Neural Importance Sampling for Rapid and Reliable Gravitational-Wave Inference

Maximilian Dax,[1,*] Stephen R. Green,[2,†] Jonathan Gair,[2] Michael Pürrer,[2,3,4]
Jonas Wildberger,[1] Jakob H. Macke,[1,5] Alessandra Buonanno,[2,6] 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] *Department of Physics, East Hall, University of Rhode Island, Kingston, RI 02881, USA*
[4] *URI Research Computing, Tyler Hall, University of Rhode Island, Kingston, RI 02881, USA*
[5] *Machine Learning in Science, University of Tübingen, 72076 Tübingen, Germany*
[6] *Department of Physics, University of Maryland, College Park, MD 20742, USA*

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 $\approx 10\%$ (two orders-of-magnitude better than standard samplers) as well as a ten-fold 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.

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

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

summarises, 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[1] [10]—a fast and accurate likelihood-free method using normalizing flows [12–15]—and treat these as a proposal for importance sampling [16]. 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 [17–20]. Our main result is an analysis of 42 events from the Second and Third Gravitational-Wave Transient Catalogs (GWTC-2 and GWTC-3) [1, 21], using two waveform models, IMRPhenomXPHM [22] and SEOBNRv4PHM [23]. Due to 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 hours 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

[1] Deep INference for Gravitational-wave Observations.

parameter estimation [24]. For parameter estimation, these methods have included variational inference [25, 26], likelihood ratio estimation [27], and posterior estimation with normalizing flows [10, 26, 28, 29]. Aside from directly estimating parameters, normalizing flows have also been used to accelerate classical samplers, with significant efficiency improvements [30].

Neural density estimation and importance sampling have previously been combined under the guise of "neural importance sampling" [31], and this has been applied to several inference problems [32]. Our contributions are to (1) extend this to the case of conditional density estimators to achieve semi-amortized results, (2) use it to improve results of 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.[2] 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 data sets (θ, d) with $\theta \sim p(\theta)$, $d \sim p(d|\theta)$—an approach called neural posterior estimation (NPE) [34]. 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) [16]. 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 = constant, but more generally the number of *effective samples* is related to the variance, $n_{\text{eff}} = (\sum_i w_i)^2/\sum_i (w_i^2)$ [35]. 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, $\sigma_{\log p(d)} = \sqrt{(1-\epsilon)/(n \cdot \epsilon)}$ (see Supplemental Material), 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 [36]. 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 [12, 37, 38].

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 [12, 13, 39]) to use as 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 [22] and SEOBNRv4PHM [23], which include higher radiative multipoles and more realistic precession. Finally, in addition to O1 networks, we also train

[2] A similar approach using convolutional networks to parametrize Gaussian and von Mises proposals was used to estimate the sky position alone [33] Using the normalizing flow proposal (as we do here) significantly improves the flexiblity of the conditional density estimator and enables inference of all parameters.

	Mean JSD	Max JSD	$\log p(d)$
DINGO	2.2	7.2 (α)	-
DINGO-IS	0.5	1.4 (d_L)	-15831.87 ± 0.01
BILBY	1.8	4.0 (d_L)	-15831.78 ± 0.10
DINGO	9.0	53.4 (M_c)	-
DINGO-IS	0.7	2.2 (α)	-16412.88 ± 0.01
BILBY	1.1	4.1 (α)	-16412.73 ± 0.09

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 [19]; here, maxima occur for right ascension α, luminosity distance d_L, and chirp mass M_c. We also report BILBY-DYNESTY results.

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 [40] detectors for all analyses, and leave inclusion of Virgo [41] data to a future publication of a complete catalog. We sample the phase parameter ϕ_c analytically, which is similar to phase marginalization [17, 42, 43], but requires no approximations (see Supplemental Material).

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 hour 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.[3]

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 [45, 46].) We generate reference posteriors using LALINFERENCE-MCMC [17], and compare one-dimensional marginalized posteriors for each parameter using the Jensen-Shannon divergence (Tab. I). For both events, the initial small deviations[4] of DINGO samples from the reference are made

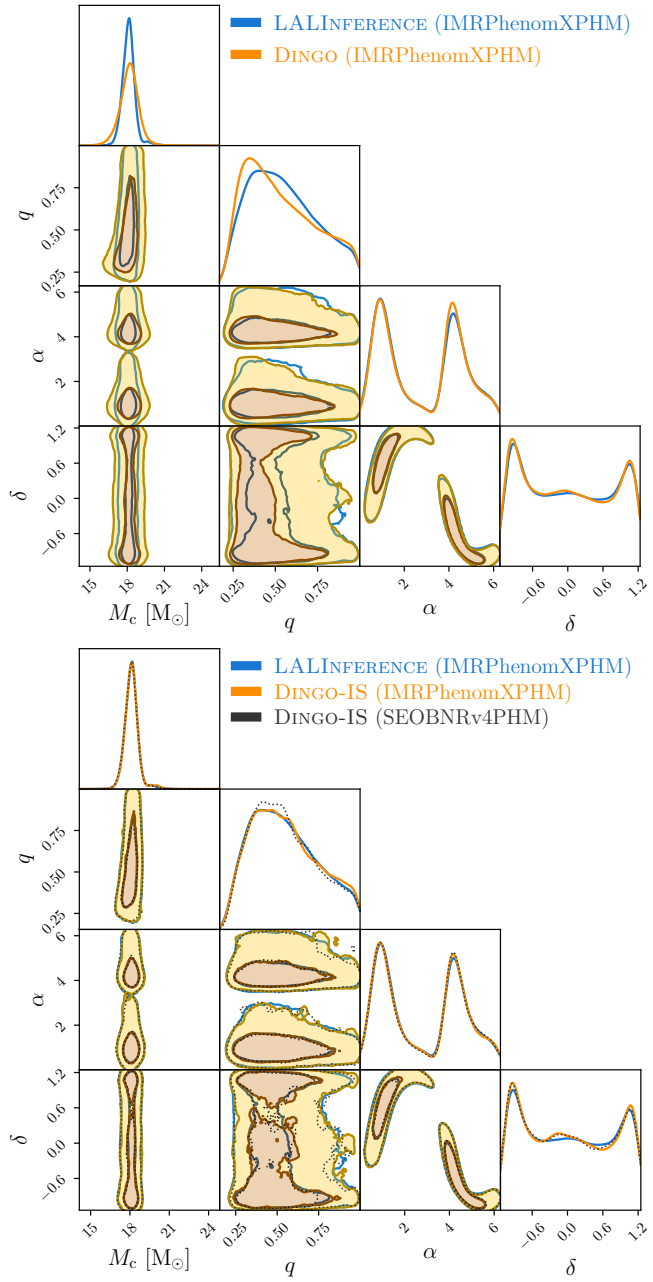

Figure 1. Chirp mass (M_c), mass ratio (q) and sky position (α, δ) parameters for GW151012, comparing inference with DINGO and LALINFERENCE-MCMC. Even when initial DINGO results deviate from LALINFERENCE posteriors (upper panel), IS leads to almost perfect agreement (lower). For comparison, the lower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

[3] It takes longer to generate the proposal than to produce low-latency DINGO samples (≈ 20 s) for several technical reasons (the use of the group-equivariant NPE (GNPE) [10, 44] algorithm that breaks access to the density, and a synthetic phase reconstruction). See Supplemental Material for details.

[4] 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.

