

Enhancing the reliability of machine learning for gravitational wave parameter estimation with attention-based models

Hibiki Iwanaga¹,* Mahoro Matsuyama¹, and Yousuke Itoh¹

Department of Physics, Osaka Metropolitan University, Osaka 558-8585, Japan



(Received 22 January 2025; accepted 10 September 2025; published 9 October 2025)

We introduce a technique to enhance the reliability of gravitational wave parameter estimation results produced by machine learning. We develop two independent machine learning models based on the Vision Transformer to estimate effective spin and chirp mass from spectrograms of gravitational wave signals from binary black hole mergers. To enhance the reliability of these models, we utilize attention maps to visualize the areas our models focus on when making predictions. This approach enables us to demonstrate that both models perform parameter estimation based on physically meaningful information. Furthermore, by leveraging these attention maps, we demonstrate a method to quantify the impact of glitches on parameter estimation. We show that as the models focus more on glitches, the parameter estimation results become more strongly biased. This suggests that attention maps could potentially be used to distinguish between cases where the results produced by the machine learning model are reliable and cases where they are not.

DOI: [10.1103/blfk-7k9f](https://doi.org/10.1103/blfk-7k9f)

I. INTRODUCTION

Since the first detection of gravitational waves from a binary black hole merger on September 14, 2015 [1], observations by LIGO [2] and Virgo [3] have continued. Combining the results of the O1, O2, and O3 observation runs, a total of 90 gravitational wave events have been detected [4–6]. With future improvements in the sensitivity of the LIGO and Virgo detectors, and with the participation of detectors like KAGRA [7] and LIGO-India [8], the number of detected events is expected to continue increasing [9].

In gravitational wave data analysis, when the gravitational wave source is a binary merger event, once a detection is made, a parameter estimation based on Bayesian inference is performed [10]. To estimate the posterior distribution, parameter space sampling is typically performed using stochastic sampling methods such as MCMC [11] or Nested Sampling [12]. Tools such as LALInference [13] and Bilby [14] have been developed to implement these algorithms. However, these methods face the challenge of significant computational costs, primarily due to the time required for waveform generation and likelihood function evaluation. Depending on the signal duration and the complexity of the waveform model, it is known that the total time required for estimation can extend

to several days per event for binary black hole mergers [4] and even several weeks for binary neutron star mergers [15]. This problem is expected to become more serious as the number of detected gravitational wave events increases exponentially in the future [16].

Several efforts have recently been made to reduce the computational cost of gravitational wave parameter estimation using machine learning. In [17], the authors achieved fast analysis by using a CNN to estimate the mass parameters of binary black hole mergers. Additionally, for the events in GWTC-1, they confirmed that the results were consistent with the posterior distributions obtained from LALInference.

In [18,19], Bayes' rule is applied to enable network training using only samples from the prior distribution and gravitational wave likelihood. This approach allows for the estimation of posterior distributions without the need for posterior samples during training. As noted above, likelihood evaluation requires significant computational costs. Therefore, these so-called “likelihood-free” methods are highly valuable, leading to the exploration of many potential applications [20–22]. A notable example is the DINGO framework [23], which uses normalizing flows to estimate all 15 parameters of BBH mergers, achieving results consistent with the posterior distributions obtained from LALInference. In the DINGO framework, the neural network is conditioned not only on the strain data but also on the power spectral density (PSD) of the detector noise, effectively addressing the fact that detector noise varies from event to event. This excellent framework has also been expanded to parameter estimation for binary neutron stars [24]. Therefore, machine learning is progressing

*Contact author: iwanagahibiki7382@gmail.com

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

toward practical applications not only in gravitational wave detection and glitch classification but also in parameter estimation (See, e.g., [25–28]). (For typical approaches that do not use machine learning, such as reduced-order modeling, efficient sampling methods, and others, see, e.g., [29–32] and references therein.)

When applying machine learning to gravitational wave parameter estimation, it is important to consider the “black-box” problem [33,34]. Because machine learning models learn from large amounts of data and autonomously generate outputs, their decision-making process is unclear. For example, even if there are significant noises such as glitches nearby gravitational wave signals, machine learning models generate outputs without warning. For another example, it has not been sufficiently verified whether the models are focusing on physically meaningful information.

In this work, we used data-efficient image transformers (DeiT) [35], which is a model based on ViT [36]. ViT applies the transformer architecture [37], originally designed for natural language processing, to image data. ViT utilizes image patch embedding, positional embeddings, transformer encoder, and classification token. While ViT achieved the state-of-the-art results in image classification when it appeared, it requires huge training datasets and it is computationally expensive for small datasets. DeiT has then successfully reduced the computational cost of training while maintaining accuracy by utilizing knowledge distillation [38] and data augmentation. We adopted DeiT due to our limited availability of computing resources. We make use of attention maps produced by DeiT that show the areas a machine learning model focuses on when it makes predictions, thereby improving the reliability of outputs. See the Appendix A for further details about the DeiT network we used.

II. METHOD

Our main purpose is to verify whether the machine learning models focus on physically meaningful information through the attention maps. Therefore, we used the spectrogram as it is intuitive to understand, making it easier to review and discuss the results afterward. As a first step, we build models to estimate the effective spin and chirp mass of binary black hole mergers. To examine the differences in attention regions for the estimation of each parameter, we prepared two independent models. In this section, we describe the details of the dataset used for training our models.

A. GW dataset generation

For the generation of CBC waveforms, we used PyCBC [39] with the SEOBNRv4_opt [40] waveform model. Our models were trained with parameters from Table I. Although some parameters are restricted, we chose a simple

TABLE I. Parameters used to create the training and validation datasets. $\mathcal{M}$, q , χ_{eff} , θ_{JN} , and ψ are uniformly distributed. Source-frame component masses m_1, m_2 , are determined from $\mathcal{M}$ and q , while the dimensionless projections of the individual BH spins χ_1 and χ_2 are determined from χ_{eff} . The luminosity distance d_L and the sky position (right ascension α and declination δ) are set to ensure a uniform distribution of gravitational wave sources within a sphere.

