Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference

Maximilian Dax,[1,*] Stephen R. Green,[2,3,†] Jonathan Gair,[2] Michael Pürrer,[2,4,5] Jonas Wildberger,[1]
Jakob H. Macke,[1,6] Alessandra Buonanno,[2,7] and Bernhard Schölkopf[1]
[1]*Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany*
[2]*Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Am Mühlenberg 1, 14476 Potsdam, Germany*
[3]*School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom*
[4]*Department of Physics, East Hall, University of Rhode Island, Kingston, Rhode Island 02881, USA*
[5]*URI Research Computing, Tyler Hall, University of Rhode Island, Kingston, Rhode Island 02881, USA*
[6]*Machine Learning in Science, University of Tübingen, 72076 Tübingen, Germany*
[7]*Department of Physics, University of Maryland, College Park, Maryland 20742, USA*

(Received 14 October 2022; revised 23 January 2023; accepted 21 February 2023; published 26 April 2023)

We combine amortized neural posterior estimation with importance sampling for fast and accurate gravitational-wave inference. We first generate a rapid proposal for the Bayesian posterior using neural networks, and then attach importance weights based on the underlying likelihood and prior. This provides (1) a corrected posterior free from network inaccuracies, (2) a performance diagnostic (the sample efficiency) for assessing the proposal and identifying failure cases, and (3) an unbiased estimate of the Bayesian evidence. By establishing this independent verification and correction mechanism we address some of the most frequent criticisms against deep learning for scientific inference. We carry out a large study analyzing 42 binary black hole mergers observed by LIGO and Virgo with the SEOBNRv4PHM and IMRPhenomXPHM waveform models. This shows a median sample efficiency of ≈10% (2 orders of magnitude better than standard samplers) as well as a tenfold reduction in the statistical uncertainty in the log evidence. Given these advantages, we expect a significant impact on gravitational-wave inference, and for this approach to serve as a paradigm for harnessing deep learning methods in scientific applications.

DOI: 10.1103/PhysRevLett.130.171403

Introduction.—Bayesian inference is a key paradigm for scientific discovery. In the context of gravitational waves (GWs), it underlies analyses including individual-event parameter estimation [1], tests of gravity [2], neutron-star physics [3], populations [4], and cosmology [5]. Given a prior $p(\theta)$ and a model likelihood $p(d|\theta)$, the Bayesian posterior

$$p(\theta|d) = \frac{p(d|\theta)p(\theta)}{p(d)} \tag{1}$$

summarizes, as a probability distribution, our knowledge of the model parameters θ after observing data d. When $p(d|\theta)$ is tractable (as in the case of GWs) likelihood-based samplers such as Markov chain Monte Carlo (MCMC) [6,7] or nested sampling [8] are typically used to draw samples from the posterior. If it is possible to *sample* $d \sim p(d|\theta)$ (i.e., simulate data) one can alternatively use amortized simulation-based (or likelihood-free) inference methods [9]. These approaches are based on deep neural networks and can be several orders-of-magnitude faster at inference time. For GW inference, they have also been shown to achieve similar accuracy to MCMC [10]. In general, however, it is not clear how well such networks generalize to out-of-distribution data and they lack diagnostics to be confident in results [11]. These powerful approaches are therefore rarely used in applications where accuracy is important and likelihoods are tractable.

In this Letter, we achieve the best of both worlds by combining likelihood-free and likelihood-based methods for GW parameter estimation. We take samples from DINGO [10,12]—a fast and accurate likelihood-free method using normalizing flows [13–16]—and treat these as a proposal for importance sampling [17]. The combined method ("DINGO-IS") generates samples from the exact posterior and now provides an estimate of the Bayesian evidence $p(d)$. Moreover, the importance sampling efficiency arises as a powerful and objective performance metric, which flags potential failure cases. Importance sampling is fully parallelizable.

After describing the method more fully in the following section, we verify on two real events that DINGO-IS produces results consistent with standard inference codes

[18–21]. Our main result is an analysis of 42 events from the Second and Third Gravitational-Wave Transient Catalogs (GWTC-2 and GWTC-3) [1,22], using two waveform models, IMRPhenomXPHM [23] and SEOBNRv4PHM [24]. Because of the long waveform simulation times, SEOBNRv4PHM inference would take several months per event with stochastic samplers. However, DINGO-IS with 64 CPU cores takes just 10 h for these waveforms. (Initial DINGO samples are available typically in under a minute.) Our results indicate that DINGO(-IS) performs well for the majority of events, and that failure cases are indeed flagged by low sample efficiency. We also find that the log evidence is recovered with statistical uncertainty reduced by a factor of 10 compared to standard samplers.

Machine learning methods have seen numerous applications in GW astronomy, including to detection and parameter estimation [25]. For parameter estimation, these methods have included variational inference [26,27], likelihood ratio estimation [28], and posterior estimation with normalizing flows [10,27,29,30]. Aside from directly estimating parameters, normalizing flows have also been used to accelerate classical samplers, with significant efficiency improvements [31].

Neural density estimation and importance sampling have previously been combined [32] under the guise of "neural importance sampling" [33], and similar approaches have been applied in several contexts [34–37]. Our contributions are to (1) extend this to amortized simulation-based inference, (2) use it to improve results generated with classical inference methods such as MCMC, and (3) to highlight how the use of a forward Kullback-Leibler (KL) loss improves reliability. We also apply it to the challenging real-world problem of GW inference [38]. We demonstrate results that far outperform classical methods in terms of sample efficiency and parallelizability, while maintaining accuracy and including simple diagnostics. We therefore expect this work to accelerate the development and verification of probabilistic deep learning approaches across science.