Event	$\log p(d)$	ϵ	Event	$\log p(d)$	ϵ	Event	$\log p(d)$	ϵ
GW190408	-16178.332 ± 0.012	6.9%	GW190727	-15992.017 ± 0.009	10.3%	GW191230	-15913.798 ± 0.009	12.2%
_181802	-16178.172 ± 0.010	9.3%	_060333	-15992.428 ± 0.005	30.8%	_180458	-15913.918 ± 0.010	8.8%
GW190413	-15571.413 ± 0.006	22.5%	GW190731	-16376.777 ± 0.005	32.6%	GW200128	-16305.128 ± 0.013	6.1%
_052954	-15571.391 ± 0.005	26.3%	_140936	-16376.763 ± 0.005	31.0%	_022011	-16304.510 ± 0.007	18.3%
GW190413	-16399.331 ± 0.009	12.4%	GW190803	-16132.409 ± 0.006	21.4%	‡GW200129	-16226.851 ± 0.109	0.1%
_134308	-16399.139 ± 0.014	4.7%	_022701	-16132.408 ± 0.005	27.8%	_065458	-16231.203 ± 0.051	0.4%
GW190421	-15983.248 ± 0.008	15.3%	GW190805	-16073.261 ± 0.006	20.0%	GW200208	-16136.381 ± 0.007	16.6%
_213856	-15983.131 ± 0.010	9.4%	_211137	-16073.656 ± 0.007	16.6%	_130117	-16136.531 ± 0.009	11.2%
GW190503	-16582.865 ± 0.022	2.0%	GW190828	-16137.220 ± 0.009	12.2%	GW200208	-16775.200 ± 0.011	7.4%
_185404	-16583.352 ± 0.027	1.4%	_063405	-16136.799 ± 0.010	9.1%	_222617	-16774.582 ± 0.021	2.2%
GW190513	-15946.462 ± 0.043	0.6%	GW190909	-16061.634 ± 0.011	7.4%	GW200209	-16383.847 ± 0.009	12.5%
_205428	-15946.581 ± 0.017	3.4%	_114149	-16061.275 ± 0.016	3.8%	_085452	-16384.157 ± 0.025	1.6%
GW190514	-16556.466 ± 0.009	11.6%	GW190915	-16083.960 ± 0.015	20.8%	GW200216	-16215.703 ± 0.017	3.4%
_065416	-16556.314 ± 0.017	3.5%	_235702	-16083.937 ± 0.027	4.8%	_220804	-16215.540 ± 0.018	3.1%
GW190517	-16271.048 ± 0.027	1.3%	GW190926	-16015.813 ± 0.019	2.8%	GW200219	-16133.457 ± 0.011	9.6%
_055101	-16272.428 ± 0.034	0.9%	_050336	-16015.861 ± 0.009	12.1%	_094415	-16133.157 ± 0.017	4.0%
GW190519	-15991.171 ± 0.008	15.2%	GW190929	-16146.666 ± 0.018	3.2%	GW200220	-16303.782 ± 0.007	17.3%
_153544	-15991.287 ± 0.068	0.2%	_012149	-16146.591 ± 0.021	2.4%	_061928	-16303.087 ± 0.026	1.5%
GW190521	-16008.876 ± 0.008	13.4%	GW191109	-17925.064 ± 0.025	1.7%	GW200220	-16136.600 ± 0.008	13.2%
_074359	-16008.037 ± 0.015	4.2%	_010717	-17922.762 ± 0.041	0.6%	_124850	-16136.519 ± 0.037	0.7%
GW190527	-16119.012 ± 0.008	13.8%	GW191127	-16759.328 ± 0.019	2.7%	GW200224	-16138.613 ± 0.006	22.5%
_092055	-16118.781 ± 0.013	6.1%	_050227	-16758.102 ± 0.029	1.2%	_222234	-16139.101 ± 0.006	21.4%
GW190602	-16036.993 ± 0.006	25.0%	‡GW191204	-15984.455 ± 0.015	4.2%	‡GW200308	-16173.938 ± 0.013	6.0%
_175927	-16037.529 ± 0.006	23.5%	_110529	-15983.618 ± 0.063	0.3%	_173609	-16173.692 ± 0.025	1.7%
GW190701	-16521.381 ± 0.040	0.6%	GW191215	-16001.286 ± 0.013	5.8%	GW200311	-16117.505 ± 0.011	7.4%
_203306	-16521.609 ± 0.010	10.1%	_223052	-16000.846 ± 0.052	0.4%	_115853	-16117.583 ± 0.009	11.9%
GW190719	-15850.492 ± 0.008	13.4%	GW191222	-15871.521 ± 0.007	16.5%	‡GW200322	-16313.568 ± 0.307	0.0%
_215514	-15850.339 ± 0.011	8.0%	_033537	-15871.450 ± 0.005	25.8%	_091133	-16313.110 ± 0.105	0.1%

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/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, 21]. ‡See remarks on these events in text.

negligible 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 [18–20], since nested sampling generally provides a more accurate estimate than MCMC. In Tab. 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 $\log p(d) = -16412.89 \pm 0.01$ 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.

We now perform a large study analyzing all 42 events in GWTC-2 [21] and GWTC-3 [1] that are consistent with our prior. 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 (Tab. II). For most events, the initial DINGO results are already accurate and only deviate slightly from DINGO-IS; see the Supplemental Material for more detailed results. 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 [47–49] model and a smaller mass prior (in a study on drifting detector noise distributions [50]) 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 would likely be improved by increasing the prior range used for training. We also expect DINGO to struggle on data inconsistent with the training distribution, i.e., OOD data.

Indeed, we find that events with known data quality issues also have low sample efficiency (see Tab. II): several events analyzed are known to be contaminated by glitch artifacts (which would be mitigated in a more complete analysis [1, 21]); GW200129_065458 has been reported to not be well modeled by either of the waveform models used here due to strong precession [51]; and GW200322_091133 may have been generated by a Gaussian noise fluctuation [52]. DINGO-IS clearly identifies these events for additional investigation. Lastly, we find that DINGO-IS also identifies data that is intentionally corrupted to mislead the inference network ($\epsilon \approx 0.01\%$, see Supplemental Material), commonly referred to as adversarial examples [53, 54]. DINGO-IS thereby addresses some of the most common failure modes of neural networks.

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 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 GNPE 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. 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 two 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 10 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.

Acknowledgments.—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 (gw-openscience.org), 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 number 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` [55], `nflows` [56], `LALSimulation` [57] and the adam optimizer [58]. The plots are generated with `matplotlib` [59] and `ChainConsumer` [60].

