Real-Time Gravitational Wave Science with Neural Posterior Estimation

Maximilian Dax,[1,*] Stephen R. Green,[2,†] Jonathan Gair,[2,‡] Jakob H. Macke,[1,3] Alessandra Buonanno,[2,4] 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]*Machine Learning in Science, University of Tübingen, 72076 Tübingen, Germany*
[4]*Department of Physics, University of Maryland, College Park, Maryland 20742, USA*

(Received 1 July 2021; accepted 17 November 2021; published 8 December 2021)

We demonstrate unprecedented accuracy for rapid gravitational wave parameter estimation with deep learning. Using neural networks as surrogates for Bayesian posterior distributions, we analyze eight gravitational wave events from the first LIGO-Virgo Gravitational-Wave Transient Catalog and find very close quantitative agreement with standard inference codes, but with inference times reduced from $O(\text{day})$ to 20 s per event. Our networks are trained using simulated data, including an estimate of the detector noise characteristics near the event. This encodes the signal and noise models within millions of neural-network parameters and enables inference for any observed data consistent with the training distribution, accounting for noise nonstationarity from event to event. Our algorithm—called "DINGO"—sets a new standard in fast and accurate inference of physical parameters of detected gravitational wave events, which should enable real-time data analysis without sacrificing accuracy.

DOI: 10.1103/PhysRevLett.127.241103

Introduction.—Since the first detection of a signal from a pair of merging black holes [1], gravitational waves have quickly emerged as an important new probe of gravitational theory [2], neutron star physics [3], cosmology [4], and black hole astrophysics [5]. These scientific successes were made possible by a growing rate of detections by the LIGO [6] and Virgo [7] observatories and their subsequent analysis and characterization as signals from merging compact binary systems. The LIGO and Virgo Collaborations (LVC) have now published results from over 50 such systems [8,9], and this number promises to grow ever faster as detectors are made more sensitive in the future [10].

Given a detection, Bayesian inference is used to characterize the originating source [11]. This is based on having models for the signals and the detector noise. For gravitational waves, signal models take the form of waveform predictions $h(\theta)$ depending on the source parameters θ (masses, location, etc.). Waveform models are based on solutions to Einstein's equations (and any relevant matter equations) for the two-body dynamics and gravitational radiation, using a combination of numerical relativity and perturbative calculations [12–14] and phenomenological fitting [14–16]. Detector noise is typically modeled as stationary and Gaussian, with some spectrum that can be estimated empirically. Together, these "forward" models give rise to the likelihood $p(d|\theta)$ for the observed strain data d, which is assumed to consist of a signal plus noise. With the choice of a prior $p(\theta)$ over parameters, the posterior distribution is given via Bayes' theorem,

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

where $p(d)$ is a normalizing factor called the evidence. The posterior gives our belief about the source parameters, given the observed data.

The task of inference is to characterize the posterior by drawing samples from it. This can be accomplished with stochastic algorithms like Markov chain Monte Carlo (MCMC) methods. The LVC have developed software tools such as LALINFERENCE [17] and BILBY [18–20] to carry this out. However, these algorithms are computationally expensive as they require many likelihood evaluations for each independent posterior sample $\theta \sim p(\theta|d)$, and each likelihood requires a waveform simulation. An analysis producing $\sim 10^4$ independent samples typically requires millions of waveform evaluations and a total inference time of hours to months, depending on the signal duration and waveform model. More physically realistic waveform models [21] are also more costly, so carrying out inference for all events with the best models is an enormous computational effort. When rapid results are desired—for alerts to trigger electromagnetic follow-up of transient phenomena [22] or when processing large numbers of

events—accuracy usually has to be traded off for speed, by restricting to a limited set of fast models [23,24] or specialized inference algorithms [25–27].

In this Letter, we describe an alternative approach to gravitational wave inference that delivers both dramatically reduced analysis time *and* high accuracy, in stark contrast to the trade-off intrinsic to standard algorithms. The basic idea is to produce a large number of simulated datasets (with associated parameters) and use these to train a type of neural network known as a "normalizing flow" to approximate the posterior. The trained network can then generate new posterior samples extremely quickly once a detection is made. This bypasses the need to generate waveforms at inference time, thereby amortizing the expensive training costs over all future detections. The general approach of building such "surrogate" inverse models is called neural posterior estimation (NPE) [28–30] and is beginning to see application in several scientific domains [31]. When applied to gravitational waves, with all of the optimizations we describe, we call the method deep inference for gravitational wave observations (DINGO).

