Dilated convolutional neural network for detecting extreme-mass-ratio inspirals

Tianyu Zhao,[1,2,3] Yue Zhou,[2] Ruijun Shi,[1,3] Zhoujian Cao,[1,3,4,*] and Zhixiang Ren[2,†]

[1]Department of Astronomy, Beijing Normal University, Beijing 100875, China
[2]Peng Cheng Laboratory, Shenzhen, 518055, China
[3]Institute for Frontiers in Astronomy and Astrophysics, Beijing Normal University, Beijing 102206, China
[4]School of Fundamental Physics and Mathematical Sciences, Hangzhou Institute for Advanced Study, UCAS, Hangzhou 310024, China

(Received 1 September 2023; accepted 28 March 2024; published 23 April 2024)

The detection of extreme-mass-ratio inspirals (EMRIs) is intricate due to their complex waveforms, extended duration, and low signal-to-noise ratio (SNR), making them more challenging to be identified compared to compact binary coalescences. While matched filtering-based techniques are known for their computational demands, existing deep learning-based methods primarily handle time-domain data and are often constrained by data duration and SNR. In addition, most existing work ignores time delay interferometry (TDI) and applies the long-wavelength approximation in detector response calculations, thus limiting their ability to handle laser frequency noise. In this study, we introduce dilated convolutional neural network for detecting extreme-mass-ratio inspirals (DECODE), an end-to-end model focusing on EMRI signal detection by sequence modeling in the frequency domain. Centered around a dilated causal convolutional neural network, trained on synthetic data considering TDI-1.5 detector response, DECODE can efficiently process a year's worth of multichannel TDI data with an SNR of around 50. We evaluate our model on one-year data with accumulated SNR ranging from 50 to 120 and achieve a true positive rate of 96.3% at a false positive rate of 1%, keeping an inference time of less than 0.01 seconds. With the visualization of three showcased EMRI signals for interpretability and generalization, DECODE exhibits strong potential for future space-based gravitational wave data analyses.

DOI: 10.1103/PhysRevD.109.084054

I. INTRODUCTION

The groundbreaking detection of gravitational waves (GWs) in 2015, exemplified by the GW150914 event, has profoundly impacted the field of astrophysics [1]. Enabled by the Laser Interferemeter Gravitational Wave Observatory (LIGO) [2] and Virgo [3], this remarkable achievement unequivocally confirmed the existence of GWs, providing empirical validation of general relativity (GR) [4]. Beyond enriching our knowledge of the cosmos, this seminal discovery has ushered in a new era of astronomical observation [5]. With the spotlight now turning to space-based GW observatories [6,7], the absence of terrestrial disturbances allows for a more dedicated exploration of the low-frequency GWs [8]. This exciting pursuit carries the potential to reveal hitherto unobserved phenomena, offering profound insights into the nature of our Universe [5].

Space-based GW detection, a largely unexplored domain, marks the next epoch in astrophysics [6]. Pioneering this exciting venture are projects such as the Laser Interferometer Space Antenna (LISA) [9] by the European Space Agency (ESA), with NASA's participation, and Asian projects including Japan's DECi-hertz Interferometer Gravitational wave Observatory (DECIGO) and B-DECIGO [10,11], as well as China's Taiji [12,13] and TianQin [14] missions. Targeting the millihertz frequency band, these endeavors offer a novel perspective for the exploration of diverse astrophysical and cosmological phenomena through the detection of low-frequency GWs [6,15,16]. The scientific goals are broad, with the intent to shed light on the enigmas of massive black hole binaries (MBHBs), extreme-mass-ratio inspirals (EMRIs), continuous waves from galactic binaries (GBs), and the stochastic GW backgrounds produced by the early Universe's myriad of unresolved sources [17].

In the spectrum of potential discoveries, EMRIs hold a unique position. These events, initiated when a compact stellar remnant spirals into a massive black hole (MBH), provide opportunities to investigate the MBH characteristics and the nature of the surrounding environments [18]. EMRIs emit low-frequency GWs throughout their extended inspiral phase, serving as a rich source of information for understanding system physical parameters and the MBH's spacetime geometry [19]. The successful detection and

*Corresponding author: zjcao@bnu.edu.cn
†Corresponding author: renzhx@pcl.ac.cn

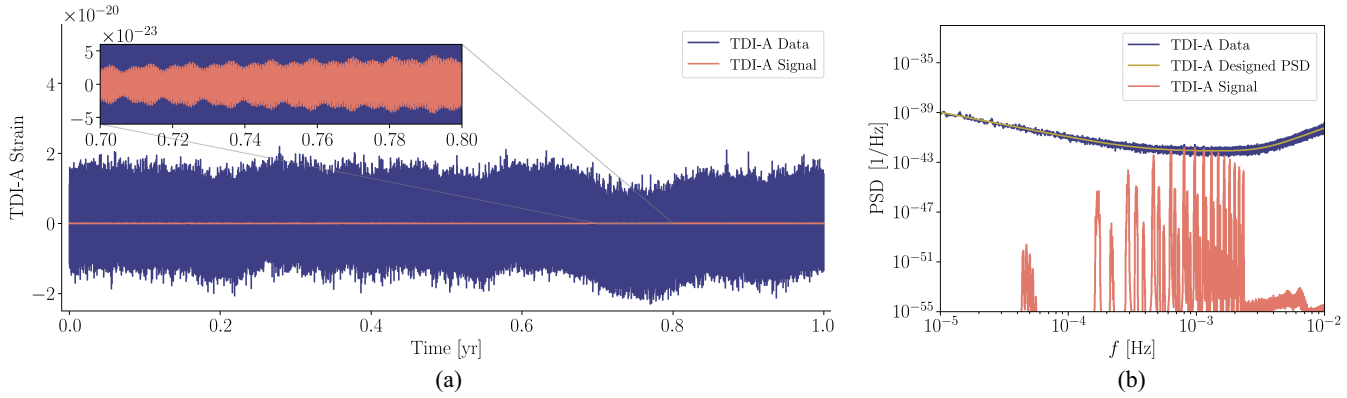

FIG. 1. Visualization of a training data sample. This depicts an EMRI signal from the TDI-A channel spanning one year with an SNR of 70. (a) Time-domain representation of the TDI-A strain, showcasing both the combined data (signal + noise) and the signal. The signal's amplitude is about 3 orders of magnitudes lower than the noise, which makes the detection challenging. (b) Welch PSD of the combined data and the signal; the signal contains lots of modes (peaks), with some reaching the noise level, highlighting the suitability of the frequency domain detection method. The designed detector noise PSD is also presented for reference.

parameter estimation of EMRI signals could provide novel insights into the astrophysics of MBHs and the foundational principles of gravity [20,21]. Traditional methods for EMRI detection, which include both time-domain and time-frequency-domain techniques, have been widely studied in prior research [22–26]. These strategies mainly employ matched filtering [22,25] and the Short Time Fourier Transform [23,24,26]. However, the inherent complexities of EMRI signals present significant obstacles. Characterized by their complex waveform templates, high-dimensional parameter space, and multiple modes within a single waveform, EMRI signals require over $\sim 10^{35}$ templates for matched filtering search [19], resulting in a computationally intensive and time-consuming procedure. An example of single EMRI in both the time and frequency domains can be seen in Fig. 1, showcasing the aforementioned challenges of signal detection. Additionally, EMRI signals are typically faint and buried within detector and confusion noise, necessitating extended observation durations to achieve an adequate signal-to-noise ratio (SNR) for detection [19]. Time-frequency techniques, offering representations in both time and frequency domains, are frequently less sensitive than matched filtering, which limits their ability to identify weak signals [26]. Given these challenges, exploring alternative methods, such as deep learning, becomes crucial for potentially improving the efficiency of EMRI signal detection.