Method.—DINGO trains a conditional density-estimation neural network $q(\theta|d)$ to approximate $p(\theta|d)$ based on simulated datasets (θ, d) with $\theta \sim p(\theta)$, $d \sim p(d|\theta)$—an approach called neural posterior estimation (NPE) [40]. Once trained, DINGO can rapidly produce (approximate) posterior samples for any measured data d. In practice, results may deviate from the true posterior due to insufficient training, lack of network expressivity, or out-of-distribution (OOD) data (i.e., data inconsistent with the training distribution). Although it was shown in [10] that these deviations are often negligible, verification of results requires comparing against expensive standard samplers.

Here, we describe an efficient method to *verify* and *correct* DINGO results using importance sampling (IS) [17]. Starting from a collection of n samples $\theta_i \sim q(\theta|d)$ (the "proposal") we assign to each one an importance weight $w_i = p(d|\theta_i)p(\theta_i)/q(\theta_i|d)$. For a perfect proposal, $w_i = \text{constant}$, but more generally the number of *effective samples* is related to the variance, $n_{\text{eff}} = (\sum_i w_i)^2/\sum_i(w_i^2)$ [41]. The *sample efficiency* $\epsilon = n_{\text{eff}}/n \in (0, 1]$ arises naturally as a quality measure of the proposal.

Importance sampling requires evaluation of $p(d|\theta)p(\theta)$ rather than the normalized posterior. The Bayesian evidence can then be estimated from the normalization of the weights as $p(d) = 1/n \sum_i w_i$. The standard deviation of the log evidence, $\sigma_{\log p(d)} = \sqrt{(1-\epsilon)/(n\epsilon)}$ (see Supplemental Material [42]), scales with $1/\sqrt{n}$, enabling very precise estimates. The evidence is furthermore unbiased if the support of the posterior is fully covered by the proposal distribution [43]. The *log* evidence does have a bias, but this scales as $1/n$, and in all cases considered here is completely negligible (see Supplemental Material). If $q(\theta|d)$ fails to cover the entire posterior, the evidence itself would also be biased, toward lower values.

NPE is particularly well suited for IS because of two key properties. First, by construction the proposal has tractable density, such that we can not only sample from $q(\theta|d)$, but also evaluate it. Second, the NPE proposal is expected to always cover the entire posterior support. This is because, during training, NPE minimizes the *forward* KL divergence $D_{\text{KL}}(p(\theta|d)||q(\theta|d))$. This diverges unless $\text{supp}(p(\theta|d)) \subseteq \text{supp}(q(\theta|d))$, making the loss "probability-mass covering." Probability mass coverage is not guaranteed for finite sets of samples generated with stochastic samplers like MCMC (which can miss distributional modes), or machine learning methods with other training objectives like variational inference [13,44,45].

Neural importance sampling can in fact be used to improve posterior samples from *any* inference method provided the likelihood is tractable. If the method provides only samples (without density) then one must first train an (unconditional) density estimator $q(\theta)$ (e.g., a normalizing flow [13,14,46]) to use as the proposal. This is generally fast for an unconditional flow, and using the forward KL loss guarantees that the proposal will cover the samples. Success, however, relies on the quality of the initial samples: if they are light tailed, sample efficiency will be poor, and if they are not mass covering, the evidence will be biased. Nevertheless, for initial samples that well represent the posterior, this technique can provide quick verification and improvement.

In the context of GWs, we refer to neural importance sampling with DINGO as DINGO-IS. Although this technique requires likelihood evaluations at inference time, in practice it is much faster than other likelihood-based methods because of its high sample efficiency and parallelizability. Indeed, DINGO samples are independent and identically distributed, trivially enabling full parallelization of likelihood evaluations. This is a crucial advantage compared to inherently sequential methods such as MCMC.

Results.—For our experiments, we prepare DINGO networks as described in [10], with several modifications. First, we extend the priors over component masses to $m_1, m_2 \in [10, 120] M_\odot$ and dimensionless spin magnitudes to $a_1, a_2 \in [0, 0.99]$. We also use the waveform models IMRPhenomXPHM [23] and SEOBNRv4PHM [24], which include higher radiative multipoles and more realistic precession. Finally, in addition to networks for the first observing run of LIGO and Virgo (O1), we also train networks based on O3 noise. For the O3 analyses, we found performance improved by training separate DINGO models with distance priors [0.1, 3] Gpc, [0.1, 6] Gpc, and [0.1, 12] Gpc. We continue to use frequency-domain strain data in the range [20, 1024] Hz with $\Delta f = 0.125$ Hz and identical data conditioning as in [10]. The network architecture, hyperparameters, and training algorithm are also unchanged. We consider the two LIGO [47] detectors for all analyses, and leave inclusion of Virgo [48] data to a future publication of a complete catalog.

In our experiments, we found that DINGO often has difficulty resolving the phase parameter ϕ_c. Although ϕ_c itself is of little physical interest, it is nevertheless needed to evaluate the likelihood for importance sampling. We therefore sample ϕ_c synthetically, by first evaluating the likelihood across a ϕ_c grid and caching the waveform modes for efficiency (see Supplemental Material). This approach is similar to standard phase marginalization [18,49,50], but it is valid even with higher modes; it can therefore be adapted also to stochastic samplers.