* Equal contribution; maximilian.dax@tuebingen.mpg.de
† Equal contribution; stephen.green@aei.mpg.de

[1] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA), GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run, (2021), arXiv:2111.03606 [gr-qc].
[2] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA), Tests of General Relativity with GWTC-3, (2021), arXiv:2112.06861 [gr-qc].
[3] B. P. Abbott *et al.* (LIGO Scientific, Virgo), GW170817: Measurements of neutron star radii and equation of state, Phys. Rev. Lett. **121**, 161101 (2018), arXiv:1805.11581 [gr-qc].
[4] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA), The population of merging compact binaries inferred using gravitational waves through GWTC-3, (2021), arXiv:2111.03634 [astro-ph.HE].
[5] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA), Constraints on the cosmic expansion history from GWTC-3, (2021), arXiv:2111.03604 [astro-ph.CO].
[6] 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 **21**, 1087 (1953).
[7] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika **57**, 97 (1970), https://academic.oup.com/biomet/article-pdf/57/1/97/23940249/57-1-97.pdf.
[8] J. Skilling, Nested sampling for general Bayesian computation, Bayesian Analysis **1**, 833 (2006).
[9] K. Cranmer, J. Brehmer, and G. Louppe, The frontier of simulation-based inference, Proceedings of the National Academy of Sciences **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), arXiv:2106.12594 [gr-qc].
[11] P. Cannon, D. Ward, and S. M. Schmon, Investigating the impact of model misspecification in neural simulation-based inference, arXiv preprint arXiv:2209.01845 (2022).
[12] D. Rezende and S. Mohamed, Variational inference with normalizing flows, in *International Conference on Machine Learning* (2015) pp. 1530–1538, 1505.05770 [stat.ML].
[13] 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* (2016) pp. 4743–4751, arXiv:1606.04934 [cs.LG].
[14] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, in *Advances in Neural Information Processing Systems* (2019) pp. 7509–7520, arXiv:1906.04032 [stat.ML].
[15] 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).
[16] S. T. Tokdar and R. E. Kass, Importance sampling: a review, Wiley Interdisciplinary Reviews: Computational Statistics **2**, 54 (2010).
[17] J. Veitch *et al.*, Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library, Phys. Rev. **D91**, 042003 (2015), arXiv:1409.7215 [gr-qc].
[18] G. Ashton *et al.*, BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy, Astrophys. J. Suppl. **241**, 27 (2019), arXiv:1811.02042 [astro-ph.IM].
[19] 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. Roy. Astron. Soc. **499**, 3295 (2020), arXiv:2006.00714 [astro-ph.IM].
[20] J. S. Speagle, dynesty: a dynamic nested sampling package for estimating bayesian posteriors and evidences, Monthly Notices of the Royal Astronomical Society **493**, 3132–3158 (2020), arXiv:1904.02180 [astro-ph.IM].
[21] R. Abbott *et al.* (LIGO Scientific, Virgo), 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), arXiv:2010.14527 [gr-qc].
[22] 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), arXiv:2004.06503 [gr-qc].
[23] S. Ossokine *et al.*, Multipolar Effective-One-Body Waveforms for Precessing Binary Black Holes: Construction and Validation, Phys. Rev. D **102**, 044055 (2020), arXiv:2004.09442 [gr-qc].
[24] 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, Machine Learning: Science and Technology **2**, 011002 (2020), arXiv:2005.03745 [astro-ph.HE].
[25] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy (2019), arXiv:1909.06296 [astro-ph.IM].
[26] S. R. Green, C. Simpson, and J. Gair, Gravitational-wave parameter estimation with autoregressive neural network flows, Phys. Rev. D **102**, 104057 (2020), arXiv:2002.07656 [astro-ph.IM].
[27] 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, (2020), arXiv:2010.12931 [astro-ph.IM].
[28] S. R. Green and J. Gair, Complete parameter inference for GW150914 using deep learning, Mach. Learn. Sci. Tech. **2**, 03LT01 (2021), arXiv:2008.03312 [astro-ph.IM].
[29] C. Chatterjee, L. Wen, D. Beveridge, F. Diakogiannis, and K. Vinsen, Rapid localization of gravitational wave sources from compact binary coalescences using deep learning, (2022), arXiv:2207.14522 [gr-qc].
[30] M. J. Williams, J. Veitch, and C. Messenger, Nested sampling with normalizing flows for gravitational-wave inference, Phys. Rev. D **103**, 103006 (2021), arXiv:2102.11056 [gr-qc].
[31] T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák, Neural importance sampling, ACM Transactions

on Graphics (TOG) **38**, 1 (2019).
[32] H. Sun, K. L. Bouman, P. Tiede, J. J. Wang, S. Blunt, and D. Mawet, alpha-deep probabilistic inference (alpha-dpi): efficient uncertainty quantification from exoplanet astrometry to black hole feature extraction, arXiv preprint arXiv:2201.08506 (2022).
[33] 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), arXiv:2111.00833 [gr-qc].
[34] G. Papamakarios and I. Murray, Fast ε-free inference of simulation models with bayesian conditional density estimation (2016), arXiv:1605.06376 [stat.ML].
[35] A. Kong, A note on importance sampling using standardized weights, University of Chicago, Dept. of Statistics, Tech. Rep **348** (1992).
[36] A. B. Owen, *Monte Carlo theory, methods and examples* (2013).
[37] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, An introduction to variational methods for graphical models, Machine learning **37**, 183 (1999).
[38] M. J. Wainwright, M. I. Jordan, *et al.*, Graphical models, exponential families, and variational inference, Foundations and Trends® in Machine Learning **1**, 1 (2008).
[39] G. Papamakarios, T. Pavlakou, and I. Murray, Masked autoregressive flow for density estimation, in *Advances in Neural Information Processing Systems* (2017) pp. 2338–2347, arXiv:1705.07057 [stat.ML].
[40] J. Aasi *et al.* (LIGO Scientific), Advanced LIGO, Class. Quant. Grav. **32**, 074001 (2015), arXiv:1411.4547 [gr-qc].
[41] F. Acernese *et al.* (VIRGO), Advanced Virgo: a second-generation interferometric gravitational wave detector, Class. Quant. Grav. **32**, 024001 (2015), arXiv:1408.3978 [gr-qc].
[42] J. Veitch and W. Del Pozzo, Analytic marginalisation of phase parameter, URL: https://dcc. ligo. org/LIGO-T1300326/public (2013).
[43] E. Thrane and C. Talbot, An introduction to Bayesian inference in gravitational-wave astronomy: parameter estimation, model selection, and hierarchical models, Publ. Astron. Soc. Austral. **36**, e010 (2019), [Erratum: Publ.Astron.Soc.Austral. 37, e036 (2020)], arXiv:1809.02293 [astro-ph.IM].
[44] M. Dax, S. R. Green, J. Gair, M. Deistler, B. Schölkopf, and J. H. Macke, Group equivariant neural posterior estimation, in *International Conference on Learning Representations* (2022) arXiv:2111.13139 [cs.LG].
[45] 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), arXiv:1502.04370 [gr-qc].
[46] J. Lange, R. O'Shaughnessy, and M. Rizzo, Rapid and accurate parameter inference for coalescing, precessing compact binaries, (2018), arXiv:1805.10457 [gr-qc].
[47] 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), arXiv:1308.3271 [gr-qc].
[48] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jimńez Forteza, and A. Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era, Phys. Rev. **D93**, 044007 (2016), arXiv:1508.07253 [gr-qc].
[49] A. Bohé, M. Hannam, S. Husa, F. Ohme, M. Pürrer, and P. Schmidt, PhenomPv2 – technical notes for the LAL implementation, LIGO Technical Document, LIGO-T1500602-v4 (2016).
[50] TBA, Tba, (2022).
[51] M. Hannam, C. Hoy, J. E. Thompson, S. Fairhurst, and V. Raymond (VIRGO), Measurement of general-relativistic precession in a black-hole binary, (2021), arXiv:2112.11300 [gr-qc].
[52] 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, (2022), arXiv:2209.05475 [gr-qc].
[53] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks, in *International Conference on Learning Representations* (2014) arXiv:1312.6199 [cs.LG].
[54] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, in *International Conference on Learning Representations* (2015) arXiv:1412.6572 [cs.LG].
[55] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, 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.
[56] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, nflows: normalizing flows in PyTorch (2020).
[57] LIGO Scientific Collaboration, LIGO Algorithm Library - LALSuite, free software (GPL) (2018).
[58] D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, (2014), arXiv:1412.6980 [cs.LG].
[59] J. D. Hunter, Matplotlib: A 2d graphics environment, Computing in Science & Engineering **9**, 90 (2007).
[60] S. R. Hinton, ChainConsumer, The Journal of Open Source Software **1**, 00045 (2016).