	Range
Source-frame chirp mass $\mathcal{M}/M_\odot$	[20, 40]
Mass ratio q	[0.4, 1.0]
Effective spin χ_{eff}	[−0.88, 0.88]
Luminosity distance d_L/Mpc	[50, 600]
Inclination angle θ_{JN}	[0, π]
Polarization angle ψ	[0, π]

configuration for the sake of simplicity, considering our main purpose.

Regarding the data generation procedure, we first created strain data by injecting a waveform into 4 second long data of stationary Gaussian noise colored by the LIGO design sensitivity (aLIGOZeroDetHighPower) [39]. During this process, the central time of the noise data was aligned with the merger time of the waveform. Then, we whitened the strain time series and cut the first and last 0.5 seconds of data to minimize possible edge effects. After the whitening process, we applied the constant-Q transform to create the spectrogram [41]. The constant-Q transform is a type of time-frequency transform similar to the Fourier transform. The constant-Q transform has high frequency resolution at low frequencies and high time resolution at high frequencies. Depending on the properties of data, the constant-Q transform is beneficial for catching detailed spectral structures than the usual Fourier transform.

This spectrogram covers a frequency range of $f_{\text{min}} = 20$ [Hz] to $f_{\text{max}} = 1024$ [Hz] and includes 3 seconds of data. Following this process, we generated 10,000 spectrograms and split them into training and validation sets in a 9 : 1 ratio to train the models.

B. Training and validation

As was mentioned in the Introduction, we used data-efficient image transformers (DeiT) [35] that is pretrained on ImageNet [42] and we trained it on the dataset we created.

Our models were trained for 50 epochs with a batch size of 16. The Adam optimizer [43] was used for optimization. Figs. 1 and 2 presents scatter plots illustrating the relationship between predicted and actual values for the validation dataset, after 50 epochs of training and validation.

Based on these results, it was confirmed that our models successfully learned the characteristics of the effective spin and chirp mass. The training took approximately one hour on an NVIDIA GeForce RTX 3090.

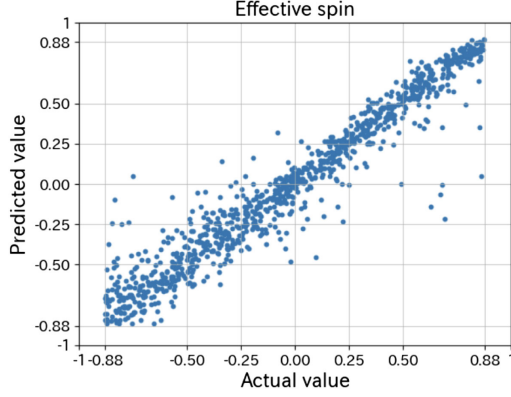


FIG. 1. A predicted vs actual plot after 50 epochs for the effective spin estimation model. The model was trained for 50 epochs with a learning late of 1×10^{-4} and a batch size of 16.

C. Evaluation of uncertainty

To evaluate the uncertainty in the predictions of our models, we developed an uncertainty evaluation method based on a primitive Monte Carlo approach. See the Appendix B for further explanation and possible use cases of our method. The flowchart of this method is shown in Fig. 3.

- (1) Initially, a waveform was generated using parameters consistent with the estimates obtained from LALInference for GW150914. After injecting this waveform into stationary Gaussian noises colored by the same PSD as in Sec. II A, a constant-Q transform was applied to create a spectrogram.
- (2) Using this spectrogram as input to our models, we estimated the effective spin and chirp mass.
- (3) Subsequently, a waveform was generated based on the values estimated in (2). For parameters that were not estimated, the values were set to those used in (1). This waveform was injected into stationary Gaussian noises colored by the same PSD as in

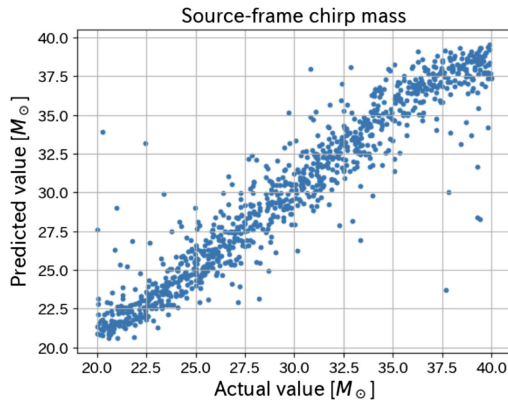


FIG. 2. A predicted vs actual plot after 50 epochs for the chirp mass estimation model. The model was trained for 50 epochs with a learning late of 5×10^{-5} and a batch size of 16.

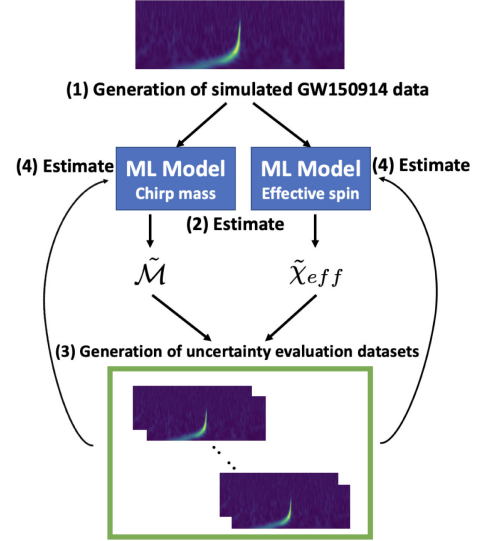


FIG. 3. Flow chart of uncertainty evaluation.

Sec. II A, and a constant-Q transform was applied to create spectrograms. In this way, we created an uncertainty evaluation dataset that contains 1,000 spectrograms.

- (4) We evaluated absolute errors by using our models to make predictions on this dataset. Then we calculated 90% confidence intervals for our models.

III. RESULTS

A. Prediction results

Table II shows the 90% confidence intervals, while Figs. 4 and 5 illustrate the distributions of absolute errors. In Table II, it can be seen that the predictions of our models were approximately consistent with those obtained by LALInference for GW150914. However, we did not analyze real GW events, and since the PSD used corresponds to the design sensitivity, this represents the result of a rough comparison. We also remark that since we used a frequentist approach while the results by LALInference are based on a Bayesian approach, two results are not necessarily expected to be the same.