For DINGO-IS, with 10^5 proposal samples per event, the total time for inference using one NVIDIA A100 GPU and 64 CPU cores is typically less than 1 h for IMRPhenomXPHM and ≈10 hours for SEOBNRv4PHM. In both cases, the computation time is dominated by waveform simulations, which could be further reduced using more CPUs. The rest of the time is taken up to generate the initial DINGO proposal samples [51].

TABLE I. Performance for GW150914 (upper block) and GW151012 (lower) with waveform model IMRPhenomXPHM. The Jensen-Shannon divergence (JSD) quantifies the deviation from LALINFERENCE-MCMC for one-dimensional marginal posteriors (all values in 10^{-3} nat). The mean is taken across all parameters. Posteriors with a maximum JSD $\leq 2 \times 10^{-3}$ nat are considered indistinguishable [20]; here, maxima occur for right ascension α, luminosity distance d_L, and chirp mass M_c. We also report BILBY-DYNESTY results.

	Mean JSD	Max JSD	$\log p(d)$
DINGO	2.2	7.2 (α)	
DINGO-IS	0.5	1.4 (d_L)	$-15\,831.87 \pm 0.01$
BILBY	1.8	4.0 (d_L)	$-15\,831.78 \pm 0.10$
DINGO	9.0	53.4 (M_c)	
DINGO-IS	0.7	2.2 (α)	$-16\,412.88 \pm 0.01$
BILBY	1.1	4.1 (α)	$-16\,412.73 \pm 0.09$

We first validate DINGO-IS against standard inference codes for two real events, GW150914 and GW151012, using IMRPhenomXPHM. (For SEOBNRv4PHM it is not feasible to run classical samplers, and one would instead need to use faster methods such as RIFT [53,54].) We generate reference posteriors using LALINFERENCE-MCMC [18], and compare one-dimensional marginalized posteriors for each parameter using the Jensen-Shannon

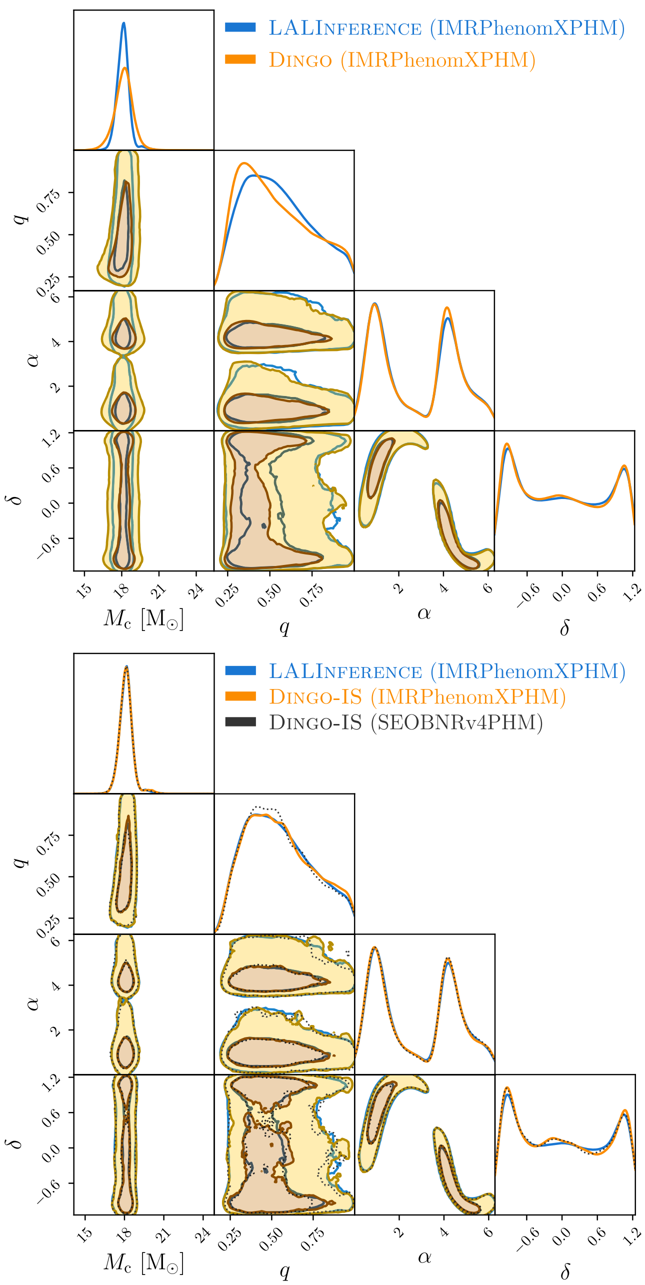

FIG. 1. Chirp mass (M_c), mass ratio (q) and sky position (α, δ) parameters for GW151012, comparing inference with DINGO and LALINFERENCE-MCMC. Even when initial DINGO results deviate from LALINFERENCE posteriors (upper panel), IS leads to almost perfect agreement (lower). For comparison, the lower panel also shows results for SEOBNRv4PHM.

divergence (Table I). For both events, the initial small deviations of DINGO samples from the reference are made negligible [55] using DINGO-IS (see Fig. 1 for a qualitative demonstration). We find sample efficiencies of $\epsilon = 28.8\%$ and $\epsilon = 12.5\%$ for GW150914 and GW151012, respectively.