Supplemental Material

IMPORTANCE-SAMPLED BAYESIAN EVIDENCE

The Bayesian evidence is given by

$$p(d) = \int d\theta p(d|\theta)p(\theta) = \int d\theta \frac{p(d|\theta)p(\theta)}{q(\theta|d)} q(\theta|d), \quad (2)$$

which can be estimated using n samples $\theta_i \sim q(\theta|d)$ in the Monte Carlo approximation as $p(d) = \hat{\mu}_w$ with

$$\hat{\mu}_w = \frac{1}{n}\sum_i \frac{p(d|\theta_i)p(\theta_i)}{q(\theta_i|d)} = \frac{1}{n}\sum_i w_i \quad (3)$$

where $w_i = p(d|\theta_i)p(\theta_i)/q(\theta_i|d)$ are the weights used for importance sampling. The variance for this Monte Carlo estimate is given by

$$\begin{aligned}\sigma_w^2 &= \mathrm{Var}\left[\frac{p(d|\theta)p(\theta)}{q(\theta|d)}\right] \approx \frac{1}{n}\sum_i (w_i - \hat{\mu}_w)^2 \\ &= \hat{\mu}_w^2 \cdot \frac{1}{n}\sum_i [\bar{w}_i - 1]^2 = \hat{\mu}_w^2 \cdot \left(\frac{1}{n}\sum_i \bar{w}_i^2 - 1\right) \\ &= \hat{\mu}_w^2 \cdot \left(\frac{n - n_{\mathrm{eff}}}{n_{\mathrm{eff}}}\right) = \hat{\mu}_w^2 \cdot \left(\frac{1-\epsilon}{\epsilon}\right),\end{aligned} \quad (4)$$

where we denote normalized weights with $\bar{w}_i = w_i/\hat{\mu}_w$ and the sample efficiency with $\epsilon = n_{\mathrm{eff}}/n$. Since we use n samples to estimate $p(d) = \hat{\mu}_w$, the standard deviation of the evidence is given by

$$\sigma_{p(d)} = \frac{\sigma_w}{\sqrt{n}} = p(d)\sqrt{\frac{1-\epsilon}{n\cdot\epsilon}}. \quad (5)$$

In practice, we are interested in the log evidence, for which the uncertainty is

$$\sigma_{\log p(d)} = \frac{\sigma_{p(d)}}{p(d)} = \sqrt{\frac{1-\epsilon}{n\cdot\epsilon}}. \quad (6)$$

Bias

Since $p(\theta)$ and $q(\theta|d)$ are normalized, Eq. (2) provides an unbiased estimate for $p(d)$ [36],

$$\mathbb{E}\left[\frac{1}{n}\sum_i w_i\right] = \mathbb{E}[\hat{\mu}_w] = p(d). \quad (7)$$

The logarithm of the evidence however has a bias. Defining $Y = \hat{\mu}_w - p(d)$, we find

$$\begin{aligned}\mathbb{E}[\log\hat{\mu}_w] &= \mathbb{E}\left[\log\left(p(d) + p(d)\cdot\frac{\hat{\mu}_w - p(d)}{p(d)}\right)\right] \\ &= \log p(d) + \mathbb{E}\left[\log\left(1 + \frac{Y}{p(d)}\right)\right] \\ &= \log p(d) + \mathbb{E}\left[\frac{Y}{p(d)} - \frac{1}{2}\left(\frac{Y}{p(d)}\right)^2\right] \\ &= \log p(d) - \frac{\sigma_w^2}{2p(d)^2 n} = \log p(d) - \frac{1-\epsilon}{2n\epsilon}\end{aligned} \quad (8)$$

where we used $\mathbb{E}[Y] = 0$ and $\mathrm{Var}[Y] = \sigma_w^2/n$ and neglected terms of order $\mathcal{O}((Y/p(d))^3)$. The bias of the log evidence thus depends on the sample efficiency $\epsilon = n_{\mathrm{eff}}/n$ and scales with $1/n$. Given that the uncertainty of $\log\hat{\mu}_w$ scales with $1/\sqrt{n}$, this bias is completely negligible in practice.

ANALYTIC ESTIMATE OF THE PHASE PARAMETER

The parameter ϕ_c describes the phase of the gravitational wave at a fixed reference frequency. It provides no physical insight, but it is necessary to define a complete likelihood [42]. While the marginal $p(\phi_c|d)$ usually has a simple structure, the *conditional* distribution $p(\phi_c|d,\tilde{\theta})$, where $\tilde{\theta}$ denotes the 14 remaining parameters, is typically very tightly constrained. Furthermore, ϕ_c is strongly correlated with $\tilde{\theta}$. We observed that DINGO has difficulties learning the phase parameter, and often infers the prior instead, $q(\phi_c|d,\tilde{\theta}) = p(\phi_c)$. While we did not find this to have a negative impact on the remaining parameters, it leads to a substantially reduced sample efficiency.

Inspired by phase marginalization [42, 43], a technique commonly used to increase the efficiency of stochastic samplers, we analytically estimate ϕ_c. The approach outlined below differs in two ways from typical phase marginalization—(1) we retrieve ϕ_c instead of marginalizing over it, and (2) this technique is exact even in the presence of higher modes, where phase marginalization is an approximation.

We decompose our posterior estimate into

$$q(\theta|d) = p(\phi_c|d,\tilde{\theta})q(\tilde{\theta}|d), \quad (9)$$

where $q(\tilde{\theta}|d)$ is estimated with DINGO. For each DINGO sample $\tilde{\theta} \sim q(\tilde{\theta}|d)$, we then *synthetically* sample ϕ_c using the analytic likelihood. This is done by evaluating $p(\phi_c|d,\tilde{\theta})$ on a uniform grid over ϕ_c with 5001 points in the range $[0, 2\pi]$ and interpolating in between.

Each likelihood evaluation requires a waveform simulation, which accounts for the bulk of the computational cost. As we outline below, by caching suitable combinations of the waveform modes, we can cheaply evaluate

waveform polarizations for arbitrary ϕ_c. Hence sampling the synthetic ϕ_c is barely more expensive than a single likelihood evaluation.

Phase transformations

We work in the L_0 frame, which aligns the z axis with the orbital angular momentum of the binary at the reference frequency, and takes ϕ_c as the azimuthal angle of the observer relative to the axis connecting the two bodies. In these coordinates, the observer is located at $(\theta, \phi) = (\iota, \pi/2 - \phi_c)$, where ι is the inclination of the binary. This is convenient for caching the modes, since ϕ_c enters the waveform entirely via the spin-weighted spherical harmonics (as opposed to the modes themselves).

