

Flexible Gravitational-Wave Parameter Estimation with Transformers

Annalena Kofler^{1,2,*}, Maximilian Dax^{1,3,4}, Stephen R. Green⁵, Jonas Wildberger³, Nihar Gupte^{2,6},
Jakob H. Macke^{1,7}, Jonathan Gair², Alessandra Buonanno^{2,6} and Bernhard Schölkopf^{1,3}

¹Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany

²Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Am Mühlenberg 1, 14476 Potsdam, Germany

³ELLIS Institute Tübingen, Maria-von-Linden-Straße 2, 72076 Tübingen, Germany

⁴Tübingen AI Center, Maria-von-Linden-Straße 1, 72076 Tübingen, Germany

⁵Nottingham Centre of Gravity and School of Mathematical Sciences, University of Nottingham,
University Park, Nottingham, NG7 2RD, United Kingdom

⁶Department of Physics, University of Maryland, College Park, Maryland 20742, USA

⁷Machine Learning in Science, University of Tübingen and Tübingen AI Center, 72076 Tübingen, Germany



(Received 12 December 2025; accepted 12 May 2026; published 23 June 2026)

Gravitational-wave data analysis relies on accurate and efficient methods to extract physical information from noisy detector signals, yet the increasing rate and complexity of observations represent a growing challenge. Deep learning provides a powerful alternative to traditional inference, but existing neural models typically lack the flexibility to handle variations in data analysis settings. Such variations accommodate imperfect observations or are required for specialized tests, and could include changes in detector configurations, overall frequency ranges, or localized cuts. We introduce a flexible transformer-based architecture paired with a training strategy that enables adaptation to diverse analysis settings at inference time. Applied to parameter estimation, we demonstrate that a *single* flexible model—called DINGO-T1—can (i) analyze 48 binary black holes from the third LIGO-Virgo-KAGRA Observing Run under a wide range of analysis configurations, (ii) enable systematic studies of how detector and frequency configurations impact inferred posteriors, and (iii) perform inspiral-merger-ringdown consistency tests probing general relativity. DINGO-T1 also improves median sample efficiency on real events from a baseline of 1.4% to 4.2%. Our approach thus demonstrates flexible and scalable inference with a principled framework for handling missing or incomplete data—key capabilities for current and next-generation observatories.

DOI: [10.1103/n3ld-b4wv](https://doi.org/10.1103/n3ld-b4wv)

Introduction—Gravitational waves (GWs) provide unique access to the strong-field dynamics of compact objects and have transformed our ability to probe astrophysics, cosmology, and fundamental physics [1–8]. Drawing scientific conclusions from GW observations requires accurate and efficient data analysis, yet conventional Bayesian techniques [9,10] are computationally expensive and difficult to scale to the growing volume and complexity of data. Machine learning offers a promising solution by leveraging neural networks to process strain data and predict physical quantities, and has seen applications to detection [11–13], noise classification [14], source localization [15,16], and parameter estimation (PE) [17–27]. For PE, neural simulation-based inference (SBI,

[28]) has reduced analysis times for compact binary coalescences to the order of a few seconds—crucial for handling increased event rates [29,30] and enabling rapid multimessenger follow-up [1].

This speedup arises from front-loading computational costs: once trained on millions of simulated datasets, neural networks can analyze new observations at negligible expense. However, conventional neural architectures operate on a fixed input dimension, requiring the data representation to be fixed prior to training. A trained model is therefore tied to specific detector combinations and data-conditioning settings—a major barrier to widespread adoption. Existing models cannot readily adapt to detectors being offline, to restricted frequency bands used to mitigate noise contamination, or to data gaps expected in LISA observations [31].

In this Letter, we propose a neural architecture and training strategy that enable flexible analysis of heterogeneous GW data. The key idea is to process strain data using a transformer encoder and exploit its ability to handle input sequences of arbitrary length [32]. We partition datasets into sequences of short strain segments called

*Contact author: annalena.kofler@tuebingen.mpg.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Open access publication funded by the Max Planck Society.