For the evidence, we compare against BILBY-DYNESTY [19–21], since nested sampling generally provides a more accurate estimate than MCMC. In Table I we see that DINGO-IS is more precise by a factor of ≈ 10, but the BILBY evidence is larger for both events by roughly one standard deviation. This deviation could be statistical, but it could also indicate a bias in one of the methods. (Recall that IS requires the proposal to be mass covering for an unbiased evidence.) To further investigate for GW151012, we perform neural importance sampling starting from 10^6 BILBY samples (see Supplemental Material). This achieves a slightly lower $\epsilon = 8.3\%$ than DINGO-IS, but $\log p(d) = -16\,412.89 \pm 0.01$ in close agreement. While this does not fully rule out a bias in DINGO-IS samples (since the test is not fully independent) we take this as an indication that DINGO-IS indeed infers an unbiased evidence. More generally, it showcases how our method can be extended to improve the output of stochastic samplers.

We now perform a large study analyzing all 42 events in GWTC-2 [22] and GWTC-3 [1] that are consistent with our mass prior [56]. We stress that a study of this scope would be infeasible with standard codes, since SEOBNRv4PHM inference for a single event would take several months. Across all events we achieve a median sampling efficiency of $\epsilon = 10.9\%$ for IMRPhenomXPHM and $\epsilon = 4.4\%$ for SEOBNRv4PHM (Table II). For most events, the initial DINGO results are already accurate and only deviate slightly from DINGO-IS; furthermore, DINGO-IS shows excellent agreement between the two waveform models (see the Supplemental Material for more detailed comparisons). Note that these results are based on highly complex precessing higher-mode waveform models, and do not include any mitigation of noise transients (see below). With the simpler IMRPhenomPv2 [59–61] model and a smaller mass prior (in a study on drifting detector noise distributions [62]) DINGO-IS achieves an even larger median sample efficiency of $\epsilon = 36.8\%$ on 37 events.

Importance sampling guarantees robust results by marking failure cases with a low sample efficiency. By this metric, DINGO struggles slightly with chirp masses near the lower prior boundary (GW191204_110529 and GW200322_091133). For such systems, efficiency may be improved by increasing the prior range used for training. Events with known data quality issues also often have low sample efficiency (see Table II): several low-ϵ events are contaminated by glitch artifacts (which would be mitigated in a more complete analysis [1,22]); GW200129_065458, in addition to having a glitch [63], may not be well modeled by either of our waveform models due to having strong

TABLE II. 42 BBH events from GWTC-3 analyzed with DINGO-IS. We report the log evidence $\log p(d)$ and the sample efficiency ϵ for the two waveform models IMRPhenomXPHM (upper rows) and SEOBNRv4PHM (lower rows). Highlighting colors indicate the sample efficiency (green: high; yellow: medium; orange and red: low); DINGO-IS results can be trusted for medium and high ϵ (see Supplemental Material). Events in gray suffer from data quality issues [1,22].

Event	$\log p(d)$	ϵ (%)
GW190408_181802	$-16\,178.332 \pm 0.012$	6.9
	$-16\,178.172 \pm 0.010$	9.3
GW190413_052954	$-15\,571.413 \pm 0.006$	22.5
	$-15\,571.391 \pm 0.005$	26.3
GW190413_134308	$-16\,399.331 \pm 0.009$	12.4
	$-16\,399.139 \pm 0.014$	4.7
GW190421_213856	$-15\,983.248 \pm 0.008$	15.3
	$-15\,983.131 \pm 0.010$	9.4
GW190503_185404	$-16\,582.865 \pm 0.022$	2.0
	$-16\,583.352 \pm 0.027$	1.4
GW190513_205428	$-15\,946.462 \pm 0.043$	0.6
	$-15\,946.581 \pm 0.017$	3.4
GW190514_065416	$-16\,556.466 \pm 0.009$	11.6
	$-16\,556.314 \pm 0.017$	3.5
GW190517_055101	$-16\,271.048 \pm 0.027$	1.3
	$-16\,272.428 \pm 0.034$	0.9
GW190519_153544	$-15\,991.171 \pm 0.008$	15.2
	$-15\,991.287 \pm 0.068$	0.2
GW190521_074359	$-16\,008.876 \pm 0.008$	13.4
	$-16\,008.037 \pm 0.015$	4.2
GW190527_092055	$-16\,119.012 \pm 0.008$	13.8
	$-16\,118.781 \pm 0.013$	6.1
GW190602_175927	$-16\,036.993 \pm 0.006$	25.0
	$-16\,037.529 \pm 0.006$	23.5
GW190701_203306	$-16\,521.381 \pm 0.040$	0.6
	$-16\,521.609 \pm 0.010$	10.1
GW190719_215514	$-15\,850.492 \pm 0.008$	13.4
	$-15\,850.339 \pm 0.011$	8.0
GW190727_060333	$-15\,992.017 \pm 0.009$	10.3
	$-15\,992.428 \pm 0.005$	30.8
GW190731_140936	$-16\,376.777 \pm 0.005$	32.6
	$-16\,376.763 \pm 0.005$	31.0
GW190803_022701	$-16\,132.409 \pm 0.006$	21.4
	$-16\,132.408 \pm 0.005$	27.8
GW190805_211137	$-16\,073.261 \pm 0.006$	20.0
	$-16\,073.656 \pm 0.007$	16.6
GW190828_063405	$-16\,137.220 \pm 0.009$	12.2
	$-16\,136.799 \pm 0.010$	9.1
GW190909_114149	$-16\,061.634 \pm 0.011$	7.4
	$-16\,061.275 \pm 0.016$	3.8