TABLE II. Comparisons of predictions between our models and LALInference. Our results provide the 90% confidence intervals, while the LALInference results are the 90% credible intervals estimated using the EOBNR waveform, as reported in [44]. Note that we used stationary Gaussian noise and assuming design sensitivity.

	Our models	LALInference
χ_{eff}	$-0.09^{+0.07}_{-0.09}$	$-0.09^{+0.19}_{-0.17}$
$\mathcal{M}/M_{\odot}$	$27.6^{+0.8}_{-1.0}$	$27.9^{+2.3}_{-1.8}$

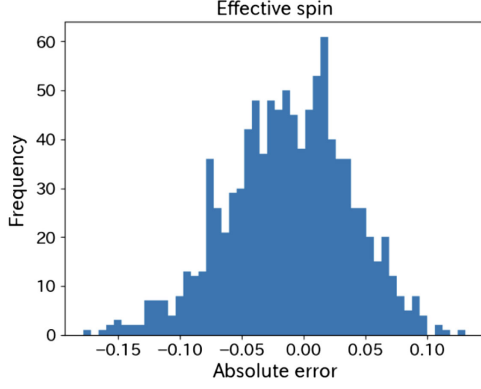


FIG. 4. Histogram of absolute error for the effective spin estimation model. The absolute errors are calculated from the effective spins estimated in Sec. II C 4 and the estimated value in Sec. II C 2.

The total estimation time was approximately six minutes, confirming that computational costs can be significantly reduced. This method does not require extensive waveform generation and likelihood evaluations, making it feasible to achieve a similar order of efficiency when applied to real GW data.

B. Attention map

We generated attention maps for the 1,000 spectrograms used in the uncertainty evaluation described in Sec. II C. Each attention map we used is for the class token at the last self-attention layer of the network. To illustrate the overall trend, we created a composite figure by overlaying these attention maps (Fig. 6). It can be seen that the effective spin estimation model appears to focus primarily on the mid-frequency range, whereas the chirp mass estimation model focuses on the low-frequency range.

Each intrinsic binary parameter has a specific frequency range where its information predominantly comes from.

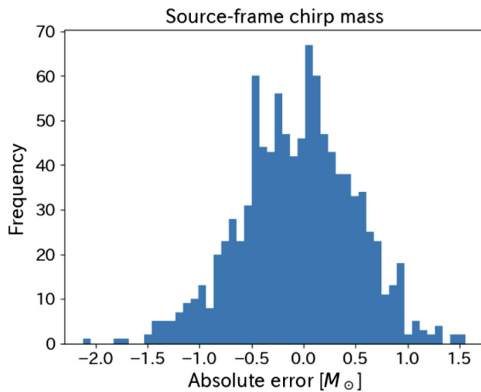


FIG. 5. Histogram of absolute error for the chirp mass spin estimation model. The absolute errors are calculated from the chirp masses estimated in Sec. II C 4 and the estimated value in Sec. II C 2.

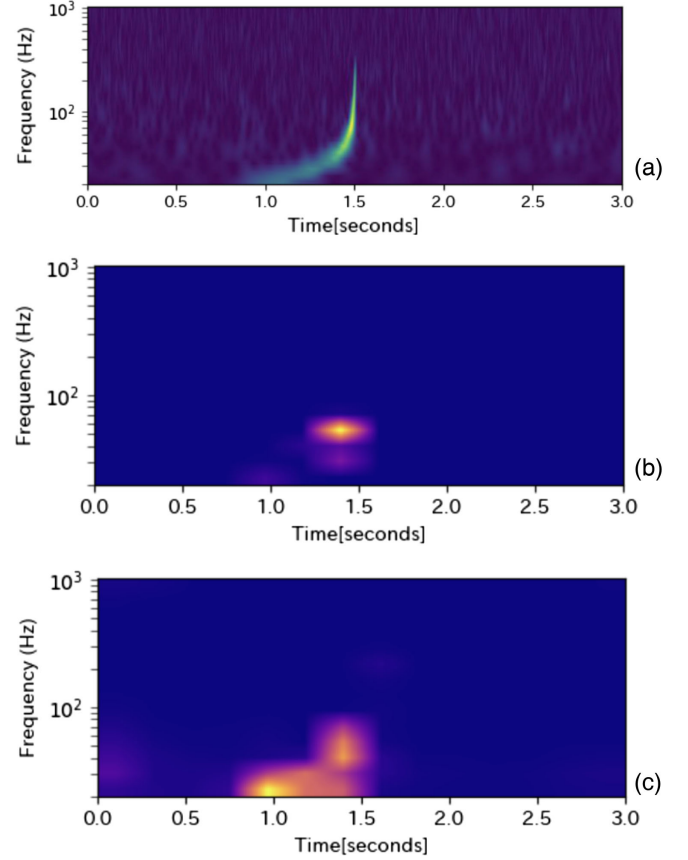


FIG. 6. Attention maps for the uncertainty evaluation dataset of simulated GW150914. (a) An example of input data. (b) Composite of 1000 overlaid attention maps for effective spin estimation. (c) Composite for chirp mass. All attention maps are composed of 14×14 pixels and are resized to match the 224×224 resolution of the input data using bilinear interpolation.

It is known that the chirp mass tends to appear more prominently at lower frequencies, while the effective spin shows its effect at mid-to-high frequencies [45,46]. From the attention map results, we confirmed that our models focus on the ranges where the effects of each parameter are predominant. These results suggest that the predictions of our models are based on physically meaningful information.

C. Attention maps for a longer GW signal

To more clearly see the differences in the attention regions of the two models, we conducted the same analysis using a waveform characterized by a large effective spin and a small chirp mass, which results in a relatively longer signal duration.