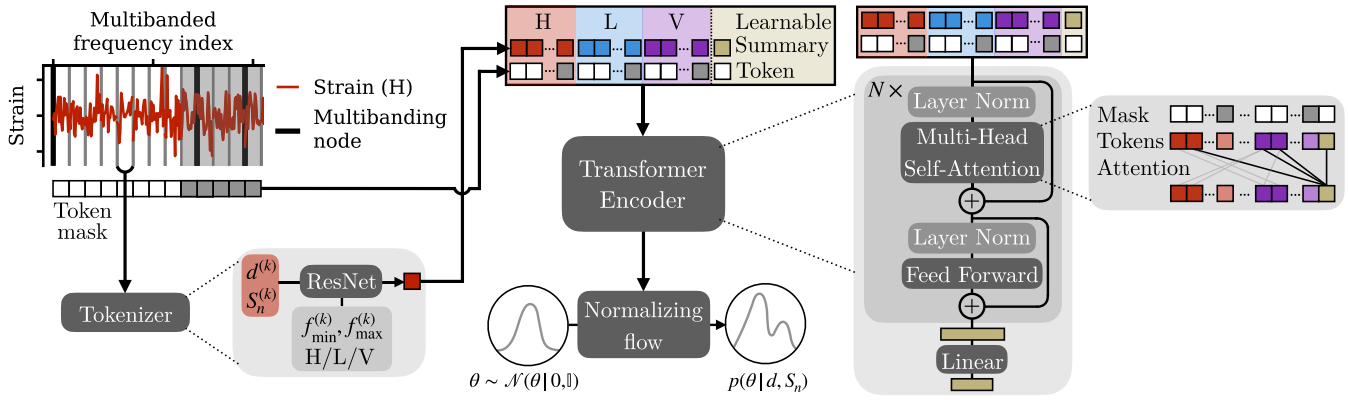


FIG. 1. Overview of DINGO-T1. After multibanding, data and detector PSD segments ($d^{(k)}, S_n^{(k)}$) are mapped by a shared tokenizer into tokens that also encode the frequency range ($f^{(k)}, f^{(k+1)}$) and detector identity I . The sequence of tokens is masked, augmented with a learnable summary token, and processed by the transformer encoder. Through self-attention, the summary token aggregates information from all unmasked segments. The final summary token is projected via a linear layer to a 128-dimensional feature vector, which conditions a normalizing flow to model the posterior $p(\theta|d, S_n)$ over source parameters.

tokens, and randomly drop tokens during training to mimic frequency cuts and missing detectors. Through the self-attention mechanism, the model learns to identify correlations between strain segments, capturing global dependencies across detectors and frequencies, and to marginalize over missing data. Although transformers have been used for a range of GW tasks [33–39], previous work has not leveraged their inherent flexibility in handling variable-length or incomplete data. Our approach is inspired by works on missing data in SBI [40–43].

We integrate the transformer encoder into the DINGO framework for GW PE, yielding a flexible model that we call DINGO-T1 (DINGO Transformer, version 1; see Fig. 1). We demonstrate its versatility and performance through injection studies and analyses of real data. In particular, with a single network, we reanalyze 48 out of 84 binary black holes (BBHs) from the LIGO-Virgo-KAGRA (LVK) [44–46] third observing run (O3) [29,30,47,48] that are consistent with our prior (generic spins, component masses $m_1, m_s \in [10, 120]M_\odot$, and luminosity distance $d_L \in [0.1, 6]$ Gpc). Together, these events span 17 combinations of detectors and frequency ranges. Compared to baseline neural posterior estimation (NPE) networks with residual network encoders [22], DINGO-T1 improves median importance sampling efficiency from 1.4% to 4.2%. We further analyze all events using all possible LIGO-Virgo detector combinations. Such studies are extremely costly for standard methods, but can be used to identify data systematics. As a final demonstration, we use frequency cuts to perform inspiral-merger-ringdown consistency tests on seven events where DINGO-T1 yields reliable posterior estimates. Previously, conducting all of these experiments would have required training 94 different models to handle the various detector combinations, and frequency ranges for events as well as tests of general relativity.

Our Letter thus represents a step towards generalized inference models in GW astronomy, enabling flexible choice of detector and frequency settings without retraining. This flexibility opens new prospects for analyzing large catalogs [29,30], facilitating rapid assessments of systematics and consistency across diverse data-analysis settings [49,50], as well as more robust approaches to missing data. In addition, our transformer model DINGO-T1 [51] serves as a powerful and general encoder for GW data which could be fine-tuned on other tasks such as GW detection, providing a foundation for efficient and scalable GW science.

Architecture—For NPE [55–57], one trains a neural network $q(\theta|d)$ to estimate the Bayesian posterior $p(\theta|d)$ over parameters θ given data d . In the context of GW single-event inference, θ denotes the source parameters, e.g., of a black hole binary, and d denotes the strain data measured at the detectors [58]. For DINGO, we also augment the strain data with an estimate of the noise power spectral density (PSD), which varies from event to event [22]. Data are generally of high dimension ($\gg 10^4$), so d is first mapped by an embedding network f_ϕ to a lower-dimensional representation. This compressed summary is then passed to a density estimator $\hat{q}_\phi$ to model the posterior, $q(\theta|d) = \hat{q}_\phi(\theta|f_\phi(d))$. The learnable parameters $\{\phi, \varphi\}$ are optimized jointly during training.

Our core architectural change is to replace the traditional embedding network with a transformer encoder (Fig. 1), which naturally handles sequences of variable length. We divide the strain datasets into small segments, which a tokenizer network maps to token embeddings (Fig. 1, left). These embeddings form a sequence that the transformer encoder processes. We now describe the tokenization scheme, model architecture, and training strategy; full technical details appear in Supplemental Material [59].

We represent GW data $d = \{d_I\}_{I=H,L,V}$ as frequency series from the Hanford, Livingston, and Virgo (HLV) detectors covering $[20, 1810]$ Hz. We use a multibanded (nonuniform) frequency grid, with coarser resolution at higher frequencies, designed to capture the chirping morphology of GW signals, and achieve an initial compression of the data [26,60,61]. To tokenize, we partition the grid into K equal-length segments with boundary nodes $(f^{(k)})_{k=0}^{K+1}$, such that each interval $(f^{(k)}, f^{(k+1)})$ has a uniform frequency grid (Fig. 1, vertical gray lines). The data d_I and PSD $S_{n,I}$ are processed identically for each detector, yielding $3K$ segments $(d_I^{(k)}, S_{n,I}^{(k)})$, each containing 16 frequency bins.