(Table continued)

TABLE II. *(Continued)*

Event	$\log p(d)$	ϵ (%)
GW190915_235702	$-16\,083.960 \pm 0.015$	20.8
	$-16\,083.937 \pm 0.027$	4.8
GW190926_050336	$-16\,015.813 \pm 0.019$	2.8
	$-16\,015.861 \pm 0.009$	12.1
GW190929_012149	$-16\,146.666 \pm 0.018$	3.2
	$-16\,146.591 \pm 0.021$	2.4
GW191109_010717	$-17\,925.064 \pm 0.025$	1.7
	$-17\,922.762 \pm 0.041$	0.6
GW191127_050227	$-16\,759.328 \pm 0.019$	2.7
	$-16\,758.102 \pm 0.029$	1.2
[a]GW191204_110529	$-15\,984.455 \pm 0.015$	4.2
	$-15\,983.618 \pm 0.063$	0.3
GW191215_223052	$-16\,001.286 \pm 0.013$	5.8
	$-16\,000.846 \pm 0.052$	0.4
GW191222_033537	$-15\,871.521 \pm 0.007$	16.5
	$-15\,871.450 \pm 0.005$	25.8
GW191230_180458	$-15\,913.798 \pm 0.009$	12.2
	$-15\,913.918 \pm 0.010$	8.8
GW200128_022011	$-16\,305.128 \pm 0.013$	6.1
	$-16\,304.510 \pm 0.007$	18.3
[a]GW200129_065458	$-16\,226.851 \pm 0.109$	0.1
	$-16\,231.203 \pm 0.051$	0.4
GW200208_130117	$-16\,136.381 \pm 0.007$	16.6
	$-16\,136.531 \pm 0.009$	11.2
GW200208_222617	$-16\,775.200 \pm 0.011$	7.4
	$-16\,774.582 \pm 0.021$	2.2
GW200209_085452	$-16\,383.847 \pm 0.009$	12.5
	$-16\,384.157 \pm 0.025$	1.6
GW200216_220804	$-16\,215.703 \pm 0.017$	3.4
	$-16\,215.540 \pm 0.018$	3.1
GW200219_094415	$-16\,133.457 \pm 0.011$	9.6
	$-16\,133.157 \pm 0.017$	4.0
GW200220_061928	$-16\,303.782 \pm 0.007$	17.3
	$-16\,303.087 \pm 0.026$	1.5
GW200220_124850	$-16\,136.600 \pm 0.008$	13.2
	$-16\,136.519 \pm 0.037$	0.7
GW200224_222234	$-16\,138.613 \pm 0.006$	22.5
	$-16\,139.101 \pm 0.006$	21.4
[a]GW200308_173609	$-16\,173.938 \pm 0.013$	6.0
	$-16\,173.692 \pm 0.025$	1.7
GW200311_115853	$-16\,117.505 \pm 0.011$	7.4
	$-16\,117.583 \pm 0.009$	11.9
[a]GW200322_091133	-16313.568 ± 0.307	0.0
	$-16\,313.110 \pm 0.105$	0.1

[a]See remarks on these events in text.

precession [64]; and GW200322_091133 may be simply a Gaussian noise fluctuation [65]. In these cases, DINGO-IS marks events for additional investigation.

Data quality issues such as non-Gaussian noise or observed signals that do not match models correspond to OOD data, i.e., data not consistent with the training distribution. Since OOD data are not seen during training, DINGO cannot be expected to return their true posterior, which results in a low sample efficiency. As an additional test, running DINGO-IS on signal-free data with a blip glitch [66] in the LIGO Hanford detector (GPS time 1 238 613 687.5) results in $\epsilon \approx 0.001\%$. Likewise, we find that DINGO-IS successfully flags adversarial examples [67,68] that are intentionally corrupted to mislead the inference network ($\epsilon \approx 0.01\%$; see Supplemental Material)—addressing a common failure mode of neural networks. Our general view, therefore, is that although there can be various reasons for low-ϵ results, it often serves as a useful heuristic to identify OOD events.

Conclusions.—We have described the use of importance sampling to improve the results of NPE in amortized inference problems, and we applied it to the case of GWs. Neural importance sampling provides rapid verification of results and corrects any inaccuracies in deep learning output; it provides an evidence estimate with precision far exceeding that of classical samplers; and it marks potentially OOD data for further investigation. With high sample efficiency and rapid initial results, DINGO-IS becomes a comprehensive inference tool for accurately analyzing the large numbers of binary black hole (BBH) events expected soon.

High sample efficiencies are predicated on a high quality proposal, which DINGO thankfully provides. A key element is the probability-mass covering property, which is guaranteed by the forward KL training loss. This tends to produce broad tails, which are downweighted in importance sampling. *Overly* broad proposals would nevertheless result in low sample efficiency, so highly expressive density estimators such as normalizing flows are essential, along with DINGO innovations such as group-equivariant NPE (GNPE) [10,52] and GW training data augmentation. DINGO posteriors are rarely light tailed, but this does occasionally lead to underestimated evidence for small n.