Deep learning, an advanced branch of machine learning, employs neural networks with multiple layers for different types of data. By facilitating the extraction of intricate patterns and representations from large datasets, it has played a crucial role in advancing various fields, from image recognition [27] to natural language processing [28]. Among the numerous architectures, the convolutional neural network (CNN) stands out for its proficiency in handling structured data, such as images and time series, by progressively learning features in a hierarchical manner. Starting with simple features like edges in the initial layers, they gradually combine these to recognize more complex patterns and structures in the deeper layers. This layered approach allows CNNs to automatically recognize and represent intricate details in the data, making them highly effective for tasks like object detection [29] and time-series classification [30]. In the area of GW data analysis, the potential of deep learning, especially CNNs, is becoming increasingly evident. A large amount of studies [31–38] has demonstrated their effectiveness in ground-based GW detection. Beyond signal detection, deep learning methods have been applied to a variety of tasks, including glitch classification [39–41], denoising [42–44], and parameter estimation [45,46]. However, the application of these methods to space-based GW detection is still in its early stages. While there have been some exploratory efforts, such as the adoption of MFCNN [33] to detect MBHB contaminated by confusion noise [47] and the application of dictionary learning to low-SNR space-based binary black hole (BBH) detection [48], these approaches have concentrated on BBH signals. Notably, Zhang *et al.* [49] pioneered the detection of EMRIs using CNN, without incorporating the time delay interferometry (TDI) technique. Therefore, further research is needed to harness the full capabilities of deep learning in space-based GW analysis.

In this paper, we introduce the dilated convolutional neural network for detecting extreme-mass-ratio inspirals (DECODE), an end-to-end model designed for detecting EMRI signals in the frequency domain with an SNR of around 50. As showed in Fig. 2, the model incorporates dilated causal convolutional layers, which expand its receptive field, allowing it to efficiently process data covering an entire year in one pass. We trained our model using synthetic data that considers the TDI-1.5 detector response, accounting for unequal arm lengths. The results

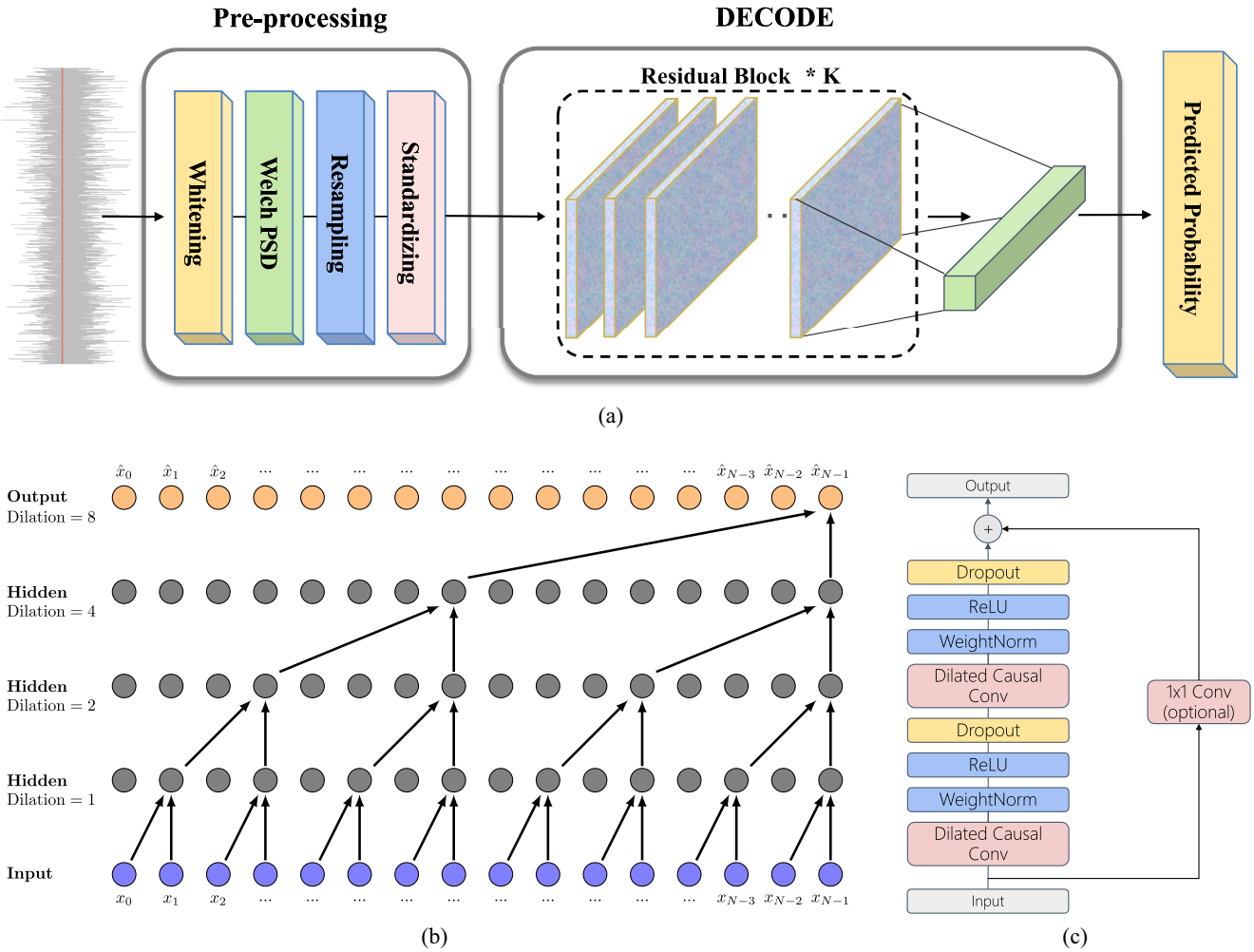

FIG. 2. Comprehensive EMRI detection framework. (a) Depicts the entire EMRI detection process, from initial data preprocessing to the end-to-end DECODE model. (b) Highlights the mechanism of dilated causal convolution with dilation factors of (1, 2, 4, 8) and a kernel size of 2, emphasizing the exponential growth of the receptive field. (c) Detailed architecture of the residual block in DECODE, comprising two dilated causal convolutional layers, weight normalization, ReLU, and dropout layers. A 1×1 convolution is introduced to address any dimension discrepancies between the residual input and output.

are promising: The DECODE detects EMRI signals with a one-year accumulated SNR between 50 and 120, achieving a true positive rate (TPR) of 96.3% with a false positive rate (FPR) of 1%. Notably, our model can evaluate one batch of data samples within 10^{-2} seconds. Visualizations of the model's intermediate outputs highlight its interpretable feature extraction process and its ability to generalize beyond GR. These findings emphasize the potential of DECODE in future space-based GW data analysis.