Each data segment is converted into a token embedding $t(d_I^{(k)}, S_{n,I}^{(k)}, f^{(k)}, f^{(k+1)}, I)$ by a shared tokenizer network t (see Fig. 1, bottom left). Transformer-based architectures are invariant under token permutations, so conditioning tokens on frequency range and detector identity is necessary for interpreting each segment [62]. Finally, we append one additional learnable summary token to the sequence so that the model has a dedicated place to store relevant information extracted across tokens [63,64].

The full token sequence is passed to a transformer encoder [32,65,66], which extracts information relevant for parameter estimation. After the final transformer layer, we compress the summary token to a 128-dimensional feature vector, which conditions a normalizing flow with a rational-quadratic spline coupling transform [67], using the same hyperparameters as in [23].

Training—We train the DINGO-T1 model—the tokenizer, transformer encoder, and normalizing flow—end to end using a negative log likelihood loss. To enable inference under missing data, we mask tokens randomly during training [68]. This has the effect of dropping segments of data, so that the network learns to interpret inputs with missing data. By applying a token mask m to (d, S_n) , our loss takes the form

$$\mathcal{L} = \mathbb{E}_{p(\theta)p(S_n)p(d|\theta,S_n)p(m)} \{-\log q[\theta|m(d), m(S_n)]\}, \quad (1)$$

where $p(m)$ denotes the probability of sampling mask m .

We employ a data-based approach to masking, i.e., we choose $p(m)$ to reflect realistic variations in data analysis settings. This includes dropping all tokens corresponding to certain detectors, removing tokens to update minimum and maximum frequencies, and applying random cuts within the frequency range. We also compare this approach to a masking strategy where tokens are dropped uniformly at random, but find that the data-based approach gives better performance. (See Supplemental Material [59] for complete details on masking.)

We train on simulated data consisting of waveforms generated from the IMRPhenomXPHM model [69] with additive stationary Gaussian noise. Our dataset includes

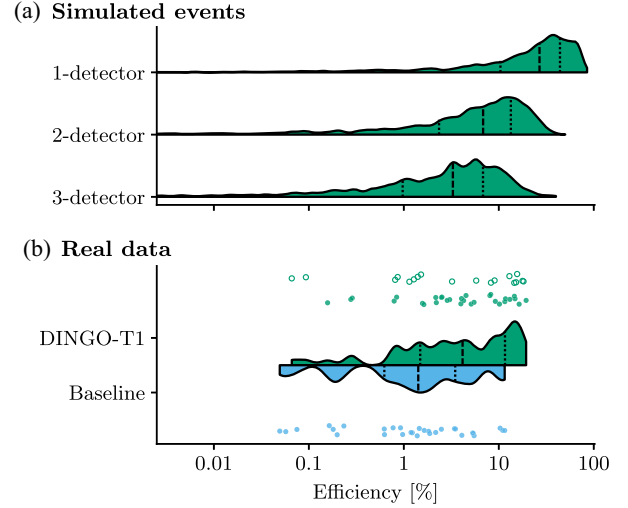


FIG. 2. (a) Sample efficiency distribution of 1000 simulated events per detector configuration for the DINGO-T1 model, shown as violin plots. (b) Sample efficiency distribution of DINGO-T1 and the DINGO baseline across 48 real GW events. Dots denote three-detector events, while circles refer to two-detector events which cannot be analyzed with the inflexible DINGO baseline. The violin plots summarize the distributions, while the points for the individual events make the limited sample size explicit. The dashed line marks the median and the dotted lines the quartiles. The efficiencies between simulated (a) and real data (b) are not directly comparable since the parameters are drawn from different distributions.

2.5×10^7 intrinsic waveforms, with extrinsic parameters generated on the fly during training. Priors are the same as [23], except for the luminosity distance, which we take as uniform over the range $[0.1, 6]$ Gpc; we infer all parameters except for the phase, which is reconstructed in postprocessing [22]. We generate noise during training based on a collection of Welch PSDs estimated across O3 [22]. Models are trained on 8 NVIDIA-A100 GPUs using distributed data parallelization [70,71] which takes approximately 9.5 days (183 epochs) for DINGO-T1.

At inference time, we only include tokens corresponding to available data that we wish to analyze, omitting those associated with missing information (e.g., a nonobserving detector or masked frequency ranges). To validate and correct the resulting posterior, we apply importance sampling (IS) [23], reweighting samples from $q(\theta|d, S_n)$ to the exact posterior defined by the ranges of the original uniform-frequency data. Because of variance in the distribution of importance weights, the effective number of weighted samples is smaller than the number of unweighted samples. This ratio (the sample efficiency) serves as a performance metric.

Results—We first evaluate DINGO-T1 on 1000 *simulated* GW signals for each subset of the three detectors HLV, with injections performed over the full frequency range. For each injection, we generate 10^5 posterior samples and

