



University
of Glasgow

Zhang, X.-T., Messenger, C., Korsakova, N., Chan, M. L., Hu, Y.-M. and Zhang, J.-d. (2022) Detecting gravitational waves from extreme mass ratio inspirals using convolutional neural networks. *Physical Review D*, 105(12), 123027. (doi: [10.1103/PhysRevD.105.123027](https://doi.org/10.1103/PhysRevD.105.123027)).

This is the Author Accepted Manuscript.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/294564/>

Deposited on: 27 March 2023

Enlighten – Research publications by members of the University of Glasgow
<http://eprints.gla.ac.uk>

Detecting Gravitational-waves from Extreme Mass Ratio Inspirals using Convolutional Neural Networks

Xue-Ting Zhang,¹ Chris Messenger,² Natalia Korsakova,³ Man Leong Chan,⁴ Yi-Ming Hu,^{1,*} and Jing-dong Zhang¹

¹*MOE Key Laboratory of TianQin Mission, TianQin Research Center for Gravitational Physics & School of Physics and Astronomy, Frontiers Science Center for TianQin, Gravitational Wave Research Center of CNSA, Sun Yat-sen University (Zhuhai Campus), Zhuhai 519082, China*

²*SUPA, School of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ, United Kingdom*

³*ARTEMIS, Observatoire de la Côte d'Azur, Boulevard de l'Observatoire, 06304 Nice, France*

⁴*Department of Applied Physics, Fukuoka University, Nanakuma 8-19-1, Fukuoka 814-0180, Japan*

(Dated: February 16, 2022)

Extreme mass ratio inspirals (EMRIs) are among the most interesting gravitational wave (GW) sources for space-borne GW detectors. However, successful GW data analysis remains challenging due to many issues, ranging from the difficulty of modeling accurate waveforms, to the impractically large template bank required by the traditional matched filtering search method. In this work, we introduce a proof-of-principle approach for EMRI detection based on convolutional neural networks (CNNs). We demonstrate the performance with simulated EMRI signals buried in Gaussian noise. We show that over a wide range of physical parameters, the network is effective for EMRI systems with a signal-to-noise ratio larger than 50, and the performance is most strongly related to the signal-to-noise ratio. The method also shows good generalization ability towards different waveform models. Our study reveals the potential applicability of machine learning technology like CNNs towards more realistic EMRI data analysis.

I. INTRODUCTION

The successful detection of Gravitational Waves (GWs) has opened a new window of understanding the Universe [1–4]. However, a wide band of GW frequencies is still inaccessible for observation. For example, systems like Double White Dwarfs (DWDs) [5, 6], Massive Binary Black Holes (MBBHs) [7, 8], stellar-mass Binary Black Holes (sBBHs) [9, 10], Extreme Mass Ratio Inspirals (EMRIs) [11, 12], and stochastic gravitational-wave background [13, 14] are expected to emit GW signals in the m Hz frequency band. The future detections of these signals can help us better understand the nature of gravity [e.g. 15, 16] and the Universe [e.g. 17, 18]. In order to properly understand the physics and astronomy behind GW events, sophisticated and efficient algorithm to perform data analysis plays a significant role.

Space-borne GW missions are capable of detecting GW signals in the m Hz band [19, 20]. Of all expected GW sources, EMRIs are among the most interesting ones. A typical EMRI system is composed of a stellar-mass Compact Object (CO) and a Massive Black Hole (MBH) (10^4 – $10^7 M_\odot$). Such systems are believed to be formed in the center of galaxies. The detection of EMRIs, among other things, can help us better understand the growth mechanism of MBHs, and also put a more stringent constraint on their population properties, like the mass function and the distribution of the surrounding COs [21, 22]. In addition, the EMRI signal can last for thousands to millions of cycles in the m Hz frequency band. This feature empowers the EMRI systems to be ideal laboratories to study gravity in a strong regime. [16, 23–26].