We generated a waveform with an effective spin of 0.8 and a source-frame chirp mass of $22M_{\odot}$, and performed uncertainty evaluation following the same procedure as in Sec. II C. We then created a composite figure by overlaying 1000 attention maps (Fig. 7). It can be seen that the two

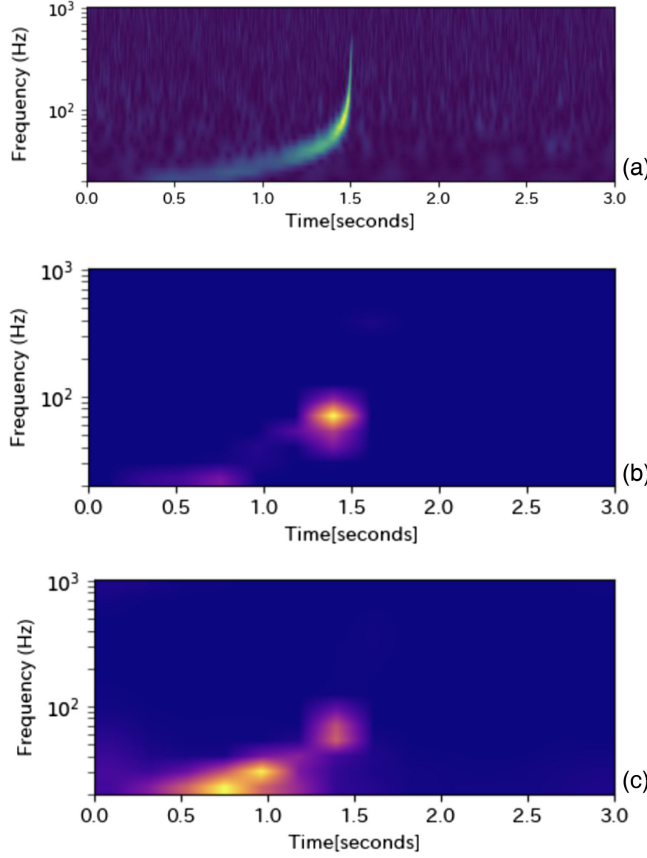


FIG. 7. Attention maps for the uncertainty evaluation dataset of simulated longer GW event. (a) An example of input data. (b) composite of 1000 overlaid attention maps for effective spin estimation. (c) Composite for chirp mass. All attention maps are composed of 14×14 pixels and are resized to match the 224×224 resolution of the input data using bilinear interpolation.

models are focusing on the midfrequency range and low-frequency range, respectively.

IV. APPLICATION OF ATTENTION MAPS FOR EVALUATING THE IMPACT OF GLITCHES

In real strain data, transient noise artifacts, commonly known as glitches, occasionally occur nearby gravitational wave signals [47–49]. These glitches are problematic as they not only generate false-positive candidates [50,51] but also bias parameter estimation [52–55]. Therefore, to apply machine learning to the parameter estimation of gravitational waves, it is essential to properly evaluate the impact of glitches and enhance the reliability of the results.

In this study, we demonstrate the potential of utilizing attention maps to quantify the impact of glitches on parameter estimation.

A. Method

We injected glitches into the simulated GW150914 data and applied the same uncertainty evaluation method as in

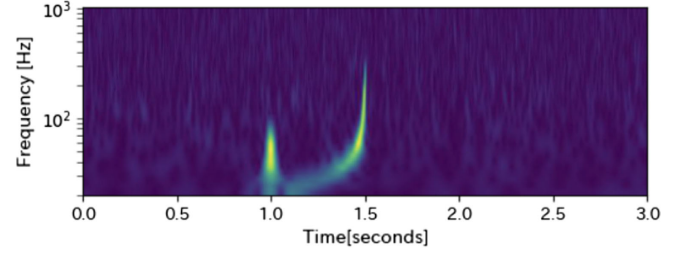


FIG. 8. An example spectrogram of simulated GW150914 data with an injected glitch of amplitude 1A.

Sec. II C to estimate the extent to which our model’s predictions would be biased if glitches were present in the simulated GW event. Furthermore, we examined the relationship between the attention map and the prediction results.

For simplicity, we modeled glitches by sine-Gaussian waveforms [56]. The sine-Gaussian waveform can be written in the time domain as

$$h_{\text{sg}} = A_{\text{sg}} e^{-\frac{(t-t_0)^2}{\tau^2}} \cos(2\pi f_0(t-t_0)) \quad (1)$$

A_{sg} is an overall amplitude, t_0 and f_0 are the location of the sine-Gaussian in time and frequency, respectively, and τ is the damping time.

In real GW data analysis, it is problematic when glitches directly overlap GW signals as in the case of GW170817 [15]. However, injecting around the merger phase seemed likely to cause significant bias, thereby making it difficult to identify clear trends. Therefore, we selected $\tau = 10^{-2}$ [s], $f_0 = 40$ [Hz] and $t_0 = 1.0$ [s] to configure the glitches to slightly overlap with the inspiral phase. For the amplitude A_{sg} , we prepared three glitches with amplitudes of 0.5A, 0.8A, and 1A, using the maximum amplitude A of the simulated GW signal. In this manner, we generated spectrograms for three simulated GW events with injected glitches. Figure 8 presents an example.

B. Predictions

We compared the prediction results from Sec. III A with three simulated GW events with glitches of three different amplitudes.

Table III shows the 90% confidence intervals, while Figs. 9 and 10 illustrates the distributions of the predicted

TABLE III. 90% confidence intervals for a simulated GW150914 event without glitches and for three simulated GW events with varying glitch amplitudes.

	No glitches	0.5A	0.8A	1A
χ^2_{eff}	$-0.09^{+0.07}_{-0.09}$	$-0.09^{+0.09}_{-0.07}$	$-0.07^{+0.08}_{-0.07}$	$0.04^{+0.07}_{-0.05}$
$\mathcal{M}/M_{\odot}$	$27.6^{+0.8}_{-1.0}$	$30.5^{+0.8}_{-1.1}$	$32.6^{+0.6}_{-1.4}$	$33.3^{+0.7}_{-1.3}$

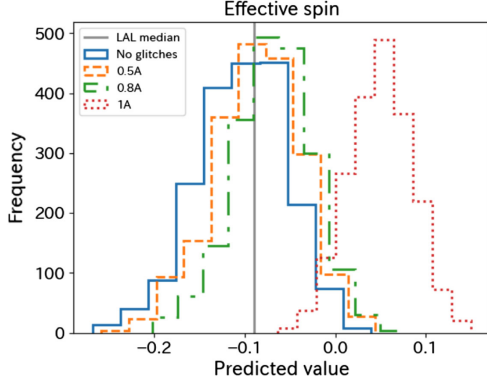


FIG. 9. Histograms of the predicted effective spins for a simulated GW150914 event without glitches and for three simulated GW events with varying glitch amplitudes. The solid vertical lines are the median estimates by LALInference [44].

values. It can be seen that for both effective spin and chirp mass, glitches with larger amplitudes more strongly biased parameter estimation. This tendency is also commonly seen in conventional methods.