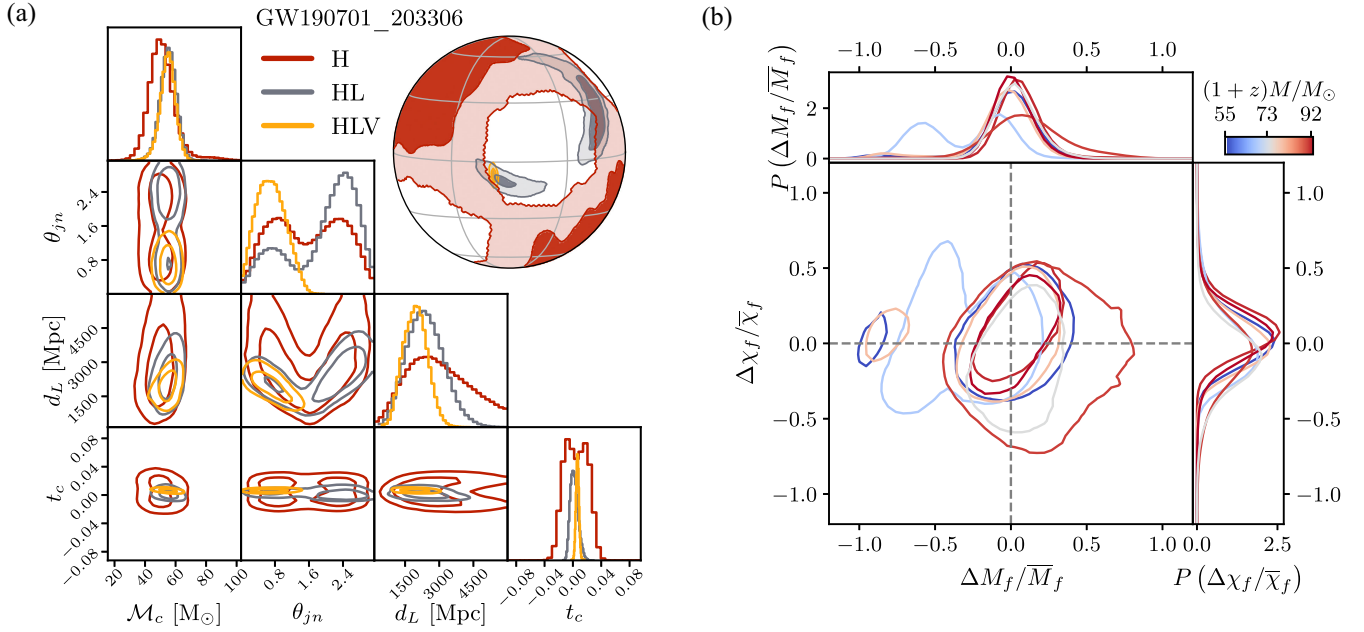


FIG. 3. (a) Posterior distribution for GW190701_203306, showing DINGO-T1 analyses for three different detector configurations. (b) Inspiral-merger-ringdown consistency tests for seven events where the signal was analyzed with the DINGO-T1 model on the inspiral and postinspiral part of the signal. The main panel shows the 90% credible regions of the 2D posteriors on $(\Delta M_f / \bar{M}_f, \Delta \chi_f / \bar{\chi}_f)$ and the side panels show the marginalized posteriors. The color corresponds to the median redshifted total mass obtained from posterior samples of the full signal.

perform IS. Initial sampling takes under five seconds, with phase reconstruction and IS requiring an additional five to ten minutes on 64 CPU cores. Median efficiencies are 26.9%, 6.8%, and 3.3% for one-, two-, and three-detector configurations, respectively [Fig. 2(a)]. Efficiencies decrease as more detectors are included, consistent with expectations: posteriors constrained by more detectors are narrower and therefore more challenging to learn for normalizing flows. To show that DINGO-T1 is well calibrated, we provide P-P plots for each detector configuration and additional information in Supplemental Material [59].

Second, we evaluate DINGO-T1 on all BBHs from O3 that fall within our prior, employing the specific data analysis settings from the official catalogs [29,30]. For each event, we obtain posterior samples as above and summarize sample efficiencies in Fig. 2(b). Across the 48 events, we find a median efficiency of 4.2%; two-detector events perform slightly better at 4.5%, while three-detector events reach 4.2%. This outperforms a baseline DINGO NPE model trained for fixed data analysis settings, which obtains a median efficiency of 1.4% across 30 HLV events evaluated on the full frequency range. In general, sample efficiencies of real data (in contrast to injections) can be negatively impacted by misspecification of the signal and noise model. Excluding such events increases the median sample efficiency from 4.2% to 5.2% for all events (see Supplemental Material [59]).

It should be noted that our results are not directly comparable to past work [23] by some of us, which

reported $\sim 10\%$ median efficiency on O3 events. That study performed HL analyses only; when restricting to HL detectors we find 11% median efficiency with DINGO-T1 (see Supplemental Material [59]). Reference [23] also incorrectly accounted for the window factor in the likelihood, likely resulting in lower efficiencies [54]. More importantly, Ref. [23] used *group-equivariant* NPE (GNPE, [22,72]), an algorithm which simplifies data representations by iteratively aligning merger times in each detector. GNPE improves accuracy, but results in substantially longer analysis times as it breaks access to the probability density. DINGO-T1 is based on standard NPE, making inference faster and simpler.

Applications—To explore the impact of detector configuration on the posterior, we evaluate every event on all possible combinations. While such studies are computationally infeasible with standard samplers, total inference times of under ten minutes per event make it a reasonable analysis with DINGO-T1. We illustrate the effect on the posterior distribution for an example event in Fig. 3(a). (See Supplemental Material [59] for detailed discussions.)