Waveform modes $h_{\ell m}$ combine into polarizations $h_{+,\times}$ as

$$h_+ - ih_\times = h = \sum_{\ell,m} h_{\ell m}\, {}_{-2}Y_{\ell m}(\theta, \phi), \tag{10}$$

In frequency domain,

$$\tilde{h}_+(f) = \frac{1}{2}\left[\tilde{h}(f) + \tilde{h}^*(-f)\right], \tag{11}$$

$$\tilde{h}_\times(f) = \frac{i}{2}\left[\tilde{h}(f) - \tilde{h}^*(-f)\right]. \tag{12}$$

Considering just the plus polarization and substituting for the mode expansion,

$$\tilde{h}_+(f) = \frac{1}{2}\sum_{\ell,m}\Big[\tilde{h}_{\ell m}(f)\, {}_{-2}Y_{\ell m}(\theta, \phi) + \tilde{h}^*_{\ell m}(-f)\, {}_{-2}Y^*_{\ell m}(\theta, \phi)\Big]. \tag{13}$$

Now we use the fact that the ϕ-dependence enters the spin-weighted spherical harmonics as ${}_{-2}Y_{\ell m}(\theta, \phi) = {}_{-2}Y_{\ell m}(\theta, 0)e^{im\phi}$. Since h_+ is real, we only need to consider $f > 0$. In the L_0 frame, we can then write

$$\tilde{h}_+(f > 0) = \sum_m \tilde{h}_{+,m}(f) e^{-im\phi_c}, \tag{14}$$

where we have grouped the terms according to their m-dependence,

$$\tilde{h}_{+,m}(f) = \frac{1}{2}\sum_\ell \Big[\tilde{h}_{\ell m}(f)\, {}_{-2}Y_{\ell m}\left(\iota, \frac{\pi}{2}\right) + \tilde{h}^*_{\ell,-m}(-f)\, {}_{-2}Y^*_{\ell,-m}\left(\iota, \frac{\pi}{2}\right)\Big]. \tag{15}$$

Notice that we combined the positive frequency parts of modes with azimuthal number m together with negative frequency modes of azimuthal number $-m$. With this decomposition, we only need to cache the $\tilde{h}_{+,m}$. Likewise for the cross polarization, we have

$$\tilde{h}_\times(f > 0) = \sum_m \tilde{h}_{\times,m}(f) e^{-im\phi_c}, \tag{16}$$

where

$$h_{\times,m}(f) = \frac{i}{2}\sum_\ell \Big[\tilde{h}_{\ell m}(f)\, {}_{-2}Y_{\ell m}\left(\iota, \frac{\pi}{2}\right) - \tilde{h}^*_{\ell,-m}(-f)\, {}_{-2}Y^*_{\ell,-m}\left(\iota, \frac{\pi}{2}\right)\Big]. \tag{17}$$

One additional complication arises because waveform models are usually given in terms of Cartesian spin components, and ϕ_c also enters into their definition in terms of the spin parameters used for parameter estimation. Consequently the modes retain a dependence on ϕ_c. We overcome this by fixing the phase parameter used in effecting this transformation. This results in a slightly different definition of the spin parameters θ_{JN} and ϕ_{JL}, which we undo in post-processing. Since the standard priors are invariant under this transformation, other parameters are not affected.

This approach enables likelihood evaluations on a ϕ_c grid at the computational cost of a single likelihood evaluation, plus a small additional cost for the inner products.[5] The implementation is fully contained in the DINGO package, which uses low-level LALSimulation [57] functions to compute frequency domain modes in the L_0 frame, and combines them into the $\tilde{h}_{+/\times,m}$. For SEOBNRv4PHM, this requires Fourier transforming the time domain modes provided by LALSIMULATION in L_0 frame. For IMRPhenomXPHM it requires transforming from J to L_0 frame, such that the ϕ_c dependence enters via the spherical harmonics, not via the modes themselves.

DENSITY RECOVERY

IS requires access to the density of the inferred samples. While for NPE, this density is tractable, this is not necessarily the case for other inference methods. Below, we describe how we use neural density estimation to recover the density in these cases.

Group equivariant neural posterior estimation

DINGO uses an iterative algorithm called *group equivariant* NPE (GNPE) [10, 44] to integrate physical symmetries and thereby improve the accuracy of inference. With GNPE, we train a density estimation network $q(\theta|d, \hat{t}_I(\theta))$ that is also conditional on a set of GNPE *proxy parameters* $\hat{t}_I$. These parameters are defined as blurred versions of

[5] For IMRPhenomXPHM, computing the individual modes with the LALSIMULATION function `SimInspiralChooseFDModes` is substantially more expensive than computing the combined polarizations with `SimInspiralFD`. This is because `SimInspiralFD` caches information when internally computing the modes, whereas `SimInspiralChooseFDModes` does not.

	2 dimensions	14 dimensions
flow steps	5	20
hidden dimension	256	256
transform blocks	4	4
bins	8	8
training samples	$4 \cdot 10^5$	10^6
batch size	4096	8192
epochs	20	60
optimizer	adam [58]	adam [58]
learning rate	0.002	0.001
training time on A100 GPU	7 minutes	1 hour

Table III. Settings for the neural spline flow [14] architecture (upper part) and training (lower) used for density recovery. For DINGO-IS with GNPE, we need to estimate a two dimensional distribution over the proxy parameters, which requires a smaller network than the distribution over the 14 dimensional parameter space used for BILBY-IS.

the coalescence times t_I in the individual interferometers (which can be computed as a function of θ) as

$$\hat{t}_I = t_I + \epsilon_I, \quad \epsilon_I \sim \kappa(\epsilon), \tag{18}$$

with $\kappa = U[-1\text{ ms}, 1\text{ms}]$. With GNPE, we iteratively infer the posterior $p(\theta, \hat{t}_I|d)$ in the joint parameter space with Gibbs sampling, and obtain the posterior over θ by marginalizing over $\hat{t}_I$. We use a paralllelized Gibbs sampler that typically converges after 30 iterations, but some events require up to 500 iterations. Each iteration corresponds to a forward pass through the density estimator $q(\theta|d, \hat{t}_I(\theta))$. 500 GNPE iterations for a batch of $5 \cdot 10^4$ samples take about 6 minutes on an A100 GPU.

In contrast to NPE, GNPE does not have a tractable density. To recover the density, we first generate $4 \cdot 10^5$ GNPE samples (48 minutes on one GPU or 6 minutes on eight GPUs for 500 iterations). We then train an unconditional normalizing flow $q(\hat{t}_I)$ to estimate the distribution over the inferred proxy parameters with a maximum likelihood objective. We use a neural spline flow with rational-quadratic spline coupling transforms [14] with the hyperparameters from Tab. III. Once trained, we can sample without the need for additional GNPE iterations via

$$\theta \sim q(\theta|d, \hat{t}_I), \;\; \hat{t}_I \sim q(\hat{t}_I). \tag{19}$$

The proposal density is now tractable,

$$\log q(\theta, \hat{t}_I|d) = \log q(\theta|d, \hat{t}_I) + \log q(\hat{t}_I). \tag{20}$$

We then perform IS in the *joint* parameter space $(\theta, \hat{t}_I)$, where the target density is given by

$$\begin{aligned}\log p(\theta, \hat{t}_I|d) = &- \log p(d) + \log p(d|\theta) + \log p(\theta) \\ &+ \sum_I \log \kappa(\hat{t}_I - t_I).\end{aligned} \tag{21}$$

The last term accounts for $p(\hat{t}_I|\theta)$. As described in the main part, we omit $\log p(d)$ and estimate this from the normalization of the weights.