With the inclusion of importance sampling, the DINGO pipeline can now be used in several different ways. When low latency is desired, complete posteriors are still available without importance sampling in a matter of seconds. Results include sky position and mass parameters and could therefore play an important role in directing electromagnetic followup observations once we extend DINGO to mergers involving neutron stars (see Ref. [56]). By comparing against DINGO-IS, we have shown that in the majority of cases, initial results are already very reliable, with only minor deviations in marginal distributions.

Indeed, validation of DINGO results was a major motivation in exploring importance sampling.

When high accuracy is desired, DINGO-IS reweights results to the true posterior and includes an estimate of the evidence. Results are verified and include probability mass-covering guarantees that ensure secondary modes are not missed. Sample efficiencies are often 2 orders of magnitude higher than MCMC or nested sampling, and importance sampling is fully parallelizable. As a consequence, results are typically available within an hour for IMRPhenomXPHM, or ten hours for SEOBNRv4PHM. This represents a significant advantage when considering the event rates likely to be reached with advanced detectors (three per week or higher in the upcoming LIGO-Virgo-KAGRA observing run O4).

DINGO-IS opens several new possibilities for GW analysis: (1) rapid inference means that the most accurate waveform models, which include all physical effects, could be used for all events; (2) high-precision evidences enable detailed model comparison; and (3) low sample efficiencies can identify data that do not fit the noise or waveform model. We believe that these results have highlighted clear benefits of combining likelihood-free and likelihood-based methods in Bayesian inference. Going forward, as DINGO-IS validates and builds trust in DINGO, it will help to set the stage for noise-model-free inference, which is truly likelihood-free.

The code for DINGO and DINGO-IS is available at [70].

We thank V. Raymond for encouraging us to pursue importance sampling in the early stages of the project, and C. García Quirós, N. Gupte, S. Ossokine, A. Ramos-Buades, and R. Smith for useful discussions. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. This research has made use of data or software obtained from the Gravitational Wave Open Science Center [69], a service of LIGO Laboratory, the LIGO Scientific Collaboration, the Virgo Collaboration, and KAGRA. LIGO Laboratory and Advanced LIGO are funded by the United States National Science Foundation (NSF) as well as the Science and Technology Facilities Council (STFC) of the United Kingdom, the Max-Planck-Society (MPS), and the State of Niedersachsen/Germany for support of the construction of Advanced LIGO and construction and operation of the GEO600 detector. Additional support for Advanced LIGO was provided by the Australian Research Council. Virgo is funded, through the European Gravitational Observatory (EGO), by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale di Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by institutions from Belgium, Germany, Greece, Hungary, Ireland, Japan, Monaco, Poland, Portugal, Spain. The construction and operation of KAGRA are funded by Ministry of Education, Culture, Sports, Science and Technology (MEXT), and Japan Society for the Promotion of Science (JSPS), National Research Foundation (NRF) and Ministry of Science and ICT (MSIT) in Korea, Academia Sinica (AS) and the Ministry of Science and Technology (MoST) in Taiwan. M. D. thanks the Hector Fellow Academy for support. J. H. M. and B. S. are members of the MLCoE, EXC number 2064/1—Project No. 390727645 and the Tübingen AI Center funded by the German Ministry for Science and Education (FKZ 01IS18039A). For the implementation of DINGO we use PyTorch [71], NFLOWS [72], LALSimulation [73], and the adam optimizer [74]. The plots are generated with MATPLOTLIB [75] and ChainConsumer [76].

M. D. and S. G. contributed equally to this work.

[*]maximilian.dax@tuebingen.mpg.de
[†]stephen.green2@nottingham.ac.uk

[1] R. Abbott *et al.* (LIGO Scientific, VIRGO, and KAGRA Collaborations), GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run, arXiv:2111.03606.

[2] R. Abbott *et al.* (LIGO Scientific, VIRGO, and KAGRA Collaborations), Tests of general relativity with GWTC-3, arXiv:2112.06861.

[3] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GW170817: Measurements of Neutron Star Radii and Equation of State, Phys. Rev. Lett. **121**, 161101 (2018).

[4] R. Abbott *et al.* (LIGO Scientific, VIRGO, and KAGRA Collaborations), The population of merging compact binaries inferred using gravitational waves through GWTC-3, arXiv:2111.03634.

[5] R. Abbott *et al.* (LIGO Scientific, VIRGO, and KAGRA Collaborations), Constraints on the cosmic expansion history from GWTC-3, arXiv:2111.03604.

[6] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. **21**, 1087 (1953).

[7] W. K. Hastings, Monte Carlo sampling methods using Markov Chains and their applications, Biometrika **57**, 97 (1970).

[8] J. Skilling, Nested sampling for general Bayesian computation, Bayesian Anal. **1**, 833 (2006).

[9] K. Cranmer, J. Brehmer, and G. Louppe, The frontier of simulation-based inference, Proc. Natl. Acad. Sci. U.S.A. **117**, 30055 (2020).

[10] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, Real-Time Gravitational Wave Science with Neural Posterior Estimation, Phys. Rev. Lett. **127**, 241103 (2021).

[11] P. Cannon, D. Ward, and S. M. Schmon, Investigating the impact of model misspecification in neural simulation-based inference, arXiv:2209.01845.

[12] Deep INference for Gravitational-wave Observations.

[13] D. Rezende and S. Mohamed, Variational inference with normalizing flows, in *Proceedings of the International*

Conference on Machine Learning (2015), pp. 1530–1538, arXiv:1505.05770.