Finally, we utilize the flexibility of DINGO-T1 to perform inspiral-merger-ringdown (IMR) consistency tests probing general relativity [4,73,74]. For these tests, the posterior of the final mass M_f and spin χ_f are independently estimated for the inspiral (low) and postinspiral (high frequency) part of the signal, separated by an event-specific cut-off frequency f_{cut} . We perform IMR consistency tests on seven events from O3 where we obtain high enough efficiency to

draw 5000 posterior samples—with all analyses carried out using the same DINGO-T1 model. We display the fractional deviations between the inspiral and postinspiral parameters in Fig. 3(b) and find that our analyses show overall consistency with general relativity.

Conclusions—We introduced DINGO-T1, a transformer-based model that enables fast and flexible GW parameter inference across varying data analysis settings, including detector configurations and frequency ranges. We validated DINGO-T1 on an array of simulated and real datasets, including a realistic catalog analysis involving 17 data configurations. We argued that DINGO-T1 can be used to probe systematics, and demonstrated it can be applied for IMR consistency tests. Thanks to improved architecture and training, we obtain increased median sample efficiency on O3 HLV events of 4.2%, compared to 1.4% for a NPE baseline. As we do not leverage knowledge of time-translation equivariances in DINGO-T1, the efficiencies do not fully reach GNPE-level performance. However, DINGO-T1 has greatly increased flexibility and maintains access to the density, making it significantly more useful in practice and enabling further scaling and fine-tuning studies.

A natural next step would be to consider variations in signal duration by conditioning tokens on frequency resolution information; however, this would require substantial modifications to the current training pipeline and is therefore left to future work. Another promising direction is to explore alternative data representations such as time domain or time-frequency domain [75] (employing vision transformers [76]), where simpler signal morphology may lead to improved training. In the context of LISA, our approach offers a natural solution to the problem of expected data gaps [31]. In general, the proposed transformer framework and training strategy can be readily extended and adapted to other tasks like signal detection and the DINGO-T1 transformer encoder could serve as a general-purpose GW compression backbone. By concentrating on adaptable models that can be shared across tasks, computational effort can thereby be invested where it most benefits scaling and performance.

This Letter has made use of many open-source Python packages, including Astropy [77], BayesWave [78–80], Bilby [9], Bilby_PIPE [81], CORNER [82], GLASFLOW [83], LALSuite [84,85], LIGO.SKYMAP [86], Matplotlib [87], NFWLWS [88], NumPy [89], PANDAS [90], PyCBC [91], PyTorch [92], SEABORN [93], and WANDB [94]. For training, we employ the AdamW [70] optimizer and automatic mixed precision [71]. The accessible color schemes in our figures are based on [95,96].

Acknowledgments—We thank Michael Pürrer and Vincent Berenz for helpful discussions, and Ik Siong Heng for comments on the manuscript. We thank the reviewers for their detailed questions and comments on

the draft. A. K. thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support. S. R. G. is supported by a UKRI Future Leaders Fellowship (Grant No. MR/Y018060/1). A. B.’s research is supported in part by the European Research Council (ERC) Horizon Synergy Grant “Making Sense of the Unexpected in the Gravitational-Wave Sky” Grant Agreement No. GWSky–101167314. The computational work for this Letter was performed on the Atlas cluster at the Max Planck Institute for Intelligent Systems. This research has made use of data or software obtained from the Gravitational Wave Open Science Center [97], a service of the LIGO Scientific Collaboration, the Virgo Collaboration, and KAGRA. This material is based upon work supported by NSF’s LIGO Laboratory which is a major facility fully funded by the National Science Foundation, 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. KAGRA is supported by Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan Society for the Promotion of Science (JSPS) in Japan; National Research Foundation (NRF) and Ministry of Science and ICT (MSIT) in Korea; Academia Sinica (AS) and National Science and Technology Council (NSTC) in Taiwan.

Data availability—The instructions to reproduce the findings of this article are openly available at [53] which includes the code, the settings files, as well as information about downloading and performing inference with the neural networks [52] which were used to generate the reported results.