NPE and conventional methods both involve the same inputs: a prior and a likelihood. A key difference, however, is the way in which the likelihood is used: for conventional methods, its density is *evaluated*, whereas for NPE it is used to *simulate* data, i.e., $d \sim p(d|\theta)$. This distinction is important when dealing with nonstationary or non-Gaussian detector noise, for which an analytic likelihood is either expensive or unavailable. In this case, one could nevertheless simulate data, in a noise-model-independent way, by injecting simulated signals into real noise. Our present focus is on speed and on validating DINGO on real data with the common assumption of stationary-Gaussian noise, but the ultimate aim of more accurate inference using real noise should be kept in mind.

There have been several previous studies that applied NPE or related approaches to gravitational waves [32–39]; see also [40]. However, most of these are limited in some way: they either restrict the number of parameters or the distributional form of the posterior, they do not analyze real data, or there are clear deviations from results obtained using standard algorithms. The best performance to date was achieved in the study [36] by some of us. This was the only study to infer all 15 parameters [41] of a binary black hole (BBH) system in real data and demonstrate close agreement to standard samplers. However, even that study did not achieve full amortization, as it did not address the fact that detector noise varies from event to event. Rather, the neural network of [36] was tuned to the noise power spectral densities (PSDs) of the detectors at the time of the analyzed event, and it would require retraining for each new event. We now present for the first time completely amortized inference for BBHs using DINGO. This is achieved by *conditioning* the neural network not only on the event strain data, but also on the detector noise PSD, which can be estimated using nearby data [17]. We also achieve unprecedented accuracy thanks to a new iterative algorithm for time shifting the coalescence times, as well as various architecture improvements. We use our trained networks to analyze all events in the first Gravitational-Wave Transient Catalog (GWTC-1) [8] with component masses greater than 10 $M_\odot$ (our prior bound) and find close (sometimes indistinguishable) quantitative agreement with standard algorithms. This Letter sets a new standard for rapid gravitational wave inference, which should enable real-time gravitational wave science in the near future. It shows that NPE has moved beyond toy models and is competitive with conventional algorithms. More broadly, it provides a demonstration of these new methods in a realistic use case, which we hope will inspire wider adoption in experimental science.

Method.—The central object of DINGO is the density-estimation neural network, which defines a conditional probability distribution $q(\theta|d)$. This should be distinguished from the posterior $p(\theta|d)$, which $q(\theta|d)$ learns to approximate through training. We use so-called normalizing flows [43–45] to define a sufficiently flexible $q(\theta|d)$ via a d-dependent mapping $f_d : u \mapsto \theta$ from a simple "base" distribution $\pi(u)$,

$$q(\theta|d) = \pi[f_d^{-1}(\theta)] |\det J_{f_d^{-1}}|. \tag{2}$$

If $\pi(u)$ can be rapidly evaluated and sampled from, and if f_d is invertible and has a simple Jacobian determinant, then $q(\theta|d)$ can also be rapidly evaluated and sampled from. Following [36], we take $\pi(u)$ to be multivariate standard normal and f_d as a composition of spline coupling flows [46], each of which is defined with a neural network.

The overall structure of DINGO is illustrated in Fig. 1. This contains three key enhancements compared to the

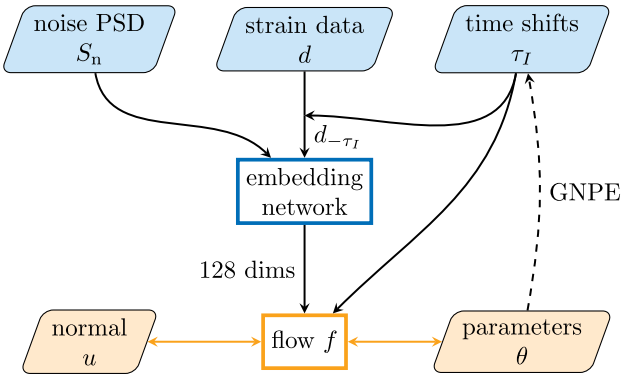

FIG. 1. DINGO flow chart. The posterior distribution is represented in terms of an invertible normalizing flow (orange), taking normally distributed random variables u into posterior samples θ. The flow itself depends on a (compressed) representation of the noise properties S_n and the data d, as well as an estimate τ_I of the coalescence time in each detector I. The data are time shifted by τ_I to simplify the representation. For inference, the iterative GNPE algorithm is used to provide an estimate of τ_I, as described in the main text.