Alternatively, we could also train an unconditional density estimator for the converged θ samples, but this is less sample efficient and more costly to train.

Stochastic samplers

We apply IS to BILBY-DYNESTY [18–20], which is based on nested sampling. To recover the density, we first generate $\approx 10^6$ posterior samples with 50 BILBY runs with identical settings. With `nlive=1000` and `nact=5`, this takes about one day per run on 10 CPUs, when using the IMRPhenomXPHM model. One typically uses larger `nact` for production results, but this substantially increases the computational cost. For reference, the runs for GW150914 and GW151012 reported in the main paper with `nlive=4000` and `nact=50` took about a week. We then estimate the distribution over the BILBY samples by training an unconditional normalizing flow $q(\theta)$, see Tab. III. To ensure a fair comparison with DINGO, we also use the analytic estimate for the phase parameter, such that we only need to estimate the distribution over the remaining 14 parameters. Due to the higher dimensional parameter space compared to DINGO (for which we only need to recover the two dimensional density over $\hat{t}_I$), we need more samples and a larger normalizing flow for the density estimate.

For GW151012, BILBY-IS achieves a sample efficiency $\epsilon = 8.3\%$, compared to $\epsilon = 12.5\%$ for DINGO-IS, and estimates an evidence of $\log p(d) = -16412.89 \pm 0.01$. Since BILBY-IS is computationally very expensive, we do not expect it to be routinely used, but rather view it as an insightful diagnostic.

IMPORTANCE SAMPLING CONVERGENCE

Due to the probability mass covering training objective, DINGO inaccuracies tend to show up as overly broad posteriors (Fig. 2). When the tails of the posterior are overestimated by DINGO, a low sample efficiency may be encountered due to many low-weight samples. These cases are straightforward to handle with DINGO-IS. The sample efficiency is approximately constant, and the statistical uncertainty of the evidence fully captures the error, even for low n_{eff}. To get smooth marginals one simply needs to generate more samples, which is cheap with DINGO.

In contrast, DINGO posteriors should rarely be light tailed. For real data, however, parts of the parameter space are occasionally strongly undersampled, which is problematic for IS. Indeed, for small n, the light tails may not be sampled at all, resulting in an underestimate of the evidence and the magnitude of its statistical error.

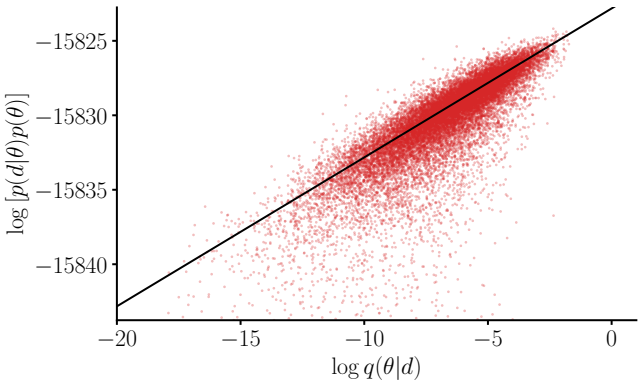

Figure 2. DINGO samples $\theta \sim q(\theta|d)$ for GW150914, comparing the inferred density $q(\theta|d)$ to the unnormalized posterior $p(d|\theta)p(\theta)$. The density ratios correspond to the importance weights, the Bayesian evidence $p(d)$ is estimated via their normalization. Samples of a perfect DINGO model would lie on the black line with offset $\log p(d) = -15831.87$. Deviations between DINGO and the true posterior are primarily found below that line, but rarely above. This is a manifestation of the probability-mass covering behavior, making DINGO particularly well-suited for importance sampling.

Moreover, when a sample from the tail is encountered it has very large importance weight, which greatly decreases the sample efficiency. In order to assess the validity of IS results with low sample efficiency it is therefore useful to check whether $\log p(d)$ has converged as a function of n (Fig. 3). If DINGO is not truly mass covering, the IS weights are not upper-bounded, and the sample efficiency approaches zero with increasing n. This happens for the OOD event GW200129_065458.

Fortunately non-convergence is rare, and for the majority of events, DINGO posteriors are indeed mass covering and heavy tailed. Even when this is not the case and the sample efficiency is very low, the DINGO marginals are often still accurate. This is because the light tailed parts of the parameter space are often negligibly small and randomly distributed throughout the parameter space. In such cases one can apply *batched* self-normalized IS: instead of normalizing the weights of all n samples simultaneously, one normalizes batches of size $k < n$. This regularizes IS by decreasing the largest possible weight from n to k. This should be done with caution, as it introduces a bias which is only small if the undersampled regions carry an overall low probability mass, or are distributed unsystematically throughout the parameter space.

ROBUSTNESS TO ADVERSARIAL EXAMPLES

An adversarial example [53, 54] refers to data $d_{\text{adv.}}$ that is specifically designed to mislead a neural network. Such examples can be generated by following gradients of the network output (or some function thereof) starting from some real data d_{true} and sequentially adding small perturbations to maximally change the output. Although the resulting adversarial example $d_{\text{adv.}}$ is often barely distinguishable from d_{true}, the neural network output can change dramatically.

In the context of posterior estimation, the output is a high-dimensional distribution which one can alter in multiple ways. We tried to shift or truncate the predicted DINGO distribution $q(\theta|d_{\text{adv.}})$ by applying only minimal modifications to the data. We found that DINGO is remarkably robust to such attacks, its output could barely be changed without significantly changing the input data d. This unusual robustness is attributed to two factors. First, the training data itself is very noisy, which regularizes DINGO models. Second, the first layer of DINGO networks is seeded with principal components of clean GW signals [10], so adversarial perturbations are projected onto the manifold of GW signals.

We thus explore a slightly different notion of adversarial attacks. Starting from strain data d initialized with random Gaussian noise, we aim to modify d such that DINGO estimates identical posteriors for $d_{\text{adv.}}$ and the real strain data d_{true} for GW150914. Specifically, we minimize the KL divergence $D_{\text{KL}}(q(\theta|d_{\text{true}})||q(\theta|d_{\text{adv.}}))$ via

$$d_{\text{adv.}} = \underset{d}{\text{argmax}}\, \mathbb{E}_{\theta\sim q(\theta|d_{\text{true}})} \log q(\theta|d). \tag{22}$$

In contrast to the technique mentioned above, we here do not constrain the difference between $d_{\text{adv.}}$ and d_{true} to be small. To optimize (22) we need to take gradients of the DINGO density with respect to d, which is intractable with the iterative GNPE [10, 44] method. Instead, we use a DINGO network trained with standard NPE. We use the adam [58] optimizer with a learning rate of 0.03 to optimize Eq. (22) with 400 gradient steps (batch size 1024). The resulting strain d_{adv} is visibly different from the true GW150914 strain d_{true}, but the estimated DINGO posteriors are almost identical (Fig. 4).

With DINGO-IS, we find a sample efficiency of $\epsilon = 1.48\%$ for the real GW150914 strain d_{true}. This is substantially smaller than the sample efficiency achieved with GNPE ($\epsilon = 28.8\%$), since standard NPE does not use the physical symmetries and is hence less accurate. However, the DINGO-IS posterior is still accurate and the evidence estimate ($\log p(d) = -15831.88 \pm 0.03$) is in good agreement with the result reported in the main paper. For $d_{\text{adv.}}$ on the other hand, DINGO-IS achieves a sample efficiency of $\epsilon = 0.006\%$, clearly identifying the adversarial example as a DINGO failure case.