-
- [1] B. P. Abbott *et al.* (LIGO Scientific, 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, SKA South Africa/MeerKAT Collaborations), Multi-messenger observations of a binary neutron star merger, *Astrophys. J. Lett.* **848**, L12 (2017).
- [2] 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).
 - [3] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), Population of merging compact binaries inferred using gravitational waves through GWTC-3, *Phys. Rev. X* **13**, 011048 (2023).
 - [4] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), Tests of general relativity with GWTC-3, *Phys. Rev. D* **112**, 084080 (2025).
 - [5] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), Constraints on the cosmic expansion history from GWTC-3, *Astrophys. J.* **949**, 76 (2023).
 - [6] A. G. Abac *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-4.0: Population properties of merging compact binaries, [arXiv:2508.18083](#).
 - [7] LIGO Scientific, Virgo, and KAGRA Collaborations, Black hole spectroscopy and tests of general relativity with GW250114, *Phys. Rev. Lett.* **136**, 041403 (2026).
 - [8] LIGO Scientific, Virgo, and KAGRA Collaborations, GWTC-4.0: Constraints on the cosmic expansion rate and modified gravitational-wave propagation, [arXiv:2509.04348](#).
 - [9] G. Ashton *et al.*, Bilby: A user-friendly Bayesian inference library for gravitational-wave astronomy, *Astrophys. J. Suppl. Ser.* **241**, 27 (2019).
 - [10] G. Ashton and C. Talbot, Bilby-MCMC: An MCMC sampler for gravitational-wave inference, *Mon. Not. R. Astron. Soc.* **507**, 2037 (2021).
 - [11] T. D. Gebhard, N. Kilbertus, I. Harry, and B. Schölkopf, Convolutional neural networks: A magic bullet for gravitational-wave detection?, *Phys. Rev. D* **100**, 063015 (2019).
 - [12] E. Marx *et al.*, Machine-learning pipeline for real-time detection of gravitational waves from compact binary coalescences, *Phys. Rev. D* **111**, 042010 (2025).
 - [13] A. E. Koloniari, E. C. Koursoumpa, P. Nousi, P. Lampropoulos, N. Passalis, A. Tefas, and N. Stergioulas, New gravitational wave discoveries enabled by machine learning, *Mach. Learn.* **6**, 015054 (2025).
 - [14] M. Razzano and E. Cuoco, Image-based deep learning for classification of noise transients in gravitational wave detectors, *Classical Quantum Gravity* **35**, 095016 (2018).
 - [15] C. Chatterjee, M. Kovalam, L. Wen, D. Beveridge, F. Diakogiannis, and K. Vinsen, Rapid localization of gravitational wave sources from compact binary coalescences using deep learning, *Astrophys. J.* **959**, 42 (2023).
 - [16] C. Chatterjee and L. Wen, Premerger sky localization of gravitational waves from binary neutron star mergers using deep learning, *Astrophys. J.* **959**, 76 (2023).
 - [17] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy, *Nat. Phys.* **18**, 112 (2022).
 - [18] A. J. K. Chua and M. Vallisneri, Learning Bayesian posteriors with neural networks for gravitational-wave inference, *Phys. Rev. Lett.* **124**, 041102 (2020).
 - [19] S. R. Green, C. Simpson, and J. Gair, Gravitational-wave parameter estimation with autoregressive neural network flows, *Phys. Rev. D* **102**, 104057 (2020).
 - [20] S. R. Green and J. Gair, Complete parameter inference for GW150914 using deep learning, *Mach. Learn.* **2**, 03LT01 (2021).
 - [21] A. Delaunoy, A. Wehenkel, T. Hinderer, S. Nissanke, C. Weniger, A. R. Williamson, and G. Louppe, Lightning-fast gravitational wave parameter inference through neural amortization, in *Third Workshop on Machine Learning and the Physical Sciences* (2020), [arXiv:2010.12931](#).
 - [22] 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).
 - [23] M. Dax, S. R. Green, J. Gair, M. Pürrer, J. Wildberger, J. H. Macke, A. Buonanno, and B. Schölkopf, Neural importance sampling for rapid and reliable gravitational-wave inference, *Phys. Rev. Lett.* **130**, 171403 (2023).
 - [24] U. Bhargwaj, J. Alvey, B. K. Miller, S. Nissanke, and C. Weniger, Sequential simulation-based inference for gravitational wave signals, *Phys. Rev. D* **108**, 042004 (2023).
 - [25] D. Chatterjee *et al.*, Rapid likelihood free inference of compact binary coalescences using accelerated hardware, *Mach. Learn.* **5**, 045030 (2024).
 - [26] M. Dax, S. R. Green, J. Gair, N. Gupte, M. Pürrer, V. Raymond, J. Wildberger, J. H. Macke, A. Buonanno, and B. Schölkopf, Real-time inference for binary neutron star mergers using machine learning, *Nature (London)* **639**, 49 (2025).
 - [27] W. van Straalen, A. Kolmus, J. Janquart, and C. Van Den Broeck, Pre-merger detection and characterization of inspiraling binary neutron stars derived from neural posterior estimation, [arXiv:2407.10263](#).
 - [28] K. Cranmer, J. Brehmer, and G. Louppe, The frontier of simulation-based inference, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30055 (2020).
 - [29] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run, *Phys. Rev. D* **109**, 022001 (2024).
 - [30] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run, *Phys. Rev. X* **13**, 041039 (2023).
 - [31] O. Burke, S. Marsat, J. R. Gair, and M. L. Katz, Addressing data gaps and assessing noise mismodeling in LISA, *Phys. Rev. D* **111**, 124053 (2025).
 - [32] A. Vaswani *et al.*, Attention is all you need, in *Advances in Neural Information Processing Systems*, edited by I. Guyon *et al.* (Curran Associates, Inc., Red Hook, NY, 2017), Vol. 30.
 - [33] L. Jiang and Y. Luo, Convolutional transformer for fast and accurate gravitational wave detection, in *2022 26th International Conference on Pattern Recognition (ICPR)* (Curran Associates, Inc., Red Hook, NY, 2022), pp. 46–53.