study [36]. First, since the data generation process depends on the detector noise PSD S_n, we include this as additional context to the neural network, i.e., $q(\theta|d, S_n)$. This allows us to tune the network at inference time to the PSD estimated just prior to the event, corresponding to standard "off-source" noise estimation [17]. An alternative would be to estimate the noise "on-source" [47], but since we consider only short-duration BBH events here, the off-source approach is sufficient.

The second enhancement addresses the problem of high-dimensional observed data by using an additional neural network to first compress to a small number of features. This network (called an "embedding network") is trained alongside the flow network. Our data are in the frequency domain, between 20 and 1024 Hz, with 0.125 Hz resolution; so, combined with the PSDs, this gives 24 096 input dimensions for each of the two or three interferometers. The first stage of the embedding network maps this linearly to 400 components per detector. To provide an inductive bias to extract signal information, we seed this layer with the principal components of clean waveforms from our training set and then allow these parameters to float during training. Following this, a fully connected residual network [48] compresses to 128 features, which are provided to the flow.

Finally, we developed a new method to treat time translations of the strain data. For standard algorithms, inference of (α, δ, t_c) requires sampling over waveforms with varying coalescence times t_I in each detector I. Likewise, for NPE, the network must learn to interpret strain data with different t_I. For frequency-domain data, however, time translations correspond to local phase shifts, which, although explicitly known, are challenging for neural networks to learn. Indeed, this occupied much of the network capacity in Ref. [36]. Our new approach—called group equivariant neural posterior estimation (GNPE)—leverages explicit knowledge of the time-translation symmetry along with *approximate* knowledge of t_I to simplify the data representation and allow the network to focus on more nontrivial parameters. For further details see [49].

For GNPE, we train the network to infer θ given perturbed coalescence times τ_I and manually time-shifted strain data $d_{-\tau_I}$. Using maximum likelihood estimation [50], this means we minimize the loss function

$$L = \mathbb{E}_{p(\theta)}\mathbb{E}_{p(S_n)}\mathbb{E}_{p(d|\theta,S_n)}\mathbb{E}_{\kappa(\delta t_I)} [-\log q(\theta|d_{-t_I(\theta)-\delta t_I}, S_n, t_I(\theta) + \delta t_I)], \quad (3)$$

with respect to the network parameters [51]. Here, $\mathbb{E}$ refers to the expected value over the specified distribution, which is evaluated stochastically using Monte Carlo draws. $\kappa(\delta t_I)$ is a uniform kernel used to perturb t_I. For inference, even though we do not have direct access to t_I, all parameters can be inferred using Gibbs sampling, starting with an approximate t_I (obtained, e.g., using standard NPE): first, convolve t_I with $\kappa(\delta t_I)$ to obtain τ_I, then use the network to infer a new estimate for t_I; then convolve again and repeat. We find that this converges after $O(10)$ iterations.

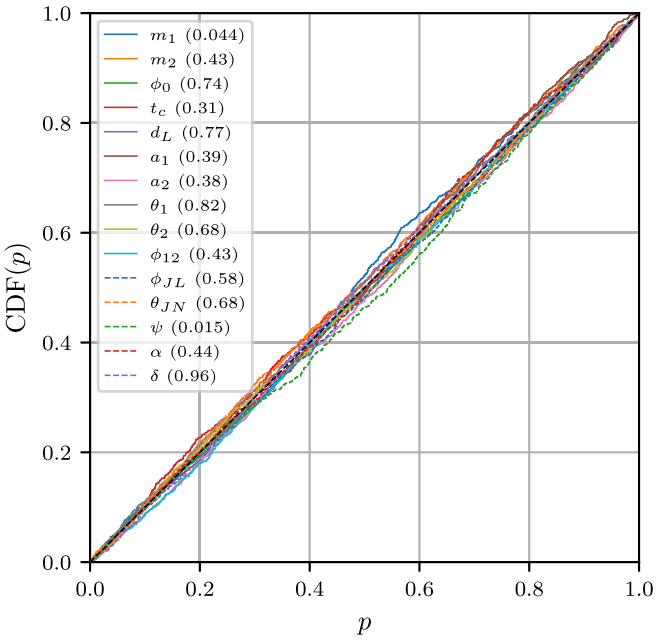

FIG. 2. *P–P* plot for 1000 injections. The legend shows the p values of the individual parameters, with a combined p value of 0.46.