Despite its great scientific potential, there remain great challenges for the end-to-end data analysis of EMRI signals. For example, the ideal EMRI waveform should consider the effects of self-force. Some progress on self-force waveform has been made recently, such as [27] for an eccentric orbit to the second-order, and [28] for a generic orbit to the adiabatic order. However, a fast and accurate EMRI waveform for the most general case is yet to be implemented. Most of the widely used waveform models are expected to quickly dephase from the physical waveform, which restricts its applicability in EMRI data analysis. The computational cost of EMRI waveforms is also usually very high, since it involves orbit integration over a period of months to years. It has been estimated that in order to implement a classical signal search with matched filtering, a bank template containing at least 10^{40} templates would be needed [29].

Some attempts have been made to explore efficient data analysis algorithms for EMRIs. Both template-based algorithms and template-free methods have been proposed to detect the EMRI signals. The former includes the semi-coherent method [29] and \mathcal{F} -statistic algorithm [30, 31], and the latter includes the time-frequency algorithm [32–35]. On the parameter estimation side, methods like Metropolis-Hastings search [36, 37], parallel tempering MCMC [38], and Gaussian process [39, 40] have been implemented.

In recent years, machine learning has gained huge popularity in various complex tasks, such as computer vision [41–46], speech recognition [47], natural language process [48], and some have been successfully applied to GW astronomy [49–62]. Deep learning is a sub-field of machine learning that is based on learning several levels of data representations [63], corresponding to a hierarchy of features, factors, or concepts, where higher-

* huyiming@mail.sysu.edu.cn

level concepts are defined from lower-level ones. The Convolutional Neural Network (CNN) is a deep learning algorithm, which has been applied in many fields because it can automatically extract data features and easily process various high-dimensional data, such as image classification, target detection, and part-of-speech tagging. In GW data analysis, CNNs has been used to detect GW signals detection, such as Binary Black Hole (BBH) mergers [49–53], Binary Neutron Star (BNS) mergers [54, 55], supernova explosions [56], and continuous gravitational waves [57, 58]. After being properly trained, deep learning methods also demand low computation cost, which improves efficiency in an end-to-end detection system. In this work, we use TianQin, a spaceborne GW detector, as a reference, to explore the EMRI signal detection with a CNN.

This paper is organized as follows. Section II introduces the characteristics of EMRI and its waveform. Section III provides the architecture of a CNN, and discusses how to prepare datasets. Section IV describes the search procedure. The results are presented in Section V. Conclusions and discussions are provided in Section VI.

II. BASICS OF EMRI DETECTION

A. Basic astronomy of EMRIs

Galactic nuclei are extreme environments for stellar dynamics [64]. The stellar densities can be as high as $10^6 \text{ M}_\odot \text{ pc}^{-3}$ and its velocity scatter can exceed 1000 km s^{-1} . Within a galactic nucleus, a large number of COs, including stellar-mass Black Holes (sBHs), White Dwarfs (WDs), and Neutron Stars (NSs), are expected to exist. They are considered to be within the “loss cone” if their orbit would lead to their capture by the central MBH, leading to the formation of EMRIs. COs outside the loss cone can fill the loss cone through dynamical processes like two-body relaxation [64].

In addition to this traditional channel, some new formation channels of EMRI systems have also been proposed. It has been conjectured that a CO can be captured by the accretion disk around central MBH(i.e., Active Galactic Nuclei (AGN) disk). Effects like wind effects, density wave generation, and dynamic friction can assist the inward migration of COs, which further boosts the formation of EMRIs [65, 66].

Although EMRIs are speculated to exist, no direct observational evidence is available yet (c.f.[67]). Therefore, currently, the population properties of EMRIs are derived from theoretical modeling. However, components within the theoretical modeling, like the mass and spin distribution of the central MBH, the erosion of the stellar cusp, the M - σ relation, etc. are uncertain, which lead to a very uncertain prediction on EMRI population [68]. Indeed, different assumptions on EMRI population lead to orders of magnitude difference in terms of detection rate by TianQin [12]. Throughout this analysis, an model is

adopted to simulate EMRI population. To be exact, we adopt the model M12 from [68].