- [34] R. Shi, Y. Zhou, T. Zhao, Z. Cao, and Z. Ren, Compact binary systems waveform generation with a generative pretrained transformer, *Phys. Rev. D* **109**, 084017 (2024).
- [35] H. Wang, Y. Zhou, Z. Cao, Z. Guo, and Z. Ren, WaveFormer: Transformer-based denoising method for gravitational-wave data, *Mach. Learn.* **5**, 015046 (2024).
- [36] C. Chatterjee *et al.*, Pre-trained audio transformer as a foundational AI tool for gravitational waves, [arXiv:2412.20789](https://arxiv.org/abs/2412.20789).
- [37] E. Khalouei, C. G. Sabiu, L. M. Hyung, and A. Gopakumar, External attention transformer: A robust AI model for identifying initial eccentricity signatures in binary black hole events in simulated Advanced LIGO data, *J. Cosmol. Astropart. Phys.* **10** (2025) 028.
- [38] P. M. Joshi and R. Prix, Transformer networks for continuous gravitational-wave searches, *Phys. Rev. D* **112**, 123052 (2025).
- [39] L. Papalini, F. D. Santi, M. Razzano, I. S. Heng, and E. Cuoco, Can Transformers help us perform parameter estimation of overlapping signals in gravitational wave detectors?, *Classical Quantum Gravity* **42**, 185012 (2025).
- [40] S. N. Shukla and B. M. Marlin, A survey on principles, models and methods for learning from irregularly sampled time series, [arXiv:2012.00168](https://arxiv.org/abs/2012.00168).
- [41] Z. Wang, J. Hasenauer, and Y. Schälte, Missing data in amortized simulation-based neural posterior estimation, *PLoS Comput. Biol.* **20**, 1 (2024).
- [42] M. Gloeckler, M. Deistler, C. D. Weibach, F. Wood, and J. H. Macke, All-in-one simulation-based inference, in *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 235, edited by R. Salakhutdinov *et al.* (PMLR, USA, 2024), pp. 15735–15766.
- [43] Y. Verma, A. Bharti, and V. Garg, Robust simulation-based inference under missing data via neural processes, in *The Thirteenth International Conference on Learning Representations* (Curran Associates, Inc., Red Hook, NY, 2025).
- [44] J. Aasi *et al.* (LIGO Scientific Collaboration), Advanced LIGO, *Classical Quantum Gravity* **32**, 074001 (2015).
- [45] F. Acernese *et al.* (Virgo Collaboration), Advanced Virgo: A second-generation interferometric gravitational wave detector, *Classical Quantum Gravity* **32**, 024001 (2015).
- [46] T. Akutsu *et al.* (KAGRA Collaboration), Overview of KAGRA: Detector design and construction history, *Prog. Theor. Exp. Phys.* **2021**, 05A101 (2021).
- [47] The LIGO Scientific and the Virgo Collaborations, GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo During the first half of the third observing run—parameter estimation data release, Zenodo (2022), [10.5281/zenodo.6513631](https://zenodo.org/record/6513631).
- [48] The LIGO Scientific, the Virgo, and the KAGRA Collaborations, GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run—parameter estimation data release, Zenodo (2021), [10.5281/zenodo.8177023](https://zenodo.org/record/8177023).
- [49] E. Payne, S. Hourihane, J. Golomb, R. Udall, D. Davis, and K. Chatziioannou, Curious case of GW200129: Interplay between spin-precession inference and data-quality issues, *Phys. Rev. D* **106**, 104017 (2022).
- [50] R. Udall, S. Hourihane, S. Miller, D. Davis, K. Chatziioannou, M. Isi, and H. Deshong, Antialigned spin of GW191109: Glitch mitigation and its implications, *Phys. Rev. D* **111**, 024046 (2025).
- [51] Weights for DINGO-T1 and DINGO baseline [52], as well as corresponding code are publicly available [53]. The model was trained taking [54] into account.
- [52] A. Kofler, M. Dax, S. R. Green, J. Wildberger, N. Gupte, J. Macke, J. Gair, A. Buonanno, and B. Schölkopf, DINGO-T1 model for “Flexible gravitational-wave parameter estimation with transformers” (2025), [10.5281/zenodo.17726076](https://zenodo.org/record/17726076).
- [53] A. Kofler, M. Dax, S. R. Green, J. Wildberger, N. Gupte, J. Macke, J. Gair, A. Buonanno, and B. Schölkopf, DINGO-T1 (2025), <https://github.com/dingo-gw/dingo-T1>.
- [54] C. Talbot, S. Biscoveanu, A. Zimmerman, T. Baka, W. M. Farr, J. Golomb, C. Hoy, A. Lundgren, J. Tissino, M. J. Williams, J. Veitch, and A. Vijaykuma, Inference with finite time series II: The window strikes back, *Classical Quantum Gravity* **42**, 235023 (2025).
- [55] G. Papamakarios and I. Murray, Fast ϵ -free inference of simulation models with Bayesian conditional density estimation, in *Advances in Neural Information Processing Systems* (2016), [arXiv:1605.06376](https://arxiv.org/abs/1605.06376).
- [56] 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., Red Hook, NY, 2017), pp. 1289–1299.
- [57] D. Greenberg, M. Nonnenmacher, and J. Macke, Automatic posterior transformation for likelihood-free inference, in *International Conference on Machine Learning* (PMLR, USA, 2019), pp. 2404–2414.
- [58] 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).
- [59] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/n3ld-b4wv> for additional details and supporting results.
- [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] S. Morisaki, Accelerating parameter estimation of gravitational waves from compact binary coalescence using adaptive frequency resolutions, *Phys. Rev. D* **104**, 044062 (2021).
- [62] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, Language modeling with gated convolutional networks, in *Proceedings of the 34th International Conference on Machine Learning—Volume 70, ICML’17* (JMLR, USA, 2017), pp. 933–941.
- [63] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Long and Short Papers Vol. 1, edited by J. Burstein, C. Doran, and T. Solorio (Association for