C. Attention map

Our goal here is to demonstrate the potential use of attention maps for evaluating the impact of glitches on parameter estimation.

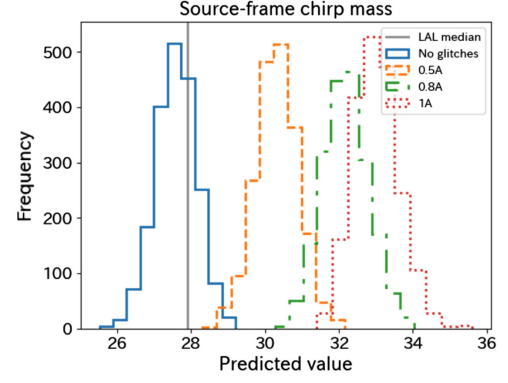


FIG. 10. Histograms of the predicted chirp masses for a simulated GW150914 event without glitches and for three simulated GW events with varying glitch amplitudes.

Figures 11 and 12 shows the attention maps for a simulated GW150914 event without glitches and for three simulated GW events with varying glitch amplitudes. It can be seen that as the amplitude of the glitch increases, the attention around 1.0 seconds and 40 Hz, where the glitch was injected, also becomes larger.

Furthermore, to quantify the impact of glitches to some extent, we calculated the total attention values for each time point. The distribution is shown in Figs. 13 and 14. The figures illustrate that as the amplitude of the glitch

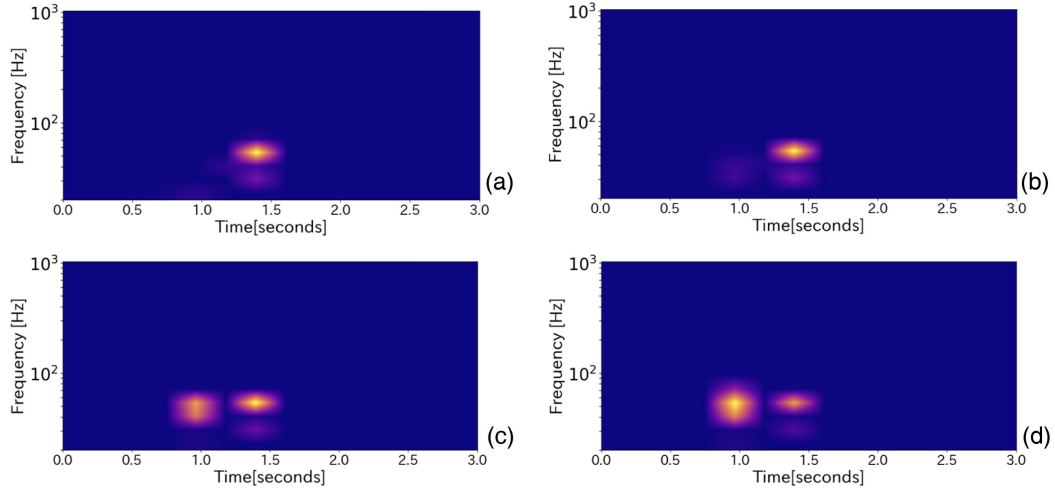


FIG. 11. Attention maps for effective spin estimation model. (a) A simulated GW150914 without glitches. Simulated GW events with glitch amplitudes of (b) 0.5A, (c) 0.8A, and (d) 1A.

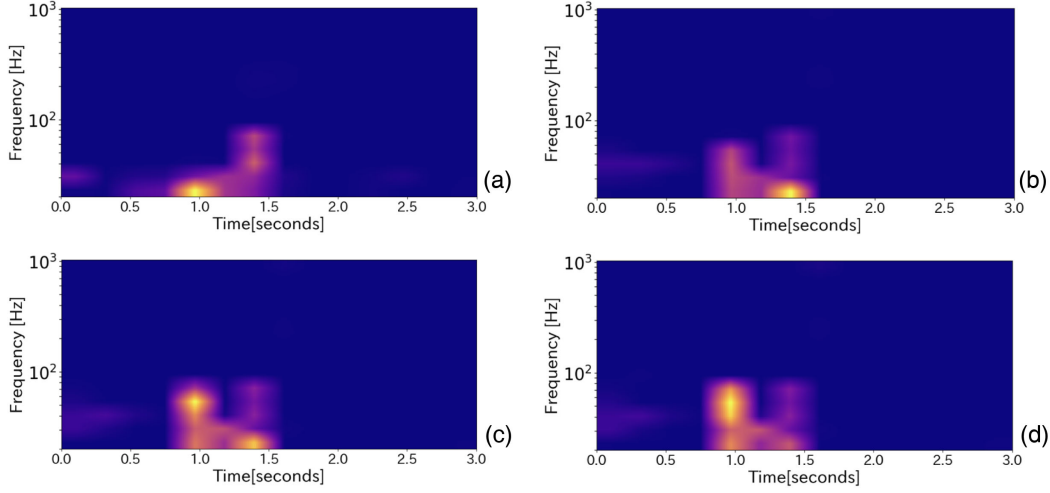


FIG. 12. Attention maps for chirp mass estimation model. (a) A simulated GW150914 without glitches. Simulated GW events with glitch amplitudes of (b) 0.5A, (c) 0.8A(c), and (d) 1A.