The remainder of this paper is organized as follows: Sec. II provides a detailed overview of the data generation procedure and outlines the architecture of our proposed model, the DECODE. In Sec. III, we present the results of our EMRI detection experiments, demonstrating the effectiveness of our approach. Finally, Sec. IV concludes the paper with a summary of our findings and a discussion on potential future work based on our findings.

II. METHOD

A. EMRI waveform modeling

Detecting EMRIs has the potential to reveal key astrophysical insights, but modeling their waveform is challenging due to the delicate balance of strong-field GR and gravitational radiation dynamics. Accurately describing EMRIs demands a solution to the self-force problem, which considers the gravitational impact of the smaller compact object on its own motion within the powerful gravitational field of the central MBH [6]. Because the self-force problem is highly nonlinear and defies analytical solutions, researchers have developed approximate waveform models, commonly referred to as kludge models [50,51].

Two commonly used kludge models in EMRI modeling are the analytic kludge (AK) [50] model and the numerical kludge (NK) [51] model. The AK model relies on

post-Newtonian expansions and perturbative calculations to evolve the orbital parameters and generate waveforms quickly. It provides computational efficiency but suffers from dephasing compared to more accurate models, leading to potential inaccuracies in parameter estimation. On the other hand, the NK model incorporates the orbital trajectory computed in curved space using Kerr geodesics and includes radiation reaction effects. Although more accurate, the NK model is computationally more expensive, making EMRI signal detection using this template highly formidable.

To address the limitations of both models, an argumented analytic kludge (AAK) [52–54] model has been proposed. The AAK model combines the computational efficiency of the AK model with improved phasing achieved through a mapping to Kerr geodesic frequencies and self-consistent post-Newtonian evolution. By incorporating self-force information and refining the phasing, the AAK model achieves higher waveform fidelity compared to the AK model while remaining computationally efficient. While its computational efficiency may not be adequate for matched filtering-based signal searches, it is suitable for producing training datasets for deep neural networks (DNNs).

Despite the advancements in kludge waveform modeling, challenges remain. Incorporating second-order self-force effects into the models and refining them for orbits approaching plunge are ongoing areas of research [6]. Nonetheless, these waveform models are crucial for accurately representing the dynamics of EMRIs and enabling the detection, parameter estimation, and data analysis of these elusive astrophysical sources.

B. Data curation

The process of curating training and testing datasets for the identification of EMRI signals using a DNN is a multistep procedure consisting of signal generation, detector response simulation, and preprocessing.

a. Waveform generation. The first step involves the generation of signal templates. The AAK model used for generating these templates is based on Ref. [55]. The waveform, denoted as $h(t) = h_+(t) - ih_\times(t)$, is typically characterized by 14 physical parameters. The parameter space used for sampling the training and testing dataset parameters in this study is detailed in Table I. Here, M and a represent the mass and the spin parameter of the MBH, respectively. The semilatus rectum is denoted by p, while e stands for orbital eccentricity, and ι signifies the orbit's inclination angle from the equatorial plane. $Y = \cos\iota \equiv L_z/\sqrt{L_z^2 + Q}$, where Q is the Carter constant, and L_z is the z component of the specific angular momentum. For the orbital parameters p, e, and Y, the initial values are designated as p_0, e_0, and Y_0, respectively. The polar and azimuthal sky location angles are represented by θ_S and ϕ_S. The orientation of the spin angular momentum vector of the MBH is described by the azimuthal and polar angles θ_K and ϕ_K. These parameters are uniformly sampled for our dataset. It is important to note that the mass of the compact object, denoted by μ, is fixed at $10M_\odot$, and the initial phases for the azimuthal ($\Phi_{\varphi,0}$), polar ($\Phi_{\theta,0}$), and radial ($\Phi_{r,0}$) modes are all manually set to 0, respectively.

TABLE I. Summary of parameter setups in EMRI signal simulation.

Parameter	Lower bound	Upper bound
$\log_{10}(M/M_\odot)$	5	8
a	10^{-3}	0.99
e_0	10^{-3}	0.8
p_0/M	15	25
Y_0	-1	1
SNR	50	120
θ_S	0	π
ϕ_S	0	2π
θ_K	0	π
ϕ_K	0	2π

b. TDI response. The next stage involves simulating the detector's response to these signals. The specific detector configurations utilized in this study are detailed in Table II. For the breathing arm length, we employed the TDI-1.5 technique, which yielded the GW strain of TDI A and E channels, denoted as $h_A(t)$ and $h_E(t)$, respectively. A detailed derivation of this technique can be found in Ref. [56]. Their CUDA-based implementation enable us to calculate the response cost in seconds. The signal is then rescaled according to the desired SNR using the formula:

$$\mathrm{SNR}^2 = (h_A|h_A) + (h_E|h_E). \tag{1}$$

Here, the inner product $(a|b)$ is defined as

TABLE II. Summary of configurations of training and testing dataset.

Parameter	Configuration
Size of training dataset	5000
Size of testing dataset	1000
Cadence	15 s
Duration	1 year
Re-sampled data length N	1024/2048/4096
Arm length L	2.5×10^9 m
Detector orbit	1st order Keplerian orbit
TDI	TDI-1.5
Acceleration noise A_{acc}	3 fm/$\sqrt{\mathrm{Hz}}$
OMS noise A_{oms}	15 pm/Hz

$$(a|b) = 2\int_{f_{\min}}^{f_{\max}} \frac{\tilde{a}^*(f)\tilde{b}(f) + \tilde{a}(f)\tilde{b}^*(f)}{S_n(f)}\mathrm{d}f. \quad (2)$$

In this equation, $f_{\min} = \frac{1}{\text{Duration}} \simeq 3.17 \times 10^{-8}$ Hz and $f_{\max} = \frac{1}{2\cdot\text{Cadence}} = \frac{1}{30}$ Hz. $\tilde{a}(f)$ and $\tilde{b}(f)$ represent the frequency domain signals, and the superscript $*$ denotes the complex conjugate. $S_n(f)$ is the one side noise power spectral density (PSD), which will be specified later.