- Computational Linguistics, Minneapolis, Minnesota, 2019), pp. 4171–4186.
- [64] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, Vision transformers need registers, in *The Twelfth International Conference on Learning Representations* (2024), <https://dblp.org/rec/conf/iclr/2024.html>.
 - [65] Q. Wang *et al.*, Learning deep transformer models for machine translation, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by A. Korhonen, D. Traum, and L. Márquez (Association for Computational Linguistics, Red Hook, NY, 2019), pp. 1810–1822.
 - [66] R. Xiong *et al.*, On layer normalization in the transformer architecture, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 119, edited by H. Daumé III and A. Singh (PMLR, 2020), pp. 10524–10533.
 - [67] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, in *Advances in Neural Information Processing Systems*, edited by H. Wallach *et al.* (Curran Associates, Inc., Red Hook, NY, 2019), Vol. 32.
 - [68] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, Masked-attention mask transformer for universal image segmentation, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE Computer Society, Los Alamitos, CA, USA, 2022), pp. 1280–1289.
 - [69] 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).
 - [70] I. Loshchilov and F. Hutter, Decoupled weight decay regularization, in *International Conference on Learning Representations* (Curran Associates, Inc., Red Hook, NY, 2019).
 - [71] P. Micikevicius *et al.*, Mixed precision training, in *International Conference on Learning Representations* (Curran Associates, Inc., Red Hook, NY, 2018).
 - [72] 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](https://arxiv.org/abs/2111.13139).
 - [73] A. Ghosh, N. K. Johnson-McDaniel, A. Ghosh, C. K. Mishra, P. Ajith, W. D. Pozzo, C. P. L. Berry, A. B. Nielsen, and L. London, Testing general relativity using gravitational wave signals from the inspiral, merger and ringdown of binary black holes, *Classical Quantum Gravity* **35**, 014002 (2017).
 - [74] 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).
 - [75] N. J. Cornish, Time-frequency analysis of gravitational wave data, *Phys. Rev. D* **102**, 124038 (2020).
 - [76] A. Dosovitskiy *et al.*, An image is worth 16×16 words: Transformers for image recognition at scale, in *International Conference on Learning Representations* (2021), <https://dblp.org/db/conf/iclr/iclr2021.html>.
 - [77] Astropy Collaboration, Astropy Project Contributors, The Astropy Project: Sustaining and growing a community-oriented open-source project and the latest major release (v5.0) of the core package, *Astrophys. J.* **935**, 167 (2022).
 - [78] N. J. Cornish and T. B. Littenberg, BayesWave: Bayesian inference for gravitational wave bursts and instrument glitches, *Classical Quantum Gravity* **32**, 135012 (2015).
 - [79] T. B. Littenberg and N. J. Cornish, Bayesian inference for spectral estimation of gravitational wave detector noise, *Phys. Rev. D* **91**, 084034 (2015).
 - [80] N. J. Cornish, T. B. Littenberg, B. Bécsy, K. Chatziioannou, J. A. Clark, S. Ghonge, and M. Millhouse, BayesWave analysis pipeline in the era of gravitational wave observations, *Phys. Rev. D* **103**, 044006 (2021).
 - [81] 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).
 - [82] D. Foreman-Mackey, CORNER.PY: Scatterplot matrices in Python, *J. Open Source Software* **1**, 24 (2016).
 - [83] M. J. Williams, jmcginn, federicostak, and J. Veitch, UOF-GRAVITY/GLASFLOW: v0.4.1 (2024), [10.5281/zenodo.13914483](https://zenodo.org/record/13914483).
 - [84] LIGO Scientific Collaboration, LIGO Algorithm Library—LALSuite, free software (GPL) (2018), <https://lscsow.docs.ligo.org/lalsuite/dev/index.html>.
 - [85] K. Wette, SWIGLAL: Python and Octave interfaces to the LALSuite gravitational-wave data analysis libraries, *SoftwareX* **12**, 100634 (2020).
 - [86] L. Singer, LIGO.SKYMAP (2020), <https://lscsoft.docs.ligo.org/ligo.skymap/>.
 - [87] J. D. Hunter, Matplotlib: A 2D graphics environment, *Comput. Sci. Eng.* **9**, 90 (2007).
 - [88] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, NFLOWS: Normalizing flows in PyTorch (2020), [10.5281/zenodo.4296287](https://zenodo.org/record/4296287).
 - [89] C. R. Harris *et al.*, Array programming with NumPy, *Nature (London)* **585**, 357 (2020).
 - [90] T. Pandas Development Team, PANDAS-DEV/PANDAS: PANDAS (2025), [10.5281/zenodo.3509134](https://zenodo.org/record/3509134).
 - [91] A. Nitz *et al.*, GWASTRO/PyCBC: v2.3.3 release of PyCBC (2024), [10.5281/zenodo.10473621](https://zenodo.org/record/10473621).
 - [92] A. Paszke *et al.*, PyTorch: An imperative style, high-performance deep learning library, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2019).
 - [93] M. L. Waskom, SEABORN: Statistical data visualization, *J. Open Source Software* **6**, 3021 (2021).
 - [94] L. Biewald, Experiment tracking with weights and biases, software available from <http://www.wandb.com/> (2020).
 - [95] P. Tol, Accessible color schemes, <https://sronpersonalpages.nl/~pault/>.
 - [96] M. Okabe and K. Ito, Color universal design (CUD): How to make figures and presentations that are friendly to color-blind people (2008), <https://jfly.uni-koeln.de/color/#pallet>.
 - [97] <https://gwosc.org>.