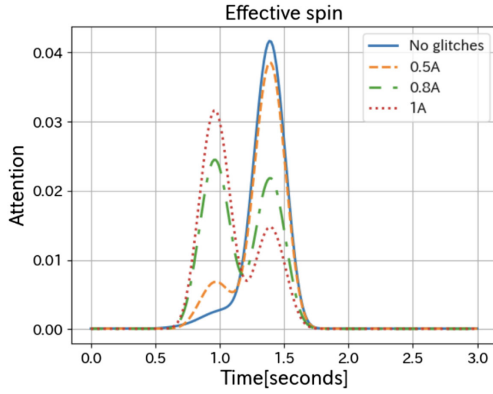


FIG. 13. Distributions of the total attention values computed at each time point for effective spin estimation model. All attention values are normalized by scaling between 0 and 1. Additionally, smoothing was performed using a 1D Gaussian filter to account for the resolution of the attention map.

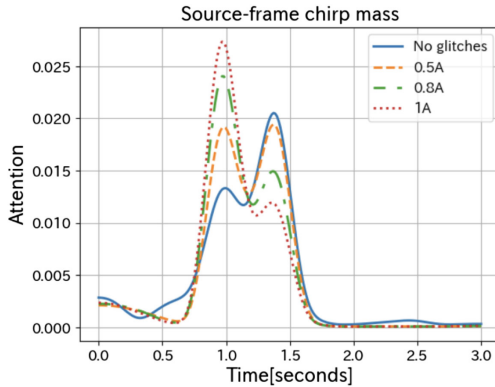


FIG. 14. Distributions of the total attention values computed at each time point for chirp mass estimation model.

increases, the attention values around it also tend to increase.

V. CONCLUSION

We developed two machine learning models based on the Vision Transformer to estimate effective spin and chirp mass from spectrograms of gravitational wave signals. Predictions of our models for simulated GW150914 were consistent with those obtained by LALInference for GW150914. The total estimation time was approximately six minutes, confirming that computational costs can be significantly reduced. This is an easier task than using observational data, which includes real noise, and therefore does not allow for a rigorous comparison. However, regarding the total estimation time, it is feasible to achieve a similar level of efficiency when applied to real GW data, as this method does not require extensive waveform generation or likelihood evaluations.

We analyzed the differences in the attention regions between the effective spin estimation model and the chirp mass estimation model using attention maps. As a result, we confirmed that our models focus on the ranges where the effects of each parameter are predominant. These results suggest that the predictions of our models are based on physically meaningful information.

Currently, our results only indicate that the attention regions of our models are roughly consistent with the ranges where the effects of each parameter are predominant. By applying the fisher analysis conducted in [46], it may be possible to quantitatively demonstrate the degree to which they are consistent.

Furthermore, we demonstrated the potential of utilizing attention maps to quantify the impact of glitches on parameter estimation. As a result, we found that as the models focus more on glitches, the parameter estimation

results become more strongly biased. This suggests that attention maps could potentially be used to distinguish between cases where the results produced by the machine learning model are reliable and cases where they are not.

It would be interesting to explore including data with glitches in the training set as a way to develop a model that is robust to glitches. In such a model, attention may not be directed toward the glitches. It would also be interesting to analyze the attention maps of our models after applying the glitch mitigation methods proposed in, e.g., [57].

In this paper, we proposed that attention maps enhance the reliability of the gravitational wave parameter estimation against transient noises in the detector output called glitches. Leveraging the power of the attention maps requires an inspection of each map for each event. If the detection rate of compact binary star mergers is small, as is the case currently, we can manually inspect the attention maps of each event. When the detection rate becomes larger, as is expected to happen in near future as the detectors' sensitivities improve, we have to develop an automatic way of assessing the attention maps. One way to do that would be to check the consistency between the time of coalescence estimated from the matched filtering stage and the time when the attention map peaks.

As for other technical details, we used Gaussian noise to estimate parameter errors. We leave use of real data as future work. This work is intended to show that attention maps enhance the reliability of the machine learning technique in gravitational wave data analysis. For this purpose, we used a spectrogram as it is easier to understand than a time series or its Fourier transform. As a possible future direction, it would be interesting to use time series in addition to spectrogram to make a network learn phases of gravitational waves and glitches. Finally, we studied attention maps of the class token at the last output stage of the network. It would be interesting to use the attention roll-out that incorporates the information of the attention map of each layer [58].

ACKNOWLEDGMENTS

We would like to thank Nobuyuki Kanda, Masaki Uematsu, Izumi Kaku, Takumi Fujimori, Ryuki Kawamoto, and Hiroyuki Nakano for helpful comments. We are grateful to Kota Tomita and Kyohei Takatani for their contribution in the early phase of this work. We would also like to thank Hirotaka Takahashi, Takashi Uchiyama, Yutaka Shikano, Shoichi Oshino and Yusuke Sakai for the valuable insights provided regarding the application of machine learning to gravitational wave data analysis. This work was supported by JSPS KAKENHI Grant No. 20H05639. We use PyTorch [59] for the implementation of our models. The plots are generated with MATPLOTLIB [60].

DATA AVAILABILITY

The data that support the findings of this article are not publicly available upon publication because it is not technically feasible and/or the cost of preparing, depositing, and hosting the data would be prohibitive within the terms of this research project. The data are available from the authors upon reasonable request.

APPENDIX A: DEiT

In this work, we used the data-efficient image transformers, specifically, the `deit_small_distilled_patch16_224` model provided by the `timm` library, and we followed the default implementation for fine-tuning and inference. As described in Sec. II B (Training and validation), we used DeiT models pretrained on ImageNet. Therefore, we did not train a separate teacher network on the same tasks of mass and spin estimation. The basic architecture of the model is publicly available at the following site [61], and as shown there, in addition to the class token, the distillation token also contributes to the model's predictions.

APPENDIX B: EVALUATION OF UNCERTAINTIES

In this appendix, we give the idea behind and possible use cases of our method of evaluating parameter uncertainties.