c. Noise generation. The third step introduces noise to the signal. This noise is modeled as a colored Gaussian noise with a PSD defined by

$$\mathrm{S_n}(f) = 16\sin^2(\omega L)(P_{\mathrm{oms}}(f) + (3 + \cos(2\omega L))P_{\mathrm{acc}}(f)), \quad (3)$$

with

$$\begin{aligned} P_{\mathrm{oms}}(f) &= A_{\mathrm{oms}}^2\left[1 + \left(\frac{2\ \mathrm{mHz}}{f}\right)^4\right]\left(\frac{2\pi f}{c}\right)^2, \\ P_{\mathrm{acc}}(f) &= A_{\mathrm{acc}}^2\left[1 + \left(\frac{0.4\ \mathrm{mHz}}{f}\right)^2\right] \\ &\quad \cdot \left[1 + \left(\frac{f}{8\ \mathrm{mHz}}\right)^4\right]\left(\frac{1}{2\pi f c}\right)^2, \end{aligned} \quad (4)$$

where A_{acc} and A_{oms} are the noise budget of test mass acceleration noise and readout noise coming from the optical metrology system (OMS), $\omega = 2\pi f/c$, L is the arm length of LISA detector, and c is the speed of light. Then the signal is injected into the noise, resulting in the synthetic data. Figure 1 is a showcase of the training data in the time and frequency domain.

d. Whitening and PSD estimation. In the final stage of data curation, the data undergoes several preprocessing steps to prepare it for input into the DNN. The first of these steps is whitening, which serves to remove the frequency-dependent variations in the noise. This process allows the DNN to concentrate on the underlying signal patterns, simplifying the learning task and enhancing the network's ability to detect subtle patterns in the data, thereby improving the overall performance of the EMRI signal identification. Following whitening, the PSD of the data is estimated using Welch's method. The data then undergoes subsampling, where it is resampled onto a log-uniform frequency grid. This step is aimed at reducing the computational load of subsequent analyses by decreasing the number of data points. Three different grid densities are selected as listed in Table II. The final preprocessing step is standardization, which ensures that all input features are on a uniform scale, a fundamental requirement for most deep learning algorithms. This step is crucial in enhancing the learning efficiency of the neural network and improving the overall performance of the model.

C. DECODE

In this work, we introduce the DECODE, a novel architecture for sequence modeling tasks, as illustrated in Fig. 2. The DECODE is inspired by the TCN architecture [57], which has been shown to outperform traditional recurrent architectures across a diverse range of tasks and datasets. The DECODE architecture leverages the strengths of convolutional networks, which have been proven to be highly effective for sequence modeling. It incorporates dilated convolutions, which are a powerful tool for capturing long-range dependencies in sequence data. The causal nature of the DECODE ensures that the model's output at each step is conditioned on all previous steps, making it suitable for tasks that require an understanding of sequential dependencies. While the TCN and other sequence modeling architectures have predominantly been applied to time series data, the DECODE stands out in its application to frequency domain data. Detecting EMRI in the time domain presents challenges due to the extended duration of the signals and their low SNR. As illustrated in Fig. 1(a), the amplitude of the signal is typically 3 orders of magnitude lower than the noise, and the data spans a full year. However, as shown in Fig. 1(b), in the frequency domain, the signal's PSD has lots of peaks, with some even reaching the noise level. Despite this shift from time to frequency domain, the core principles of sequence modeling remain applicable. The DECODE effectively exploits these principles, achieving notable performance in EMRI signal detection.

a. Causal sequence modeling. The DECODE framework is designed for sequence modeling, with a focus on maintaining causality throughout its structure. Central to DECODE's design are two fundamental principles. Firstly, the architecture ensures that the output sequence's length aligns with the input sequence. This alignment is achieved via a 1D-convolutional network design, where each hidden layer matches the length of the input layer. To maintain this length consistency, zero padding of length $(\text{kernel size} - 1)$ is applied. Following this, the architecture emphasizes the causality of the sequence. This is achieved by using causal convolutions, which ensure that the output at a particular time step is convolved only with preceding elements in the previous layer.

b. Dilated convolution. Incorporated into the DECODE architecture, dilated convolutions play a pivotal role in capturing long-range dependencies in sequence data. Drawing inspiration from the WaveNet [58], the DECODE employs dilated convolutions to exponentially expand the receptive field without a significant increase in computational complexity or number of parameters. We provide an illustration in Fig. 2(b). More formally, for a 1D sequence input $\mathbf{x} \in \mathbb{R}^n$ and a filter $f: \{0, \ldots, k-1\} \longrightarrow \mathbb{R}$, the dilated convolution operation F on element s of the sequence is defined as

$$F(s) = (\mathbf{x} *_d f)(s) = \sum_{i=0}^{k-1} f(i) \cdot \mathbf{x}_{s-d\cdot i}, \tag{5}$$

where d is the dilation factor, k is the filter size (i.e. kernel size), and $s - d \cdot i$ accounts for the direction of the past. When $d = 1$, a dilated convolution reduces to a regular convolution. By employing larger dilations, the receptive field of a DECODE is effectively expanded, allowing it to capture long-range dependencies within the sequence data more effectively.

c. Residual connections. Residual connections, another key feature of the DECODE architecture, are designed to facilitate the training of deep networks. These connections, introduced by He *et al.* [59], allow the gradient to flow directly through the network, mitigating the problem of vanishing gradients that often jeopardize deep networks. In the DECODE, a residual block is composed of two dilated causal convolutional layers, with a residual connection skipping over them. If we denote the input to the residual block as $\mathbf{x}$, the output of the block, $\mathbf{y}$, can be computed as

$$\mathbf{y} = \text{Activation}(\mathbf{x} + \mathcal{F}(\mathbf{x})), \tag{6}$$

where $\mathcal{F}(\mathbf{x})$ represents the transformations performed by the dilated causal convolutional layers. This design choice has been shown to improve the performance of deep networks and is a key component of the DECODE architecture. The residual block used in the DECODE model is illustrated in Fig. 2(c). Each block comprises two layers of dilated causal convolution, followed by the rectified linear unit (ReLU) activation function. Weight normalization [60] and dropout [61] are incorporated after each dilated convolution within the residual block.

d. Loss function. In our DECODE model, the output of the residual block has a shape of (H, N), where H represents the hidden size of our model, and N is the length of the input sequence. The last column of this output is then passed through a linear layer to generate the predicted probability for EMRI signal detection. To train the model, we use the cross-entropy loss, a common choice for classification tasks. One of the advantages of using the cross-entropy loss is its ability to accelerate convergence during training, especially when compared to other loss functions like mean squared error [62]. The cross-entropy loss for a binary classification problem is given by

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\mathcal{P}_i) + (1 - y_i)\log(1 - \mathcal{P}_i). \tag{7}$$

In this equation, y_i denotes the actual label, while $\mathcal{P}_i$ is the predicted probability for the ith sample, with n representing the total number of samples in the training dataset. The cross-entropy loss quantifies the divergence between the actual label and the predicted probability.