Evaluating (3) requires sampling $\theta^{(i)} \sim p(\theta)$ and $S_n^{(i)} \sim p(S_n)$ and then simulating data $d^{(i)} \sim p(d|\theta^{(i)}, S_n^{(i)})$. Aside from the PSD sampling, this follows Ref. [36] very closely. In particular, we use the same prior over parameters, with $m_1, m_2 \in [10, 80]\ M_\odot$. We train separate networks for the noise distributions in the first (O1) and second (O2) observing runs of LIGO and Virgo, with PSD samples estimated empirically from stretches of interferometer noise data [52]. For O1, we choose the distance prior [100, 2000] Mpc. For O2, we train one network for loud events with distance prior [100, 2000] Mpc and another for quieter events with [100, 6000] Mpc. In addition to these two-detector networks, we train a three-detector network with distance prior [100, 1000] Mpc to analyze GW170814. With future enhancements of network architecture we expect to cover the entire distance range with a single network. Finally, as in Ref. [36], training data are generated from a fixed set of spin-precessing, frequency-domain waveforms, described by the IMRPhenomPv2 [16,53,54] model, but with extrinsic parameters and noise realizations drawn randomly during training. With training sets of 5×10^6 waveforms, there is no indication of overfitting. Training takes roughly 10 days on a single NVIDIA A100. Further details on the networks and training are provided in the Supplemental Material [55].

Results.—As a first test, we evaluate DINGO on data entirely consistent with the training distribution, i.e.,

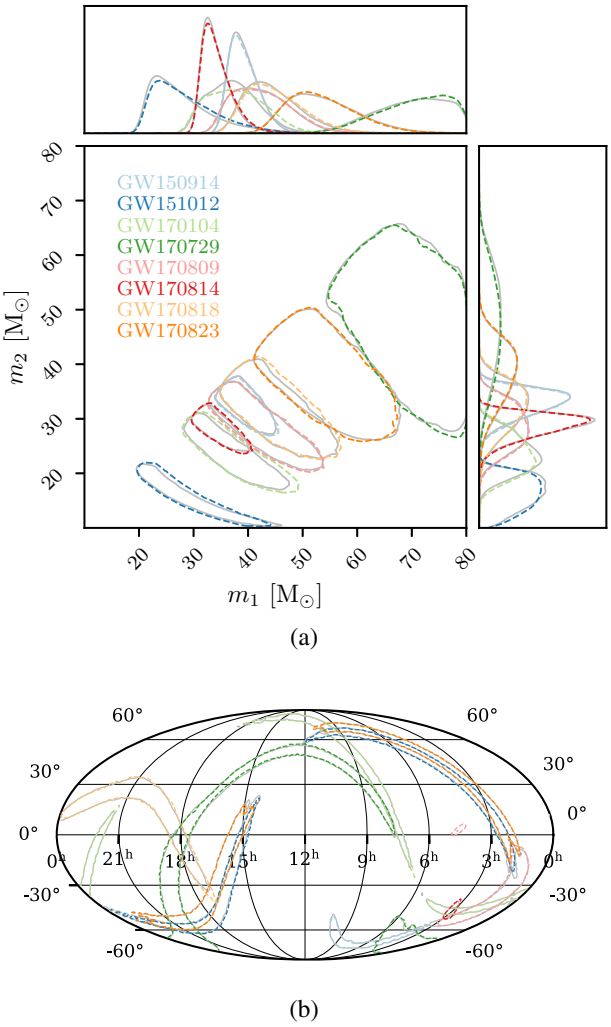

FIG. 3. Comparison of (a) detector-frame component mass and (b) sky position posteriors from DINGO (colored) and LALINFERENCE (gray) for eight GWTC-1 events. 90% credible regions shown.

simulated waveforms in stationary-Gaussian noise. This is an easier task than using observational data, which include real signals in noise that is neither strictly stationary nor Gaussian and therefore lie outside the training distribution. We sample posteriors from 1000 simulated datasets and construct a *P–P* plot (see Fig. 2). For each parameter, we compute the percentile score of the true value within its marginalized posterior, and then we plot the cumulative distribution function (CDF) of these scores. For true posteriors, the percentiles should be uniformly distributed, so the CDF should be diagonal. Kolmogorov-Smirnov test p values are indicated in the legend, with combined p value of 0.46. This shows that DINGO is performing properly on simulated data.