We have used a simple frequentist approach. Let us explain it using a simple example. Suppose our data x consists of a noise n and signal μ like $x = n + \mu$. We may estimate the value of μ from x . Now, if we have lots of statistically independent x 's ($x_1, x_2, \dots, x_N$) for a fixed μ , then we obtain N point-estimates of μ from which we obtain an error estimate of our point estimate of μ . To realize this, we need to have lots of noise realizations $\{n_i\}_{i=1}^N$.

In the real-world application of our method, we propose that such noise realizations are obtained by injecting a simulated signal h with fixed signal parameters $\hat{p}$ into time series starting at many different times, assuming that the statistical properties of the detector output time series are stable in time. Indeed, this is what is done to estimate the false alarm rate of gravitational wave events in the literature.

The question then is how we find $\hat{p}$. A possible flow of the data analysis would be the following. The first process is the matched filtering, with which one searches for gravitational wave events and, if detected, which gives us point estimates of the astrophysical parameters such as the masses of the stars. One usually uses, e.g., Markov-Chain Monte-Carlo or nested sampling techniques to estimate the parameter errors, but that takes huge computational costs. One of the problems we are tackling in this paper is this computational cost problem. As such, to some

extent, we can assume that we are given point estimates of parameters from the search pipeline (matched filtering stage) before we start. We then use that set of estimated parameters to generate a simulated waveform and inject it

into sets of time series at many different time stamps (where no signal is known to be present). We obtain astrophysical parameters from each time series and error estimates from them.

-
- [1] B. Abbott *et al.* (The LIGO Scientific Collaboration and the Virgo Collaboration), Observation of gravitational waves from a binary black hole merger, *Phys. Rev. Lett.* **116**, 061102 (2016).
 - [2] J. Aasi *et al.* (The LIGO Scientific Collaboration), Advanced LIGO, *Classical Quantum Gravity* **32**, 074001 (2015).
 - [3] F. Acernese *et al.*, Advanced Virgo: A second-generation interferometric gravitational wave detector, *Classical Quantum Gravity* **32**, 024001 (2015).
 - [4] B. Abbott *et al.* (The LIGO Scientific Collaboration and the Virgo Collaboration), 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).
 - [5] R. Abbott *et al.* (The LIGO Scientific Collaboration and the Virgo Collaboration), 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).
 - [6] R. Abbott *et al.*, 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).
 - [7] Y. Aso, Y. Michimura, K. Somiya, M. Ando, O. Miyakawa, T. Sekiguchi, D. Tatsumi, and H. K. Yamamoto, Interferometer design of the KAGRA gravitational wave detector, *Phys. Rev. D* **88**, 043007 (2013).
 - [8] C. S. Unnikrishnan, IndiGO and Ligo-India scope and plans for gravitational wave research and precision metrology in India, *Int. J. Mod. Phys. D* **22**, 1341010 (2013).
 - [9] B. Abbott *et al.*, Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA, *Living Rev. Relativity* **23**, 3 (2020).
 - [10] B. P. Abbott *et al.*, A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals, *Classical Quantum Gravity* **37**, 055002 (2020).
 - [11] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, emcee: The MCMC Hammer, *Publ. Astron. Soc. Pac.* **125**, 306 (2013).
 - [12] J. S. Speagle, DYNESTY: A dynamic nested sampling package for estimating Bayesian posteriors and evidences, *Mon. Not. R. Astron. Soc.* **493**, 3132 (2020).
 - [13] 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).
 - [14] G. Ashton *et al.*, Bilby: A user-friendly bayesian inference library for gravitational-wave astronomy, *Astrophys. J.* **241**, 27 (2019).
 - [15] B. Abbott *et al.*, Properties of the binary neutron star merger GW170817, *Phys. Rev. X* **9**, 011001 (2019).
 - [16] F. S. Broekgaarden, S. Banagiri, and E. Payne, Visualizing the number of existing and future gravitational-wave detections from merging double compact objects, *Astrophys. J.* **969**, 108 (2024).
 - [17] A. McLeod, D. Jacobs, C. Chatterjee, L. Wen, and F. Panther, Rapid mass parameter estimation of binary black hole coalescences using deep learning, [arXiv:2201.11126](https://arxiv.org/abs/2201.11126).
 - [18] A. J. Chua and M. Vallisneri, Learning bayesian posteriors with neural networks for gravitational-wave inference, *Phys. Rev. Lett.* **124**, 041102 (2020).
 - [19] 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).
 - [20] S. R. Green and J. Gair, Complete parameter inference for GW150914 using deep learning, *Mach. Learn. Sci. Tech.* **2**, 03LT01 (2021).
 - [21] S. R. Green, C. Simpson, and J. Gair, Gravitational-wave parameter estimation with autoregressive neural network flows, *Phys. Rev. D* **102**, 104057 (2020).
 - [22] H. Shen, E. A. Huerta, E. O'Shea, P. Kumar, and Z. Zhao, Statistically-informed deep learning for gravitational wave parameter estimation, *Mach. Learn. Sci. Technol.* **3**, 015007 (2022).
 - [23] 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).
 - [24] 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).
 - [25] P. Couvares, I. Bird, E. Porter, S. Bagnasco, M. Punturo, D. Reitze, S. Katsanevas, T. Kajita, V. Kalogera, H. Lueck, D. McClelland, S. Rowan, G. Sanders, B. S. Sathyaprakash, D. Shoemaker, and J. van den Brand, Gravitational wave data analysis: Computing challenges in the 3G Era, [arXiv:2111.06987](https://arxiv.org/abs/2111.06987).
 - [26] E. Cuoco *et al.*, Enhancing gravitational-wave science with machine learning, *Mach. Learn. Sci. Technol.* **2**, 011002 (2021).
 - [27] Y. Sakai, Y. Itoh, P. Jung, K. Kokeyama, C. Kozakai, K. T. Nakahira, S. Oshino, Y. Shikano, H. Takahashi, T. Uchiyama, G. Ueshima, T. Washimi, T. Yamamoto, and T. Yokozawa, Training process of unsupervised learning