B. Waveform models of EMRIs

Theoretical modelling of EMRI waveforms is essential for the successful detection of GWs from these systems. To accurately compute the EMRI waveform, one has to consider the effects of the gravitational self-force, which leads to a deviation of CO orbit from the test particle geodesics under the Kerr metric of the MBH [69]. Unfortunately, waveform simulations with numerical relativity cannot be used for EMRI due to prohibitively high computational costs caused by the very high mass ratio of such systems. Although some progress has been made in adding self-force into the waveform generation [27, 28, 69–73], a fully realistic solution for EMRI is still yet to be achieved. Nevertheless, there are methods that can generate EMRI waveforms. For example, the Teukolsky-based waveform [74] and Numerical Kludge (NK) waveform [75] solves the Teukolsky equation with the geodesic under Kerr metric. Such methods suffer from low efficiency, while the Analytic Kludge (AK) method that adopts the post-Newtonian (PN) formula can achieve waveform generation in a much faster way [76]. The Augmented Analytic Kludge (AAK) waveform [77] can reach an accuracy similar to NK with the generating speed of AK. For our study, since a large number of simulated waveforms are needed for the training, testing, and validation of our CNN method, we focus on the analytic family of waveform models and concentrate on both AK and AAK waveforms. We remark that during the writing of this manuscript, a new and fast waveform has been implemented [78], which could be used to improve efficiency in the future, but we do not expect any significant difference in our findings. Besides, [78] adopts the Schwarzschild Black Hole (BH) assumption, while we would like to also study the impact of MBH spin.

The AK method first calculates the orbit of CO through post-Newtonian equations, considering the radiation back-reaction and the Lense-Thirring effect. The overall evolution of the CO orbit is governed by three fundamental frequencies, namely the orbital angular frequency f_o , the periapsis precession frequency f_p , and the Lense-Thirring precession frequency f_{LT} . The waveform is then calculated based on quadrupolar approximation. Notice that the overall setup of a waveform is not self-consistent, hence the name of a “kludge”. However, this choice of formulation allows a fast way of producing waveforms at the cost of its accuracy.

Under the transverse and traceless gauge, two polarizations of AK waveform are defined via a n -harmonic

waveform which is a function of physical parameters ¹:

$$h_+ = \sum_n - \left[1 + (\hat{L} \cdot \hat{n})^2 \right] [a_n \cos 2\gamma - b_n \sin 2\gamma] \\ + \left[1 - (\hat{L} \cdot \hat{n})^2 \right] c_n, \\ h_\times = \sum_n 2(\hat{L} \cdot \hat{n}) [b_n \cos 2\gamma + a_n \sin 2\gamma],$$

where \hat{L} is the orbital angular momentum, \hat{n} is the position of the source, and γ is an azimuthal angle measuring the direction of pericenter. The coefficients (a_n, b_n, c_n) with n ranges from 1 to 5, are determined by the eccentricity and mean anomaly orbital phase [76]. The orbital phase is determined by three fundamental frequencies f_o , f_p , and f_{LT} . More details on waveform generation can be found in [12, 76]. Notice that since the self-force can not be properly incorporated by the AK waveform, the spin parameter for the CO is not included.

The AK waveform quickly deviates from physical systems, due to intrinsic inconsistency. Under the AK model, instead of using the physical parameters, one can adopt the effective parameters to evolve the orbit and produce a waveform similar to the NK model. The NK model is more accurate as it combines PN orbital evolution with Kerr geodesics. Chua et al. implemented the AAK waveform by establishing the link between the physical parameter and the effective parameter. To obtain such a link, one can evolve the NK waveform for a short period and determine the mapping relationship to derive the correct fundamental frequencies f_o , f_p , and f_{LT} [79]. In this way, the AAK waveform can achieve both the accuracy of NK and the speed of AK.

C. The TianQin mission

For a given gravitational wave detector, one can use the antenna pattern to describe its response to a source at a given location, polarization, and frequency [80–82]. A typical EMRI signal lasts for a long time. If we observe from the point of view of the detector, then a given source would change in location as well as in polarisation. One needs to know the orbit of the spacecraft to derive the overall evolution of the signal as seen at the detector. In this work, we focus our attention on the proposed TianQin mission [19, 83] for the following analysis.