ADDITIONAL RESULTS

Fig. 5 shows one-dimensional marginal posteriors for a subset of GW events analyzed in the main paper, com-

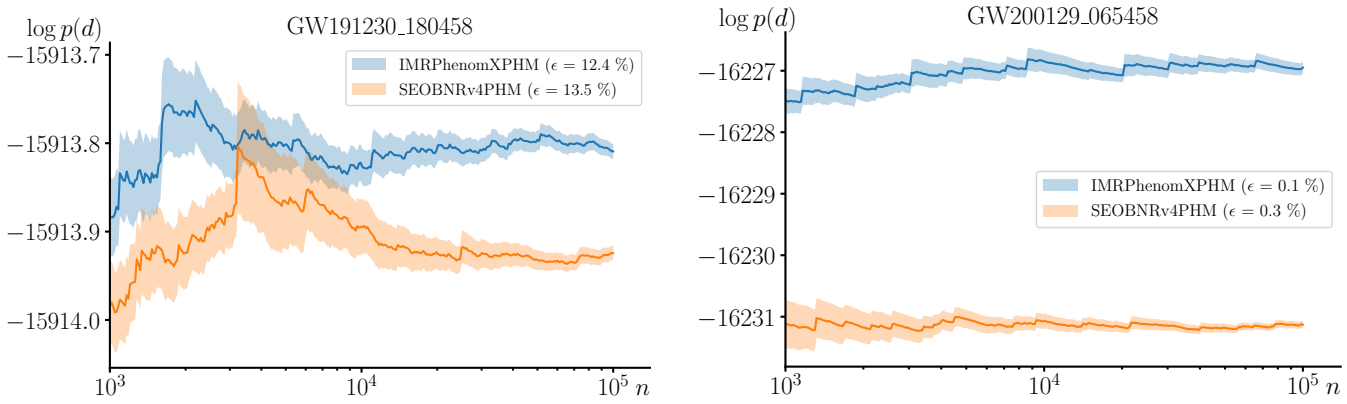

Figure 3. Evidence $\log p(d)$ as a function of the number of importance samples n. For constant sample efficiency ϵ, the statistical uncertainty scales with $1/\sqrt{n}$, leading to precise estimates when DINGO-IS works well (left). When the DINGO posterior is too light tailed (right), samples from the tails of the distribution are assigned very large IS weights, leading to bumps in the evidence whenever a high-weight sample is encountered.

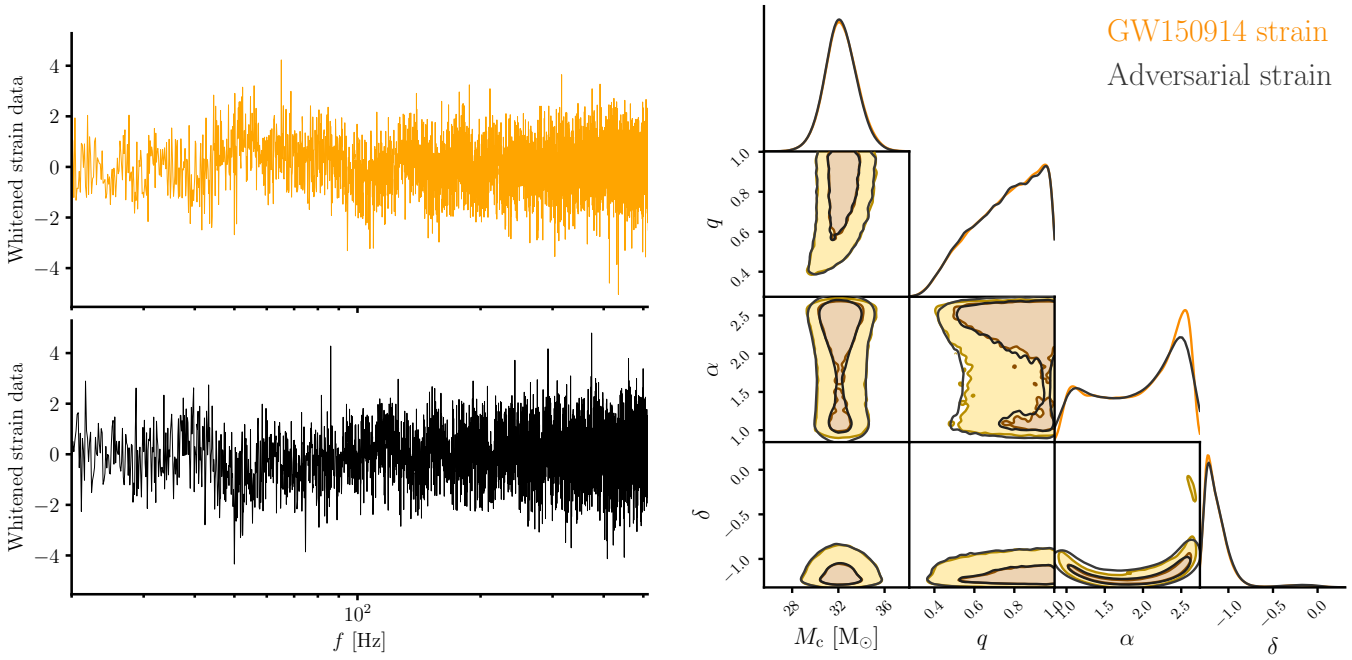

Figure 4. Left: Strain data (real part) in the LIGO Hanford detector. The upper row shows the measured data for GW150914, the lower row shows an adversarial example that is synthetically generated to mislead the inference network. Right: The inference network infers almost identical posteriors for both strain datasets.

paring the two waveform models IMRPhenomXPHM and SEOBNRv4PHM. We see that the models give results that appear to be in good agreement.

Fig. 6 shows posterior marginals for several GW events from O3. A large sample efficiency often corresponds to good agreement of the DINGO and DINGO-IS marginals. The sample efficiency is sensitive to deviations in the full 15 dimensional parameter space, so small sample efficiencies do not necessarily imply inaccurate *marginal* distributions.