[14] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, Improved variational inference with inverse autoregressive flow, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., New York, 2016), pp. 4743–4751.

[15] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., New York, 2019), pp. 7509–7520.

[16] G. Papamakarios, E. T. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, Normalizing flows for probabilistic modeling and inference., J. Mach. Learn. Res. **22**, 1 (2021).

[17] S. T. Tokdar and R. E. Kass, Importance sampling: A review, Wiley Interdiscip. Rev. **2**, 54 (2010).

[18] J. Veitch, V. Raymond, B. Farr, W. Farr, P. Graff, S. Vitale *et al.*, Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALINFERENCE software library, Phys. Rev. D **91**, 042003 (2015).

[19] G. Ashton *et al.*, BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy, Astrophys. J. Suppl. Ser. **241**, 27 (2019).

[20] I. M. Romero-Shaw *et al.*, Bayesian inference for compact binary coalescences with BILBY: Validation and application to the first LIGO–Virgo gravitational-wave transient catalogue, Mon. Not. R. Astron. Soc. **499**, 3295 (2020).

[21] J. S. Speagle, DYNESTY: A dynamic nested sampling package for estimating Bayesian posteriors and evidences, Mon. Not. R. Astron. Soc. **493**, 3132 (2020).

[22] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run, Phys. Rev. X **11**, 021053 (2021).

[23] G. Pratten *et al.*, Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes, Phys. Rev. D **103**, 104056 (2021).

[24] S. Ossokine *et al.*, Multipolar effective-one-body waveforms for precessing binary black holes: Construction and validation, Phys. Rev. D **102**, 044055 (2020).

[25] 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, Mach. Learn. **2**, 011002 (2020).

[26] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy, Nat. Phys. **18**, 112 (2022).

[27] S. R. Green, C. Simpson, and J. Gair, Gravitational-wave parameter estimation with autoregressive neural network flows, Phys. Rev. D **102**, 104057 (2020).

[28] A. Delaunoy, A. Wehenkel, T. Hinderer, S. Nissanke, C. Weniger, A. R. Williamson, and G. Louppe, Lightning-fast gravitational wave parameter inference through neural amortization, in *Proceedings of the Third Workshop on Machine Learning and the Physical Sciences* (2020), arXiv:2010.12931.

[29] S. R. Green and J. Gair, Complete parameter inference for GW150914 using deep learning, Mach. Learn. Sci. Tech. **2**, 03LT01 (2021).

[30] C. Chatterjee, L. Wen, D. Beveridge, F. Diakogiannis, and K. Vinsen, Rapid localization of gravitational wave sources from compact binary coalescences using deep learning, arXiv:2207.14522.

[31] M. J. Williams, J. Veitch, and C. Messenger, Nested sampling with normalizing flows for gravitational-wave inference, Phys. Rev. D **103**, 103006 (2021).

[32] B. Paige and F. Wood, Inference networks for sequential *Monte Carlo* in graphical models, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2016), pp. 3040–3049, arXiv:1602.06701.

[33] T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák, Neural importance sampling, ACM Trans. Graph. **38**, 1 (2019).

[34] F. Noé, S. Olsson, J. Köhler, and H. Wu, Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning, Science **365**, eaaw1147 (2019).

[35] M. S. Albergo, G. Kanwar, and P. E. Shanahan, Flow-based generative models for Markov Chain Monte Carlo in lattice field theory, Phys. Rev. D **100**, 034515 (2019).

[36] G. Kanwar, M. S. Albergo, D. Boyda, K. Cranmer, D. C. Hackett, S. Racanière, D. J. Rezende, and P. E. Shanahan, Equivariant Flow-Based Sampling for Lattice Gauge Theory, Phys. Rev. Lett. **125**, 121601 (2020).

[37] H. Sun, K. L. Bouman, P. Tiede, J. J. Wang, S. Blunt, and D. Mawet, α-deep probabilistic inference (α-DPI): Efficient uncertainty quantification from exoplanet astrometry to black hole feature extraction, Astrophys. J. **932**, 99 (2022).

[38] A similar approach using convolutional networks to parametrize Gaussian and von Mises proposals was used to estimate the sky position alone [39]. Using the normalizing flow proposal (as we do here) significantly improves the flexibility of the conditional density estimator and enables inference of all parameters.

[39] A. Kolmus, G. Baltus, J. Janquart, T. van Laarhoven, S. Caudill, and T. Heskes, Fast sky localization of gravitational waves using deep learning seeded importance sampling, Phys. Rev. D **106**, 023032 (2022).

[40] G. Papamakarios and I. Murray, Fast ϵ-free inference of simulation models with Bayesian conditional density estimation, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., New York, 2016).

[41] A. Kong, A note on importance sampling using standardized weights, University of Chicago, Department of Statistics, Technical Report, 1992, Vol. **348**.

[42] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.130.171403 for additional figures, derivations and technical details of our algorithms and neural networks.

[43] A. B. Owen, Monte Carlo theory, methods and examples (2013).

[44] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, An introduction to variational methods for graphical models, Mach. Learn. **37**, 183 (1999).

[45] M. J. Wainwright, M. I. Jordan *et al.*, Graphical models, exponential families, and variational inference, Found. Trends® Mach. Learn. **1**, 1 (2008).