D. Implementation detail

For waveform generation of training data, we employed `FastEMRIWaveform`[1] [54,55] for EMRI signal creation and `lisa-on-gpu`[2] [56] for GPU-accelerated detector response simulations, which includes TDI. We also integrated additional functionalities from the SciPy library. Our DECODE architecture consists of 10 residual blocks, each with a kernel size of 3 and a hidden size of 128. Developed using the PyTorch framework, known for its computational efficiency and speed, computations were performed on a high-performance computing cluster equipped with NVIDIA Tesla V100 GPUs. The training utilized the Adam optimizer with a learning rate of 2×10^{-4} and a batch size of 64.

III. RESULTS

A. EMRI detection proficiency

Receiver operating characteristic (ROC) curve and the area under the curve (AUC) are essential tools for evaluating the performance of models in binary classification tasks. In the context of our study, where the task is to detect EMRI signals buried in noise, these tools provide valuable insights. The ROC curve, which plots the TPR against the FPR, offers a visual representation of the model's performance across various threshold settings. The AUC, on the other hand, provides a single, overall measure of the model's performance across all thresholds. A model with perfect discrimination has an AUC of 1, while a model performing no better than random guessing has an AUC of 0.5.

In our research, we employ ROC curves as the primary benchmark to quantify the performance of the DECODE. Our test dataset used here is generated like the training datasets; i.e., the waveform parameters are uniformly distributed as shown in Table I but with different SNR range. As depicted in Figs. 3(a)–3(c), we show three separate ROC curves, with each corresponding to a unique input sample length fed into the DECODE. For the specified input lengths of $N = (1024, 2048, 4096)$, the SNR ranges are set at [50, 120], [70, 170], and [100, 240]. The associated AUC values, detailed within the figures, offer quantitative insight into the model's sensitivity in detecting EMRI signals. For clarity in visual representation, especially at lower FPR values, Fig. 3 adopts a logarithmic scale for their axes.

It is noteworthy that our test dataset comprises signals with a duration of one year, achieving twice the SNR compared to the three-month data scenario presented in Ref. [49]. While their study tested models on datasets with $\text{SNR} \in [50, 120]$, we evaluated ours on datasets with

[1]https://github.com/BlackHolePerturbationToolkit/FastEMRIWaveforms.

[2]https://github.com/mikekatz04/lisa-on-gpu.

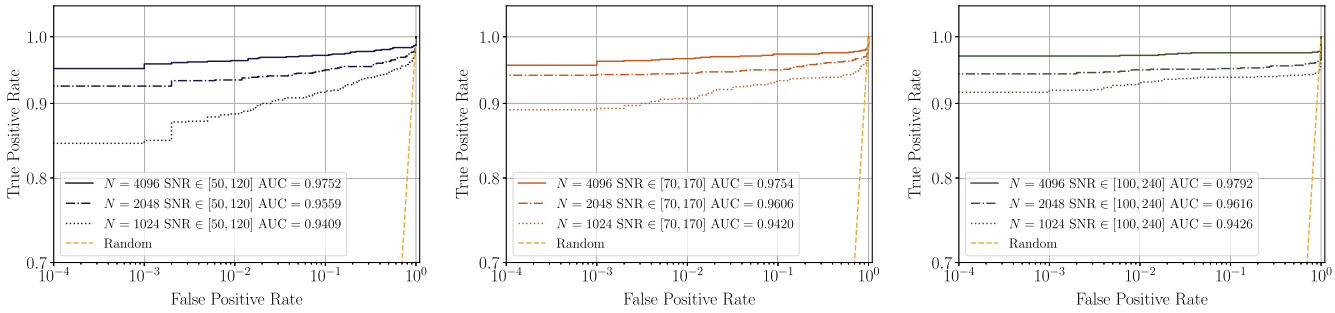

FIG. 3. EMRI detection performance across SNR and N. All subplots depict receiver operating characteristic (ROC) curves for distinct input sample lengths N within specific SNR ranges, presented on a logarithmic scale. Each line style signifies the balance between TPR and FPR for a given sample length, with the area beneath each curve representing the model's efficacy. A reference yellow dashed line indicates the random prediction. The use of logarithmic scales enhances the visibility of performance difference, especially at lower FPR levels. (a) Evaluation for SNR $\in [50, 120]$. (b) Evaluation for SNR $\in [70, 170]$. (c) Evaluation for SNR $\in [100, 240]$.

SNR $\in [100, 240]$. Both datasets, when rescaled for a one-year duration, maintain equivalent SNR values, implying consistent signal amplitudes. Impressively, our model attains a TPR of 97.5% at a FPR of 1% as showcased in Fig. 3(c).

One significant advantage of deep learning methods over matched filtering-based approaches is their speed. Once trained, the model can be rapidly deployed for inference. In our tests, conducted on a single NVIDIA Tesla V100 GPU, our model processed 2000 data samples in approximately 4 seconds, amounting to less than 10^{-2} seconds per sample.

B. EMRI detection efficacy

In Fig. 4, we provide a detailed examination of the DECODE's performance across different physical parameters. Figure 4(a) illustrates the relationship between TPR and SNR. The subfigure clearly demonstrates that as the SNR increases, the TPR increases correspondingly, particularly at the specified FPR thresholds of 0.10 and 0.01.

To gain a deeper understanding of the sensitivity of our model, we introduce the relative amplitude, denoted as $\mathcal{A}$. It is defined as

$$\mathcal{A} = \max_{i \in A,E} \sqrt{\frac{S_h^i(f)}{S_n(f)}}, \tag{8}$$

where S_h^i represents the Welch PSD of waveform h_i. This metric effectively captures the signal's amplitude in the frequency domain. Figure 4(b) plots the TPR against the relative amplitude, at FPRs of 0.1 and 0.01; this subfigure presents the model's proficiency in discerning power exceeds in the frequency domain. Notably, the DECODE can also detect signals that are entirely submerged within the noise.