- architecture for gravity spy dataset, *Ann. Phys. (Berlin)* **536**, 2200140 (2024).
- [28] Y. Sakai, Y. Itoh, P. Jung, K. Kokeyama, C. Kozakai, K. T. Nakahira, S. Oshino, Y. Shikano, H. Takahashi, T. Uchiyama, G. Ueshima, T. Washimi, T. Yamamoto, and T. Yokozawa, Unsupervised learning architecture for classifying the transient noise of interferometric gravitational-wave detectors, *Sci. Rep.* **12**, 9935 (2022).
- [29] S. Morisaki, Accelerating parameter estimation of gravitational waves from compact binary coalescence using adaptive frequency resolutions, *Phys. Rev. D* **104**, 044062 (2021).
- [30] E. Lee, S. Morisaki, and H. Tagoshi, Mass-spin reparametrization for a rapid parameter estimation of inspiral gravitational-wave signals, *Phys. Rev. D* **105**, 124057 (2022).
- [31] N. Adhikari and S. Morisaki, Accelerating gravitational-wave parametrized tests of general relativity using a multi-band decomposition of likelihood, *Phys. Rev. D* **106**, 104053 (2022).
- [32] S. Morisaki, R. Smith, L. Tsukada, S. Sachdev, S. Stevenson, C. Talbot, and A. Zimmerman, Rapid localization and inference on compact binary coalescences with the Advanced LIGO-Virgo-KAGRA gravitational-wave detector network, *Phys. Rev. D* **108**, 123040 (2023).
- [33] M. Tulio Ribeiro, S. Singh, and C. Guestrin, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session* (Association for Computing Machinery, New York, NY, 2016), pp. 97–101.
- [34] Z. C. Lipton, in *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)* (Association for Computing Machinery, New York, NY, 2016).
- [35] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, Training data-efficient image transformers & distillation through attention, *arXiv:2012.12877*.
- [36] A. Dosovitskiy *et al.*, An image is worth 16×16 words: Transformers for image recognition at scale, *arXiv:2010.11929*.
- [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA* (Curran Associates Inc., Red Hook, NY, 2017).
- [38] G. Hinton, O. Vinyals, and J. Dean, in *Proceedings of the NIPS 2014 Deep Learning Workshop* (Curran Associates Inc., Red Hook, NY, 2015).
- [39] A. Nitz *et al.*, gwastro/pycbc: v2.3.3 release of pycbc, (2024).
- [40] C. Devine, Z. B. Etienne, and S. T. McWilliams, Optimizing spinning time-domain gravitational waveforms for advanced LIGO data analysis, *Classical Quantum Gravity* **33**, 125025 (2016).
- [41] S. Chatterji, L. Blackburn, G. Martin, and E. Katsavounidis, Multiresolution techniques for the detection of gravitational-wave bursts, *Classical Quantum Gravity* **21**, S1809 (2004).
- [42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronics Engineers (IEEE), New York, NY, 2009), pp. 248–255.
- [43] D. P. Kingma and J. Ba, in *Proceedings of the 3rd International Conference for Learning Representations, San Diego, 2015* (The International Conference on Learning Representations (ICLR), Appleton, WI, 2014).
- [44] B. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Properties of the binary black hole merger GW150914, *Phys. Rev. Lett.* **116**, 241102 (2016).
- [45] L. Blanchet, Post-Newtonian theory for gravitational waves, *Living Rev. Relativity* **27**, 4 (2024).
- [46] I. Harry and T. Hinderer, Observing and measuring the neutron-star equation-of-state in spinning binary neutron star systems, *Classical Quantum Gravity* **35**, 145010 (2018).
- [47] D. Davis *et al.*, LIGO detector characterization in the second and third observing runs, *Classical Quantum Gravity* **38**, 135014 (2021).
- [48] F. Acernese *et al.*, Virgo detector characterization and data quality during the O3 run, *arXiv:2205.01555*.
- [49] T. Akutsu *et al.*, Overview of KAGRA: Calibration, detector characterization, physical environmental monitors, and the geophysics interferometer, *Prog. Theor. Exp. Phys.* **2021**, 05A102 (2021).
- [50] B. Abbott *et al.*, Effects of data quality vetoes on a search for compact binary coalescences in Advanced LIGO's first observing run, *Classical Quantum Gravity* **35**, 065010 (2018).
- [51] D. Davis, T. B. Littenberg, I. M. Romero-Shaw, M. Millhouse, J. McIver, F. Di Renzo, and G. Ashton, Subtracting glitches from gravitational-wave detector data during the third LIGO-Virgo observing run, *Classical Quantum Gravity* **39**, 245013 (2022).
- [52] J. Powell, Parameter estimation and model selection of gravitational wave signals contaminated by transient detector noise glitches, *Classical Quantum Gravity* **35**, 155017 (2018).
- [53] R. Macas, J. Pooley, L. K. Nuttall, D. Davis, M. J. Dyer, Y. Lecoecue, J. D. Lyman, J. McIver, and K. Rink, Impact of noise transients on low latency gravitational-wave event localization, *Phys. Rev. D* **105**, 103021 (2022).
- [54] 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).
- [55] 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).
- [56] L. Blackburn *et al.*, The LSC glitch group: Monitoring noise transients during the fifth LIGO science run, *Classical Quantum Gravity* **25**, 184004 (2008).
- [57] S. Ghonge, J. Brandt, J. M. Sullivan, M. Millhouse, K. Chatziioannou, J. A. Clark, T. Littenberg, N. Cornish, S. Hourihane, and L. Cadonati, Assessing and mitigating the impact of glitches on gravitational-wave parameter estimation: A model agnostic approach, *Phys. Rev. D* **110**, 122002 (2024).
- [58] S. Abnar and W. Zuidema, in *Proceedings of the 58th Annual Meeting of the Association for Computational*

- Linguistics* (Association for Computational Linguistics, Kerrville, TX, 2020), pp. 4190–4197.
- [59] A. Paszke, S. Gross, F. Massa *et al.*, in *Advances in Neural Information Processing Systems 32* (Curran Associates, Inc., Red Hook, NY, 2019), pp. 8024–8035.
- [60] J.D. Hunter, Matplotlib: A 2D graphics environment, *Comput. Sci. Eng.* **9**, 90 (2007).
- [61] <https://github.com/huggingface/pytorch-image-models/blob/main/timm/models/deit.py>.