We now proceed to our main result, which is a demonstration of performance on real events. We perform inference on the eight GWTC-1 BBH events compatible with our prior, using both DINGO and LALINFERENCE MCMC. For DINGO, generation of 50 000 sample points with 30 GNPE iterations takes roughly 20 s. Comparisons of inferred component masses and sky position for all events show good agreement (see Fig. 3), including multimodality for the sky position. The one exception is GW170104, where the mass posterior is slightly flatter. Nevertheless, 90% credible intervals are in good agreement.

For quantitative comparisons, we compute the Jensen-Shannon divergence (JSD) [56] between DINGO and LALINFERENCE one-dimensional marginalized posteriors (see Fig. 4). This is a symmetric divergence that measures the difference between two probability distributions, with values ranging from 0 to $\ln(2) \approx 0.69$ nat. We find a mean JSD across all events and parameters of 0.0009 nat, which is slightly higher than the variation (0.0007 nat) found between LALINFERENCE runs with identical settings but different random seeds [19]. By comparing such LALINFERENCE runs, Ref. [19] also established a maximum JSD of 0.002 nat for indistinguishability; our results are approaching this

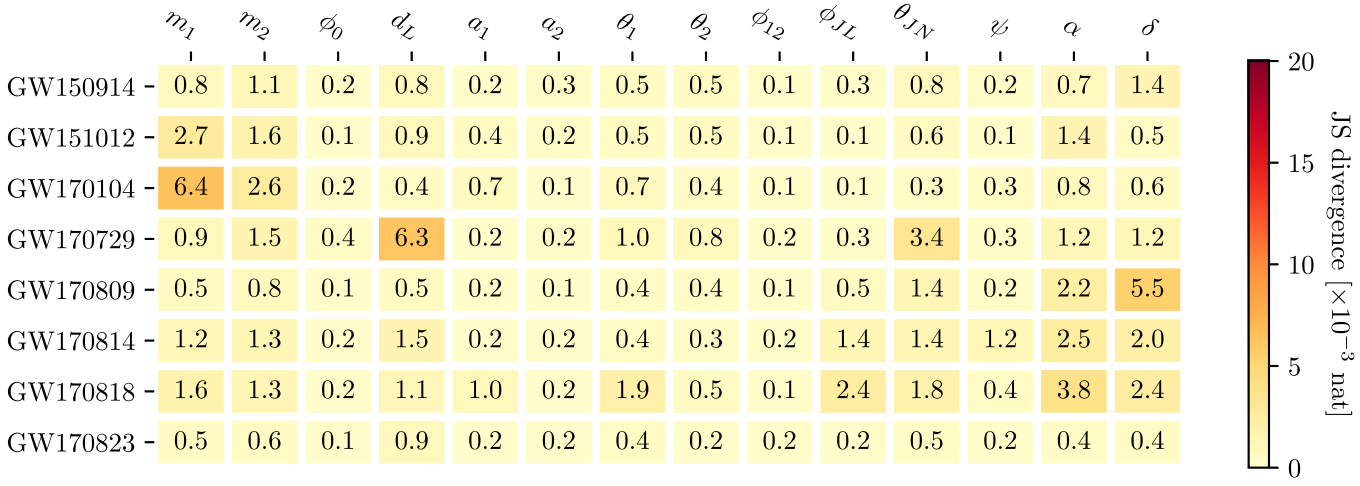

FIG. 4. JSDs between DINGO and LALINFERENCE marginalized posteriors, averaged over 100 realizations. The mean JSD across all events and parameters is 0.0009 nat.

threshold, with two events below for all parameters and the others with 1–3 parameters above. The slight visible disagreement between mass posteriors for GW170104 is also reflected in larger JSDs. For comparison, we note that PSD variations (see Supplemental Material [55]) and the choice of waveform model [19] both impact the JSD at a much higher level (0.02 nat). Additional comparisons between samplers, including posteriors for all events, are provided in the Supplemental Material [55].