[46] G. Papamakarios, T. Pavlakou, and I. Murray, Masked autoregressive flow for density estimation, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., New York, 2017), pp. 2338–2347.

[47] J. Aasi *et al.* (LIGO Scientific Collaboration), Advanced LIGO, Classical Quantum Gravity **32**, 074001 (2015).

[48] F. Acernese *et al.* (VIRGO Collaboration), Advanced Virgo: A second-generation interferometric gravitational wave detector, Classical Quantum Gravity **32**, 024001 (2015).

[49] J. Veitch and W. Del Pozzo, Analytic marginalisation of phase parameter, https://dcc.ligo.org/LIGO-T1300326/public (2013).

[50] E. Thrane and C. Talbot, An introduction to Bayesian inference in gravitational-wave astronomy: parameter estimation, model selection, and hierarchical models, Pub. Astron. Soc. Aust. **36**, e010 (2019); **37**, e036(E) (2020).

[51] It takes longer to generate the proposal than to produce low-latency DINGO samples (≈20 s) because of the group-equivariant NPE (GNPE) algorithm [10,52] (which breaks access to the density) and the synthetic phase recovery. See Supplemental Material for details.

[52] M. Dax, S. R. Green, J. Gair, M. Deistler, B. Schölkopf, and J. H. Macke, Group equivariant neural posterior estimation, in *Proceedings of the International Conference on Learning Representations* (2022), arXiv:2111.13139.

[53] C. Pankow, P. Brady, E. Ochsner, and R. O'Shaughnessy, Novel scheme for rapid parallel parameter estimation of gravitational waves from compact binary coalescences, Phys. Rev. D **92**, 023002 (2015).

[54] J. Lange, R. O'Shaughnessy, and M. Rizzo, Rapid and accurate parameter inference for coalescing, precessing compact binaries, arXiv:1805.10457.

[55] Initial deviations are larger than those reported in [10] since we use a more complicated waveform model and a larger prior, while keeping the size of the neural network and training time the same. Any remaining deviations after importance sampling can in principle also be due to sampling inaccuracies of LALINFERENCE MCMC. Note that a direct comparison to published LIGO-Virgo-KAGRA results is impeded by different data settings.

[56] Lower mass events produce longer signals, so extending DINGO to these may require improved methods for data compression [57,58]. This will be particularly relevant for binary neutron stars.

[57] S. Vinciguerra, J. Veitch, and I. Mandel, Accelerating gravitational wave parameter estimation with multi-band template interpolation, Classical Quantum Gravity **34**, 115006 (2017).

[58] K. Cannon *et al.*, Toward early-warning detection of gravitational waves from compact binary coalescence, Astrophys. J. **748**, 136 (2012).

[59] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms, Phys. Rev. Lett. **113**, 151101 (2014).

[60] 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 **93**, 044007 (2016).

[61] A. Bohé, M. Hannam, S. Husa, F. Ohme, M. Pürrer, and P. Schmidt, PhenomPv2—technical notes for the LAL implementation, LIGO Technical Document No. LIGO-T1500602-v4, 2016.

[62] J. Wildberger, M. Dax, S. R. Green, J. Gair, M. Pürrer, J. H. Macke, A. Buonanno, and B. Schölkopf, companion paper, Adapting to noise distribution shifts in flow-based gravitational-wave inference, Phys. Rev. D **107**, 084046 (2023).

[63] E. Payne, S. Hourihane, J. Golomb, R. Udall, D. Davis, and K. Chatziioannou, Curious case of GW200129: Interplay between spin-precession inference and data-quality issues, Phys. Rev. D **106**, 104017 (2022).

[64] M. Hannam *et al.*, General-relativistic precession in a black-hole binary, Nature (London) **610**, 652 (2022).

[65] G. Morras, J. F. N. n. Siles, J. Garcia-Bellido, and E. R. Morales, The false alarms induced by Gaussian noise in gravitational wave detectors, Phys. Rev. D **107**, 023027 (2023).

[66] S. Coughlin, M. Zevin, S. Bahaadini, N. Rohani, S. Allen, C. Berry, K. Crowston, M. Harandi, C. Jackson, V. Kalogera, A. Katsaggelos, V. Noroozi, C. Osterlund, O. Patane, J. Smith, S. Soni, and L. Trouille, Gravity spy machine learning classifications of LIGO glitches from observing runs O1, O2, O3a, and O3b, 10.5281/zenodo.5649212 (2021).

[67] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks, in *Proceedings of the International Conference on Learning Representations* (2014), arXiv:1312.6199.

[68] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, in *Proceedings of the International Conference on Learning Representations* (2015), arXiv:1412.6572.

[69] http://gw-openscience.org.

[70] https://github.com/dingo-gw/dingo.

[71] A. Paszke *et al.*, PyTorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019), pp. 8024–8035.

[72] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, NFLOWS: normalizing flows in PyTorch (2020), 10.5281/zenodo.4296287.

[73] LIGO Scientific Collaboration, LIGO Algorithm Library—LALSuite, free software (GPL) (2018), 10.7935/GT1W-FZ16.

[74] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *Proceedings of the International Conference on Learning Representations* (2015), arXiv:1412.6980.

[75] J. D. Hunter, MATPLOTLIB: A 2D graphics environment, Comput. Sci. Eng. **9**, 90 (2007).

[76] S. R. Hinton, ChainConsumer, J. Open Source Softwaare **1**, 45 (2016).