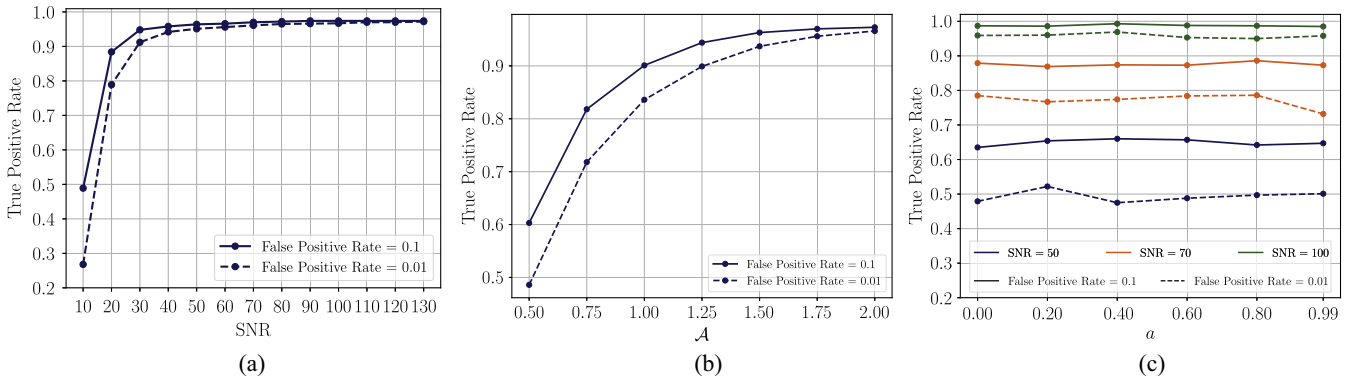

FIG. 4. Detection capability of DECODE across various parameters. (a) Illustrates the TPR as a function of SNR, highlighting the model's capability to detect signals with varying strengths. (b) Showcases the TPR plotted against the relative amplitude $\mathcal{A}$ [defined in Eq. (8)], emphasizing the model's ability to detect power excesses in the frequency domain and detect signals even when they are submerged within the noise. (c) Explores the TPR in relation to the spin parameter a, keeping the MBH mass consistent at $10^6 M_\odot$. This subfigure is evaluated at three distinct SNR levels: 50, 70, and 100, shedding light on the relationship between spin parameters and detection capabilities.

In Fig. 4(c), we evaluate the DECODE's sensitivity to varying spin parameters, while keeping the MBH mass constant at $10^6 M_\odot$. The evaluation, performed at SNR levels of 50, 70, and 100 and FPR thresholds of 0.1 and 0.01, indicates that the model's detection performance is mainly influenced by the SNR. In contrast, the spin parameter appears to have a limited effect on detection, suggesting that the spin parameter contribution to the overall strength of the EMRI signal is relatively minor.

C. Interpretability

CNN-based models are powerful tools for pattern recognition and prediction. Their unique architecture and operational mechanism make them inherently interpretable, a feature that is particularly valuable in interdisciplinary research. CNN-based models learn hierarchical patterns in the data through their convolutional layers, with each layer extracting a set of high-level features from the input data. These features are then used by subsequent layers to understand more complex patterns. This transparent process of feature extraction can be visualized, providing insight into how the network interprets the data and makes predictions.

The activation maps, often used in the context of neural networks, provide a visual representation of the features that the model identifies and emphasizes during its processing. Essentially, they capture the output values or "activations" from various layers or blocks within the network when presented with an input. These maps offer insights into which parts of the input data the model finds significant or relevant for a particular task. In the case of the DECODE, the activation maps generated at the output of each residual block reveal how the model processes and interprets the frequency-domain data of EMRI signals.

The activation maps illuminate the interpretability of the DECODE. By analyzing the outputs of multiple residual blocks, the processes of feature extraction are made transparent. Figure 5 provides a detailed visualization of these maps, demonstrating the ability of the DECODE to distinguish EMRI signals from noise. Specifically, panel *i* of each subfigure depicts activation maps for inputs with an EMRI signal, while panel *iii* depicts the corresponding frequency domain data. These maps emphasize activated neurons in regions that correspond to the frequency components of the signal. In contrast, panel *ii* of each subfigure depicts diminished activations for noise-only samples. The corresponding frequency domain data for these samples is presented in panel *iv*, validating the model's ability at identifying EMRI signals.

D. Generalization ability

Generalization ability is the capacity of a model trained on a specific dataset to perform well on new, untrained data. It indicates how well a model can extrapolate from its training data to make accurate predictions on unknown data. In practical applications, a model will frequently be presented with data that differs from its training set, so this ability is crucial. A model that generalizes well is robust and flexible, ensuring that it does not simply memorize the training data but rather understands inherent patterns and relationships.

In Figs. 5(b) and 5(c), we provide evidence of the generalization capabilities of our model. Even though the model was only trained on AAK waveform datasets, it identified the AK waveform accurately during evaluation with the output probability equal to 1, demonstrating its ability to generalize across various waveform templates. In contrast, the model's successful detection of the XSPEG waveform [63,64], which was formulated using the KRZ metric, demonstrates its generalization ability with respect to various gravitational theories. These results demonstrate the generalization ability of the model, suggesting that it is capable of handling scenarios beyond its training datasets.

IV. CONCLUSION AND DISCUSSION

The detection of EMRIs in gravitational wave astronomy presents a formidable challenge. In this paper, we introduce the DECODE, a state-of-the-art end-to-end DNN model designed for the detection of EMRI signals in the frequency domain. By leveraging dilated causal convolutional layers, the DECODE efficiently processes yearlong data. Our evaluations on synthetic datasets have revealed the model's robustness and efficiency, achieving remarkable detection rates at varied SNR levels. Furthermore, the model exhibits rapid inference capabilities and its ability to generalize beyond its training parameters, although there is still room for future advancement.

The precision of the EMRI detection model is intrinsically related to the precision of the training data. While our current training dataset employs the TDI-1.5 detector response, future developments could benefit from the incorporation of more sophisticated simulations, such as the TDI-2.0 technique. This would provide a more accurate simulation of the detector's response, potentially enhancing the model's applicability in the actual world.

Our current approach primarily focuses on the amplitude information of the EMRI signals. However, the phase information, which has been largely wasted in this research, holds considerable potential. By integrating phase-related features into the model, we could capture more intricate patterns and details of the EMRI signals. This may lead to improved detection rates and lower false alarm rates.

In conclusion, DECODE is a step forward in EMRI detection. Even though there are avenues for improvement, its foundational accomplishments demonstrate its potential as a tool for future space-based GW data analyses.