Conclusions.—In this Letter, we introduced DINGO and applied it to perform extremely fast Bayesian parameter inference for gravitational waves observed by the LIGO and Virgo detectors. We analyzed eight GWTC-1 events and showed excellent agreement with standard algorithms, with inference times reduced by factors of $10^3 - 10^4$. This was achieved by conditioning on the detector noise characteristics and making a number of architecture and algorithm improvements. We plan to release a public DINGO code in the very near future.

A critical component of DINGO is a new iterative algorithm—GNPE—to partially off-load the modeling of time translations from the neural network. Although convergence of GNPE may take 20 s, initial samples with slightly reduced accuracy can, however, be produced in just a few seconds by taking fewer iterations.

Going forward, the next steps are to extend the prior to include longer-duration binary neutron star signals [57] (for which rapid results are especially important to identify electromagnetic counterparts) and to extend to more physically realistic waveform models, which include higher multipole modes and more accurate spin-precession effects [21]. Long and complex waveforms are much more expensive for standard algorithms, so the relative improvement in performance should be even more significant. If successful, this would also enable the routine use of the most physically realistic waveforms, resulting in consistently reduced systematic errors. These extensions will likely require somewhat larger networks and improved data representation or compression [58].

Another natural extension would be to study signals without making the common stationary-Gaussian idealization for the detector noise during the training stage. For DINGO, performing inference with realistic noise is simply a matter of training with simulated signals injected into real noise realizations taken from detectors. Using real noise should lead to improved accuracy that is not possible using standard likelihood-based methods and would serve as an excellent demonstration of the advantages of NPE. For real-time analysis, it will also be necessary to develop approaches to progressively retrain networks to keep pace with changing data distributions during an observing run, e.g., as detector sensitivity is improved. All of these enhancements, particularly the treatment of nonstationary noise, will be critical for extensions to future observatories such as the Laser Interferometer Space Antenna.

Deep-learning tools are now ready to analyze the vast majority of LIGO and Virgo events. In the past, the primary challenge has been in obtaining sufficiently accurate results, but with DINGO, we have now achieved this in a realistic context. Through planned future extensions, we expect that DINGO could become one of the leading approaches to gravitational wave inference.

We thank S. Ossokine, M. Pürrer, C. Simpson, and P. Züge for helpful discussions. This research has made use of data, software, and/or web tools obtained from the Gravitational Wave Open Science Center [62], a service of LIGO Laboratory, the LIGO Scientific Collaboration, and the Virgo Collaboration. LIGO Laboratory and Advanced LIGO are funded by the U.S. 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, and Spain. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. 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. We use PYTORCH [63] and NFLOWS [64] for the implementation of our neural networks. The plots are generated with MATPLOTLIB [65], CHAINCONSUMER [66], and LIGO.SKYMAP [67].

*maximilian.dax@tuebingen.mpg.de
†stephen.green@aei.mpg.de
‡jonathan.gair@aei.mpg.de

[1] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Observation of Gravitational Waves from a Binary Black Hole Merger, Phys. Rev. Lett. **116**, 061102 (2016).
[2] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Tests of general relativity with binary black holes from the second LIGO-Virgo Gravitational-Wave Transient Catalog, Phys. Rev. D **103**, 122002 (2021).
[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] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), A Gravitational-wave measurement of the hubble constant following the second observing run of Advanced LIGO and Virgo, Astrophys. J. **909**, 218 (2021).
[5] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Population properties of compact objects from the second

