highlighted parameter's 1D PDF contains the true value in the 90% CI whereas in a certain 2D PDF it is falls on the 3-sigma credible interval boundary. Thus, while individual parameters can occur with the right probability, it is no guarantee that they occur in pairs with the correct probability.

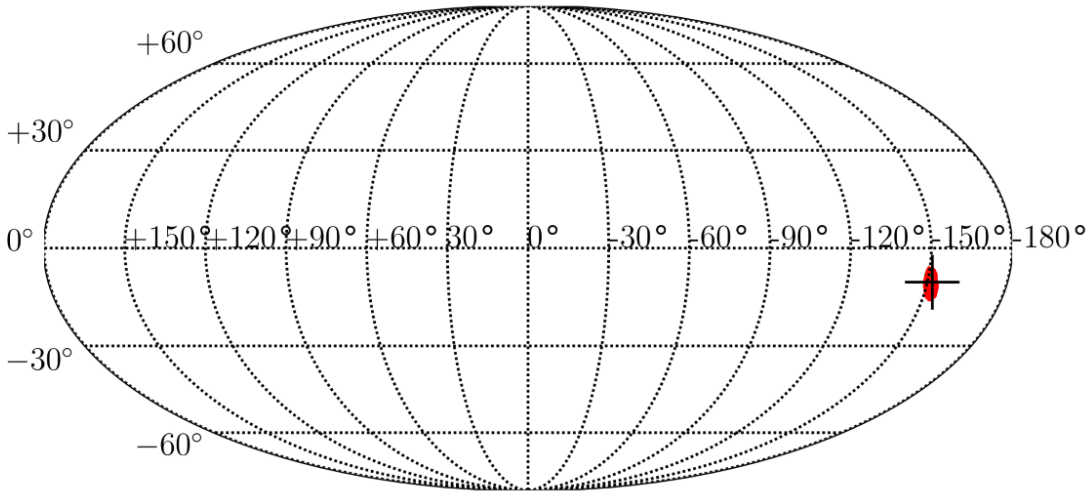
We agree with the referee that 1D PDFs can certainly hide bias in 2D PDFs, but we refer the referee to our updated posterior plots (Fig. 2) and updated JS divergence figure for an accurate accounting of said biases.

I am particularly concerned with the (distance, time), (distance, ra), (distance, dec), (ra, dec) PDFs as the authors suggest that their method could be used for rapid sky localisation on BNS mergers. It currently appears to me that VItamin results could conceivably assign high probability to sky locations which in fact have effectively zero probability. I am therefore curious if the distributions for the other software injections also display similar disparities in the 2D credible intervals. In particular, what do the 2D credible intervals look like for the highest SNR signal in the injection set (SNR ~ 75)?

We have attached an example of one of the high SNR events in our testing set (optimal network SNR~37). Our training set does extend to SNR~75 but due to the fact that the training set had 1e6 signals and the test set has 250 we do not have a test example that extends that far in to the high SNR tail. As can be seen in both the corner plot and sky map figures below, VItamin I is able to reproduce to a high degree of accuracy the posteriors for this event.


Network SNR: 37.24; IndiSNRs: [20.96, 25.22, 17.64]
time (seconds)
Dynesty
Ptemcee
Vitamin
m_1 ($M_\odot$)
m_2 ($M_\odot$)
d_L (Mpc)
t_0 (seconds)
Θ_{jn} (rad)
ψ
a_1
a_2
Θ_1
Θ_2
ϕ_{12}
ϕ_{jl}
α (rad)
δ (rad)


$1307.21^{+127.98}_{-174.37}$
$0.49^{+0.02}_{-0.02}$
$-0.19^{+0.04}_{-0.03}$
α (rad)
δ (rad)
0.56
0.48
0.0
-0.2
1250
1500
0.48
0.56
-0.2
0.0
d_L (Mpc)
α (rad)
δ (rad)



At the very least, I would like to see that the VItamin 2D credible intervals can (a) match the benchmark results, and (b) that benchmark results should be consistent in order to make a meaningful comparison. As things stand, I cannot claim that VItamin's full 7D posteriors are unbiased and I strongly dispute the author's claim that their results are indistinguishable from any of the benchmark results.

We refer the referee to our JS divergence histogram figure in order to get a sense for the overall performance of our ML approach within the context of other benchmark results. As can be seen in the updated JS figure, our results are entirely consistent with comparisons between different samplers. This is further illustrated in the corner plot example and the statistical consistency of the results is demonstrated in the updated PP plot.

The KL-statistic test also sweeps the issue of bias under the rug by saying that the KL-stat distribution between VItamin and benchmark results is similar to that of the benchmark results with themselves. I maintain that this should not be the case: the benchmark results should be converged so that the KL-statistic is close to zero.

We disagree with the referee on this issue. As has been argued earlier in this response, since only the Dynesty sampler implementation has been rigorously calibrated by PE experts there is no reason to expect that the JS divergence values between different samplers will approach zero. We also note that the 14-dimensional JS divergence distributions were estimated using an approximation technique (https://pypi.org/project/universal-divergence/) and a finite number of samples such that there was a fundamental noise of ~+/- 0.15 on the output values - hence even samples from 2 different sampler runs on the same test data would have JS divergence scatter of this magnitude around 0. We include the current version of this plot for the referee but have replaced it with the more trustworthy, accurate, and commonly used 1-D versions in the updated draft.

To use the bilby GWTC-1 paper as an example again, the authors use a Jensen-Shannon (JS) divergence statistic to measure the accuracy between bilby and LALInference. They find a maximum JS stat of 0.0019 across all parameters for the entire GWTC-1 catalogue. While the JS statistic is not identical to the KL-statistic, they both nevertheless measure the difference in bits between two distributions. The authors find KL-stats of order 1-10 between their benchmark results, again suggesting a lack of convergence of one or all benchmark samplers.

The paper the referee mentions is only computing the 1-D JS divergence, not a 14-D JS divergence, so these numbers are not directly comparable. The figures quoted are also only done comparing two sampling implementations (bilby and lalinference), both highly tuned for gravitational wave inference. This is not a comparison of different bilby samplers with each other, and is therefore not a fair benchmark to compare our results to. We would also point out that the correct maximum JS statistic (on single parameters of the 11 events considered) from the GWTC-1 paper is 0.0026 nats as quoted in Section 5 (Summary) of the paper - although we also note that in Table 1 a maximum JS statistic of 0.0064 is obtained for the BBH event GW170818. In this case the authors state that the kernel density estimation method used to calculate the statistic is at fault. We are using the exact same code to perform our estimates and should therefore also expect to see values at this level When comparing bilby's implementation of the Dynesty and PTemcee samplers across a far larger testing set (250) we see median JS divergences at the 0.002 nat level. We also see JS divergences between Dynesty and VItamin within a factor of ~few of these values. Most importantly, we feel that this level of accuracy, whilst not exactly matching existing samplers, is obtained in <1 second - 6 orders of magnitude faster than existing samplers.

As stated earlier, we have now replaced Fig. 5 with 1-D JS-divergence plots which provide a more robust and accurate comparison between Vitamin and other samplers with reference to the trusted Dynesty benchmark results.