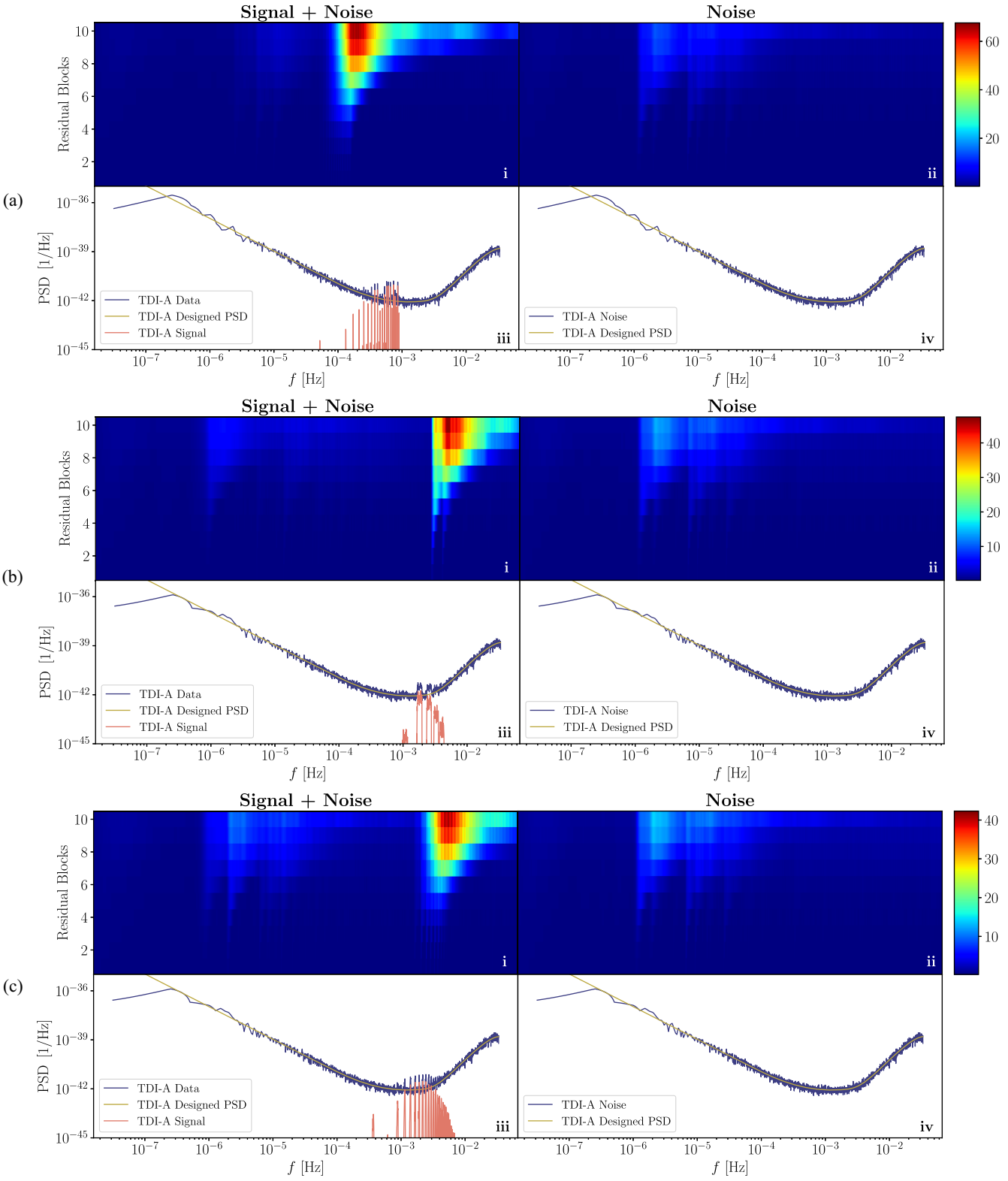

FIG. 5. Interpretability and generalization ability showcase. This figure provides an in-depth visualization of the intermediate outputs from each residual block, demonstrating the model's capability for feature extraction within the frequency domain and its generalization ability to different waveform templates and gravitational theories. For each subfigure, panels *i* and *ii* represent the intermediate results corresponding to the input data samples shown in panels *iii* and *iv*. In contrast to the faint activations in panel *ii*, the noticeable activated neurons in panel *i* indicate the extraction of essential characteristics when a signal is present in the input. (a) AAK waveform. (b) AK waveform. (c) XSPEG waveform.

ACKNOWLEDGMENTS

The research was supported by the Peng Cheng Laboratory and by Peng Cheng Laboratory Cloud-Brain. This work was also supported in part by the National Key Research and Development Program of China Grant No. 2021YFC2203001 and in part by the NSFC (No. 11920101003 and No. 12021003). Z. C. was supported by the "Interdisciplinary Research Funds of Beijing Normal University" and CAS Project for Young Scientists in Basic Research YSBR-006.