LIGO-Virgo Gravitational-Wave Transient Catalog, Astrophys. J. Lett. **913**, L7 (2021).
[6] J. Aasi *et al.* (LIGO Scientific Collaboration), Advanced LIGO, Classical Quantum Gravity **32**, 074001 (2015).
[7] F. Acernese *et al.* (VIRGO Collaboration), Advanced Virgo: A second-generation interferometric gravitational wave detector, Classical Quantum Gravity **32**, 024001 (2015).
[8] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs, Phys. Rev. X **9**, 031040 (2019).
[9] 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).
[10] B. P. Abbott *et al.* (KAGRA, LIGO Scientific, and Virgo Collaborations), Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA, Living Rev. Relativity **23**, 3 (2020).
[11] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), A guide to LIGO-Virgo detector noise and extraction of transient gravitational-wave signals, Classical Quantum Gravity **37**, 055002 (2020).
[12] A. Buonanno and T. Damour, Effective one-body approach to general relativistic two-body dynamics, Phys. Rev. D **59**, 084006 (1999).
[13] A. Bohé *et al.*, Improved effective-one-body model of spinning, nonprecessing binary black holes for the era of gravitational-wave astrophysics with advanced detectors, Phys. Rev. D **95**, 044028 (2017).
[14] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, L. E. Kidder, and H. P. Pfeiffer, Surrogate model of hybridized numerical relativity binary black hole waveforms, Phys. Rev. D **99**, 064045 (2019).
[15] M. Pürrer, Frequency domain reduced order model of aligned-spin effective-one-body waveforms with generic mass-ratios and spins, Phys. Rev. D **93**, 064041 (2016).
[16] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez 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).
[17] J. Veitch *et al.*, Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library, Phys. Rev. D **91**, 042003 (2015).
[18] G. Ashton *et al.*, BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy, Astrophys. J. Suppl. Ser. **241**, 27 (2019).
[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. R. Astron. Soc. **499**, 3295 (2020).
[20] J. S. Speagle, dynesty: A dynamic nested sampling package for estimating Bayesian posteriors and evidences, Mon. Not. R. Astron. Soc. **493**, 3132 (2020).
[21] S. Ossokine *et al.*, Multipolar effective-one-body waveforms for precessing binary black holes: Construction and validation, Phys. Rev. D **102**, 044055 (2020).
[22] B. P. Abbott *et al.* (LIGO Scientific and Virgo, Fermi GBM, INTEGRAL, IceCube, AstroSat Cadmium Zinc Telluride Imager Team, IPN, Insight-Hxmt, ANTARES, Swift, AGILE Team, 1M2H Team, Dark Energy Camera GW-EM, DES, DLT40, GRAWITA, Fermi-LAT, ATCA, ASKAP, Las Cumbres Observatory Group, OzGrav, DWF (Deeper Wider Faster Program), AST3, CAASTRO, VINROUGE, MASTER, J-GEM, GROWTH, JAGWAR, CaltechNRAO, TTU-NRAO, NuSTAR, Pan-STARRS, MAXI Team, TZAC Consortium, KU, Nordic Optical Telescope, ePESSTO, GROND, Texas Tech University, SALT Group, TOROS, BOOTES, MWA, CALET, IKI-GW Follow-up, H.E.S.S., LOFAR, LWA, HAWC, Pierre Auger, ALMA, Euro VLBI Team, Pi of Sky, Chandra Team at McGill University, DFN, ATLAS Telescopes, High Time Resolution Universe Survey, RIMAS, RATIR, and SKA South Africa/MeerKAT Collaborations), Multi-messenger observations of a binary neutron star merger, Astrophys. J. Lett. **848**, L12 (2017).
[23] G. Pratten, S. Husa, C. Garcia-Quiros, M. Colleoni, A. Ramos-Buades, H. Estelles, and R. Jaume, Setting the cornerstone for a family of models for gravitational waves from compact binaries: The dominant harmonic for nonprecessing quasi-circular black holes, Phys. Rev. D **102**, 064001 (2020).
[24] 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).
[25] L. P. Singer and L. R. Price, Rapid Bayesian position reconstruction for gravitational-wave transients, Phys. Rev. D **93**, 024013 (2016).
[26] J. Lange, R. O'Shaughnessy, and M. Rizzo, Rapid and accurate parameter inference for coalescing, precessing compact binaries, arXiv:1805.10457.
[27] N. J. Cornish, Rapid and robust parameter inference for binary mergers, Phys. Rev. D **103**, 104057 (2021).
[28] 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), pp. 1028–1036.
[29] J.-M. Lueckmann, P. J. Gonçalves, G. Bassetto, K. Öcal, M. Nonnenmacher, and J. H. Macke, Flexible statistical inference for mechanistic models of neural dynamics, in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates, Inc., New York, 2017), pp. 1289–1299.
[30] D. Greenberg, M. Nonnenmacher, and J. Macke, Automatic posterior transformation for likelihood-free inference, in *International Conference on Machine Learning* (PMLR, Long Beach, 2019), pp. 2404–2414.
[31] K. Cranmer, J. Brehmer, and G. Louppe, The frontier of simulation-based inference, Proc. Natl. Acad. Sci. U.S.A. **117**, 30055 (2020).
[32] H. Gabbard, C. Messenger, I. Siong Heng, F. Tonolini, and R. Murray-Smith, Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy, arXiv:1909.06296.
[33] A. J. K. Chua and M. Vallisneri, Learning Bayesian Posteriors with Neural Networks for Gravitational-Wave Inference, Phys. Rev. Lett. **124**, 041102 (2020).
[34] C. Chatterjee, L. Wen, K. Vinsen, M. Kovalam, and A. Datta, Using deep learning to localize gravitational wave sources, Phys. Rev. D **100**, 103025 (2019).