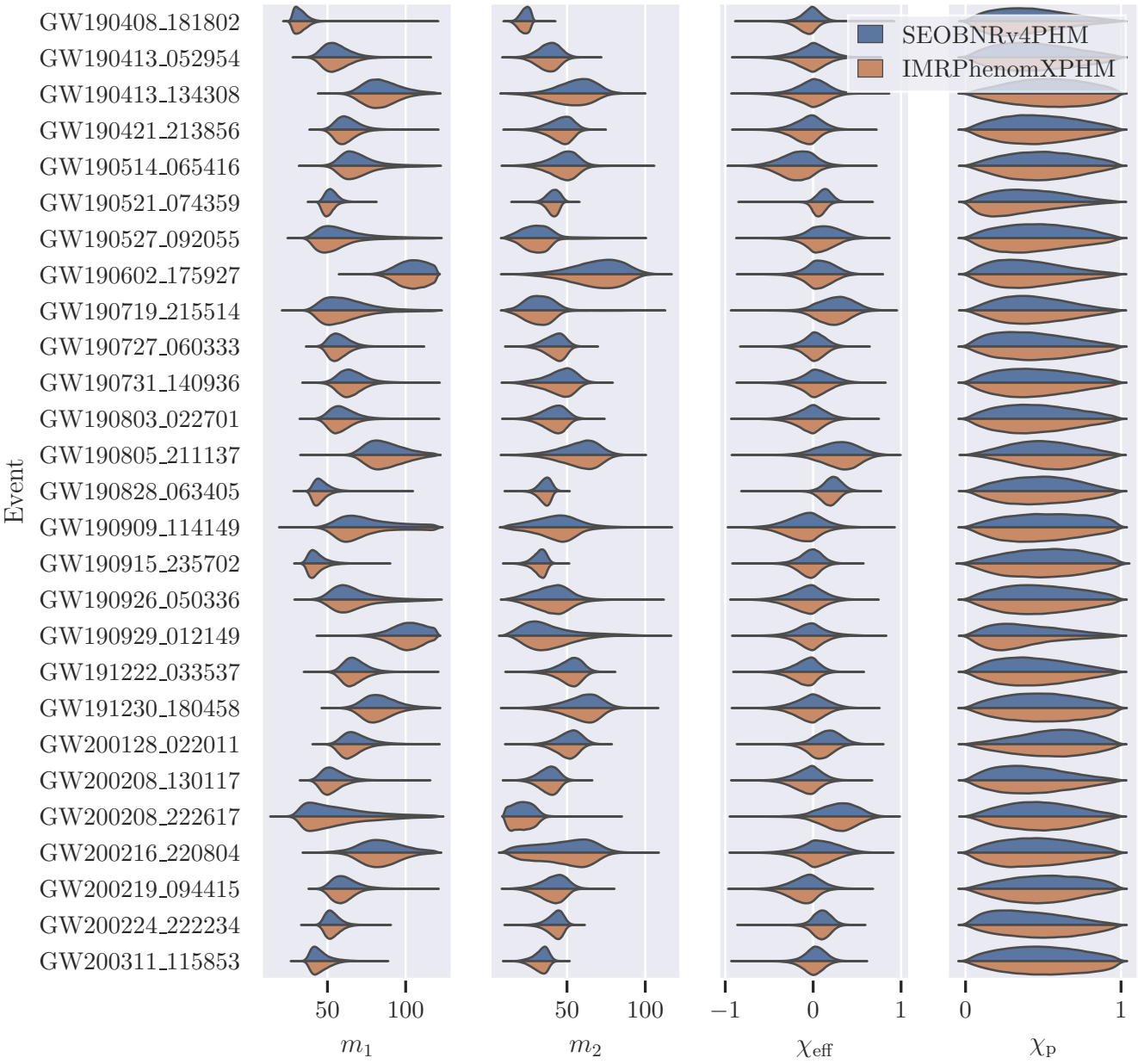

Figure 5. Posterior distributions for component masses and effective spin parameters, for those events from the main paper with $\epsilon > 2\%$ for both waveform models. This shows good agreement between the two waveform models. In a future publication we will include a more complete catalog, which incorporates Virgo data, and includes a more careful treatment of noise artifacts and data conditioning.

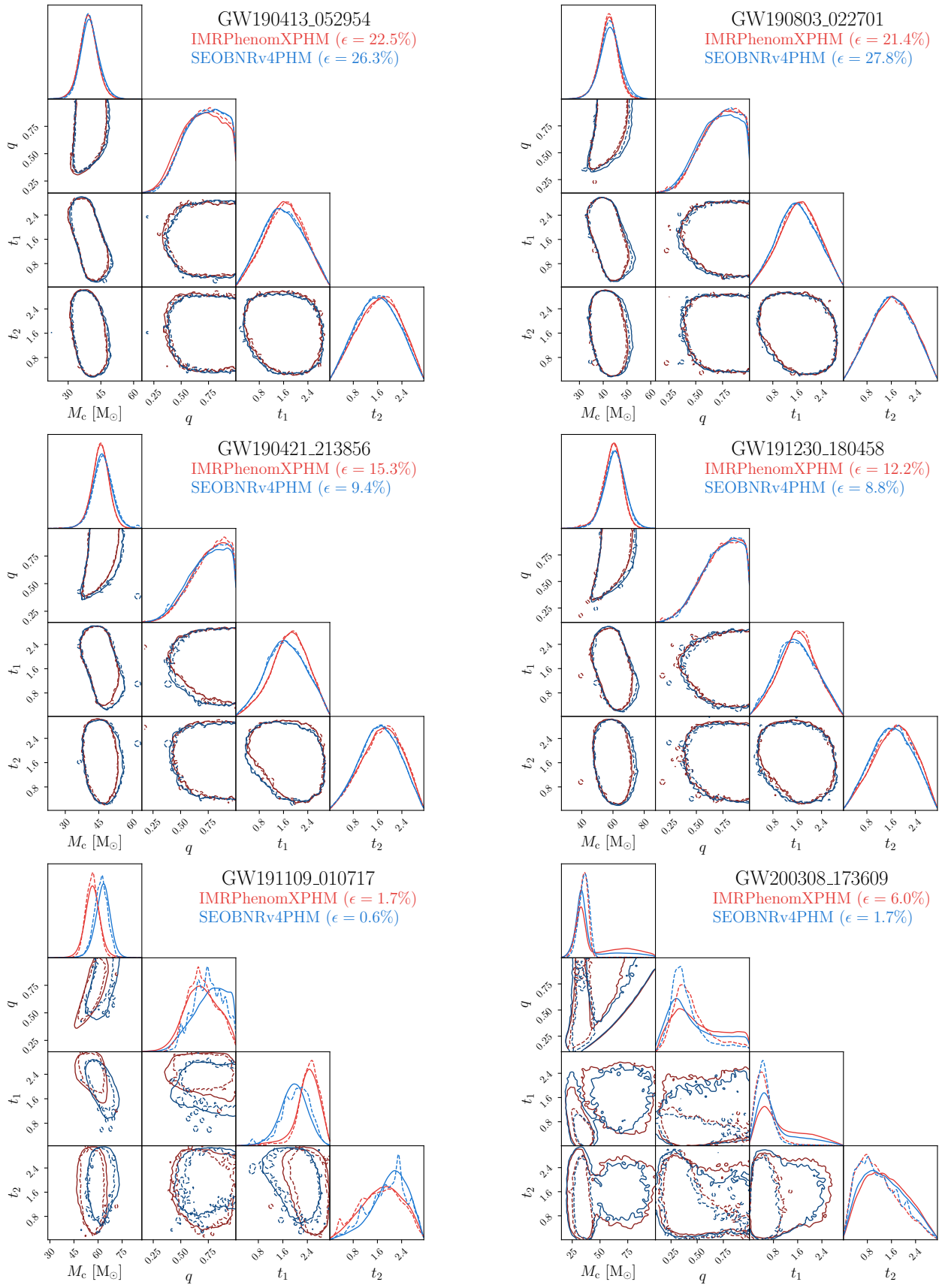

Figure 6. Marginalized one- and two- dimensional posterior distributions for selected O3 events, comparing DINGO (solid lines) and DINGO-IS (dashed) inference results with waveform models IMRPhenomXPHM and SEOBNRv4PHM. Contours represent 90% credible regions. For events with high (upper row) or medium (middle row) sample efficiency, the initial DINGO results are often accurate and only deviate slightly from DINGO-IS results. For events with low effective sample size (lower row), the DINGO-IS contours are often not smooth. Yet, the initial DINGO results may capture the marginals well, see GW191109_010717.