[1] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 061102 (2016).
[2] J. Aasi, B. P. Abbott, R. Abbott, T. Abbott *et al.* (The LIGO Scientific Collaboration), Classical Quantum Gravity **32**, 074001 (2015).
[3] F. Acernese, M. Agathos, K. Agatsuma, D. Aisa *et al.*, Classical Quantum Gravity **32**, 024001 (2015).
[4] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 221101 (2016).
[5] M. Bailes, B. K. Berger, P. R. Brady *et al.*, Nat. Rev. Phys. **3**, 344 (2021).
[6] P. Amaro-Seoane, J. Andrews, M. Arca Sedda, A. Askar, Q. Baghi, R. Balasov *et al.*, Living Rev. Relativity **26**, 2 (2023).
[7] A. Sesana, Phys. Rev. Lett. **116**, 231102 (2016).
[8] F. Matichard, B. Lantz, R. Mittleman *et al.*, Classical Quantum Gravity **32**, 185003 (2015).
[9] P. Amaro-Seoane, H. Audley, S. Babak, J. Baker, E. Barausse, P. Bender, E. Berti *et al.*, arXiv:1702.00786.
[10] S. Kawamura, T. Nakamura, M. Ando, N. Seto *et al.*, Int. J. Mod. Phys. D **28**, 1845001 (2019).
[11] S. Kawamura, M. Ando, N. Seto, S. Sato, M. Musha, I. Kawano, J. Yokoyama *et al.*, Prog. Theor. Exp. Phys. **2021**, 05A105 (2021).
[12] W.-R. Hu and Y.-L. Wu, Natl. Sci. Rev. **4**, 685 (2017).
[13] Z. Ren, T. Zhao, Z. Cao, Z.-K. Guo, W.-B. Han, H.-B. Jin, and Y.-L. Wu, Front. Phys. **18**, 64302 (2023).
[14] J. Luo, L.-S. Chen, H.-Z. Duan, Y.-G. Gong, S. Hu, J. Ji, Q. Liu, J. Mei, V. Milyukov, M. Sazhin, C.-G. Shao, V. T. Toth, H.-B. Tu, Y. Wang, Y. Wang, H.-C. Yeh, M.-S. Zhan, Y. Zhang, V. Zharov, and Z.-B. Zhou, Classical Quantum Gravity **33**, 035010 (2016).
[15] P. Auclair *et al.* (The LISA Cosmology Working Group), Living Rev. Relativity **26**, 5 (2023).
[16] E. Barausse, E. Berti, T. Hertog, S. A. Hughes, P. Jetzer, P. Pani *et al.*, Gen. Relativ. Gravit. **52**, 81 (2020).
[17] LISA Science Study Team, LISA science requirements document, Technical Report No. ESA-L3-EST-SCI-RS-001, European Space Agency, 2018.
[18] J. R. Gair, S. Babak, A. Sesana, P. Amaro-Seoane, E. Barausse, C. P. L. Berry, E. Berti, and C. Sopuerta, J. Phys. Conf. Ser. **840**, 012021 (2017).
[19] S. Babak, J. Gair, A. Sesana, E. Barausse, C. F. Sopuerta, C. P. L. Berry, E. Berti, P. Amaro-Seoane, A. Petiteau, and A. Klein, Phys. Rev. D **95**, 103012 (2017).
[20] A. J. K. Chua, S. Hee, W. J. Handley, E. Higson, C. J. Moore, J. R. Gair, M. P. Hobson, and A. N. Lasenby, Mon. Not. R. Astron. Soc. **478**, 28 (2018).
[21] W.-B. Han and X. Chen, Mon. Not. R. Astron. Soc. **485**, L29 (2019).
[22] J. R. Gair, L. Barack, T. Creighton, C. Cutler, S. L. Larson, E. S. Phinney, and M. Vallisneri, Classical Quantum Gravity **21**, S1595 (2004).
[23] J. Gair and L. Wen, Classical Quantum Gravity **22**, S1359 (2005).
[24] L. Wen and J. R. Gair, Classical Quantum Gravity **22**, S445 (2005).
[25] J. Gair and G. Jones, Classical Quantum Gravity **24**, 1145 (2007).
[26] J. R. Gair, I. Mandel, and L. Wen, J. Phys. Conf. Ser. **122**, 012037 (2008).
[27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Proc. IEEE **86**, 2278 (1998).
[28] J. Devlin, M. Chang, K. Lee, and K. Toutanova, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, edited by J. Burstein, C. Doran, and T. Solorio (Association for Computational Linguistics, Minneapolis, MN, USA, 2019), Vol. 1, pp. 4171–4186.
[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Los Alamitos, CA, USA, 2016), pp. 779–788.
[30] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Vancouver, BC, Canada, 2023), pp. 16133–16142.
[31] D. George and E. A. Huerta, Phys. Rev. D **97**, 044039 (2018).
[32] H. Gabbard, M. Williams, F. Hayes, and C. Messenger, Phys. Rev. Lett. **120**, 141103 (2018).
[33] H. Wang, S. Wu, Z. Cao, X. Liu, and J.-Y. Zhu, Phys. Rev. D **101**, 104003 (2020).
[34] P. G. Krastev, Phys. Lett. B **803**, 135330 (2020).
[35] M. López, I. Di Palma, M. Drago, P. Cerdá-Durán, and F. Ricci, Phys. Rev. D **103**, 063011 (2021).
[36] V. Skliris, M. R. K. Norman, and P. J. Sutton, arXiv:2009.14611.
[37] R. Qiu, P. G. Krastev, K. Gill, and E. Berger, Phys. Lett. B **840**, 137850 (2023).

[38] A. Ravichandran, A. Vijaykumar, S. J. Kapadia, and P. Kumar, arXiv:2302.00666.
[39] R. E. Colgan, K. R. Corley, Y. Lau, I. Bartos, J. N. Wright, Z. Márka, and S. Márka, Phys. Rev. D **101**, 102003 (2020).
[40] M. Cavaglia, K. Staats, and T. Gill, Comput. Phys. Commun. **25**, 963 (2018).
[41] M. Razzano and E. Cuoco, Classical Quantum Gravity **35**, 095016 (2018).
[42] W. Wei and E. A. Huerta, Phys. Lett. B **800**, 135081 (2020).
[43] Z. Ren, H. Wang, Y. Zhou, Z.-K. Guo, and Z. Cao, Mach. Learn. Sci. Tech. **5**, 015046 (2024).
[44] T. Zhao, R. Lyu, H. Wang, Z. Cao, and Z. Ren, Commun. Phys. **6**, 212 (2023).
[45] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, Nat. Phys. **18**, 112 (2022).
[46] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, Phys. Rev. Lett. **127**, 241103 (2021).
[47] W.-H. Ruan, H. Wang, C. Liu, and Z.-K. Guo, Phys. Lett. B **841**, 137904 (2023).
[48] C. Badger, K. Martinovic, A. Torres-Forné, M. Sakellariadou, and J. A. Font, Phys. Rev. Lett. **130**, 091401 (2023).
[49] X.-T. Zhang, C. Messenger, N. Korsakova, M. L. Chan, Y.-M. Hu, and J.-d. Zhang, Phys. Rev. D **105**, 123027 (2022).
[50] L. Barack and C. Cutler, Phys. Rev. D **69**, 082005 (2004).
[51] S. Babak, H. Fang, J. R. Gair, K. Glampedakis, and S. A. Hughes, Phys. Rev. D **75**, 024005 (2007).
[52] A. J. K. Chua and J. R. Gair, Classical Quantum Gravity **32**, 232002 (2015).
[53] A. J. K. Chua, C. J. Moore, and J. R. Gair, Phys. Rev. D **96**, 044005 (2017).
[54] A. J. K. Chua, M. L. Katz, N. Warburton, and S. A. Hughes, Phys. Rev. Lett. **126**, 051102 (2021).
[55] M. L. Katz, A. J. K. Chua, L. Speri, N. Warburton, and S. A. Hughes, Phys. Rev. D **104**, 064047 (2021).
[56] M. L. Katz, J.-B. Bayle, A. J. K. Chua, and M. Vallisneri, Phys. Rev. D **106**, 103001 (2022).
[57] S. Bai, J. Z. Kolter, and V. Koltun, in *Proceedings of the 6th International Conference on Learning Representations (ICLR)* (ICLR, Vancouver, BC, Canada, 2018), https://openreview.net/forum?id=BJEX-H1Pf.
[58] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, in *Proceedings of the 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)* (ISCA, Sunnyvale, CA, USA, 2016), p. 125, arXiv:1609.03499.
[59] K. He, X. Zhang, S. Ren, and J. Sun, in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Los Alamitos, CA, USA, 2016), pp. 770–778.
[60] T. Salimans and D. P. Kingma, in *Advances in Neural Information Processing Systems (NeurIPS)*, edited by D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., Barcelona, Spain, 2016), Vol. 29, p. 901.
[61] N. Srivastava, G. Hinton, A. Krizhevsky, and I. Sutskever, J. Mach. Learn. Res. **15**, 1929 (2014).
[62] C. M. Bishop, *Pattern Recognition and Machine Learning*, Softcover Reprint of the Original 1st ed., Information Science and Statistics (Springer, New York, New York, NY, 2016).
[63] S. Xin, W.-B. Han, and S.-C. Yang, Phys. Rev. D **100**, 084055 (2019).
[64] C. Zhang, W.-B. Han, and S.-C. Yang, Commun. Theor. Phys. **73**, 085401 (2021).