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

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

[37] 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, arXiv:2010.12931.

[38] P. G. Krastev, K. Gill, V. Ashley Villar, and E. Berger, Detection and parameter estimation of gravitational waves from binary neutron-star mergers in real LIGO data using deep learning, Phys. Lett. B **815**, 136161 (2021).

[39] H. Shen, E. A. Huerta, E. O'Shea, P. Kumar, and Z. Zhao, Statistically-informed deep learning for gravitational wave parameter estimation, arXiv:1903.01998v3.

[40] 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).

[41] Parameters consist of detector-frame component masses (m_1, m_2), time of coalescence at geocenter t_c, reference phase ϕ_c, sky position (α, δ), luminosity distance d_L, inclination angle θ_{JN}, spin magnitudes (a_1, a_2), spin angles $(\theta_1, \theta_2, \phi_{12}, \phi_{JL})$ [42], and polarization angle ψ.

[42] B. Farr, E. Ochsner, W. M. Farr, and R. O'Shaughnessy, A more effective coordinate system for parameter estimation of precessing compact binaries from gravitational waves, Phys. Rev. D **90**, 024018 (2014).

[43] D. Rezende and S. Mohamed, Variational inference with normalizing flows, in *International Conference on Machine Learning* (PMLR, Lille, 2015), pp. 1530–1538 [arXiv:1505.05770].

[44] 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.

[45] 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.

[46] 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.

[47] T. B. Littenberg and N. J. Cornish, Bayesian inference for spectral estimation of gravitational wave detector noise, Phys. Rev. D **91**, 084034 (2015).

[48] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, arXiv:1512.03385.

[49] M. Dax, S. R. Green, J. Gair, M. Deistler, B. Schölkopf, and J. H. Macke, Group equivariant neural posterior estimation, arXiv:2111.13139.

[50] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016), http://www.deeplearningbook.org.

[51] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.

[52] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Open data from the first and second observing runs of Advanced LIGO and Advanced Virgo, SoftwareX **13**, 100658 (2021).

[53] 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).

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

[55] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.127.241103 for additional figures and details regarding the architecture and training of the neural network.

[56] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inf. Theory **37**, 145 (1991).

[57] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral, Phys. Rev. Lett. **119**, 161101 (2017).

[58] Initial estimates based on a singular value decomposition [59] indicate that to accurately represent SEOBNRv4PHM BBH waveforms [21], the initial layers of our embedding network should be widened by a factor of roughly 4. Assuming the same number of iterations and fixed hardware, the total training time would increase by about 35%. For binary neutron stars, adopting a frequency-dependent resolution [60,61] would limit the expansion of the number of frequency bins to a factor of 3.

[59] K. Cannon, A. Chapman, C. Hanna, D. Keppel, A. C. Searle, and A. J. Weinstein, Singular value decomposition applied to compact binary coalescence gravitational-wave signals, Phys. Rev. D **82**, 044025 (2010).

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

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

[62] https://www.gw-openscience.org/

[63] 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., New York, 2019), pp. 8024–8035.

[64] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, NFLOWS: Normalizing flows in PyTorch (2020).

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

[66] S. R. Hinton, ChainConsumer, J. Open Source Software **1**, 00045 (2016).

[67] Leo Singer, LIGO.SKYMAP https://lscsoft.docs.ligo.org/ligo.skymap/ (2020).