The TianQin mission is designed to be a constellation with three satellites, which accommodate test masses in free fall and use drag-free control to follow these test masses. The satellites share an Earth-central orbit and form a regular triangle. They are connected by laser links and can form Michelson interferometers to detect GW. The geocentric nature of the TianQin orbit makes it

adopts a “3-month on 3-month off” work scheme, that the detectors would operate continuously for three months followed by three months of break. Each pair of neighboring laser arms can form interference on the vertex satellite, whose noise can be described by the power spectral density (PSD):

$$S_n(f) = \frac{1}{L_{\text{arm}}^2} \left[\frac{4S_a}{(2\pi f)^4} \left(\frac{1 + 10^{-4} \text{ Hz}}{f} \right) + S_x \right] \\ \left[1 + 0.6 \left(\frac{f}{f_*} \right)^2 \right], \quad (1)$$

where f is the frequency, $S_a = 1 \times 10^{-30} \text{ m}^2 \text{ s}^{-4}/\text{Hz}$ describes the acceleration noise, $S_x = 1 \times 10^{-24} \text{ m}^2/\text{Hz}$ describes the positional noise, and $f_* = \frac{c}{2\pi L_{\text{arm}}}$ is the transfer frequency with c being the speed of light and $L_{\text{arm}} = \sqrt{3} \times 10^8 \text{ m}$ is the arm length.

For the low frequency, where $f < f_*$, the three interferometers can be combined into two orthogonal channels, corresponding to the two GW polarisations. So the recorded GW signal can be expressed as:

$$h_{I,II}(t) = \frac{\sqrt{3}}{2} \left(h_+(t) F_{I,II}^+(\theta_S, \phi_S, \psi_S) \right. \\ \left. + h_\times(t) F_{I,II}^\times(\theta_S, \phi_S, \psi_S) \right), \quad (2)$$

where $F_{I,II}^+$ and $F_{I,II}^\times$ are the antenna pattern for both plus and cross polarisations, and

$$F_I^+(\theta_S, \phi_S, \psi_S) = \frac{1}{2} (1 + \cos^2 \theta_S) (\cos 2\phi_S \cos 2\psi_S \\ - \cos \theta_S \sin 2\phi_S \sin 2\psi_S). \quad (3) \\ F_I^\times(\theta_S, \phi_S, \psi_S) = \frac{1}{2} (1 + \cos^2 \theta_S) (\cos 2\phi_S \sin 2\psi_S \\ + \cos \theta_S \sin 2\phi_S \cos 2\psi_S).$$

Here θ_S and ϕ_S are the sky location of the source, and ψ_S is the polarization angle. All quantities are defined in the detector frame, which makes it time-dependent [9].

The two effective channels differ only in polarisation angles by $\pi/4$, which means

$$F_{II}^+(\theta_S, \phi_S, \psi_S) = F_I^+(\theta_S, \phi_S, \psi_S - \pi/4), \\ F_{II}^\times(\theta_S, \phi_S, \psi_S) = F_I^\times(\theta_S, \phi_S, \psi_S - \pi/4). \quad (4)$$

Finally, a correction term is added to account for the Doppler effect induced by the orbital motion of the detectors [9].

$$\Phi_{\text{Doppler}} = 2\pi f(t) \times R \times \sin \bar{\theta}_S \times \cos (\bar{\Phi}(t) - \bar{\phi}_S), \quad (5)$$

where f is the frequency of GW, $R = 1\text{AU}$ and barred symbols denote parameters in Solar System Barycentre frame, with $\bar{\Phi}(t) = \bar{\phi}_0 + 2\pi t/T$, $T = 1\text{year}$ is Earth’s orbital period around the Sun, and $\bar{\phi}_0$ is the initial location of TianQin at time $t = 0$. Notice that in higher

¹ The details of the parameters adopted are presented in Table II

frequencies when $f \geq f_*$, the long-wavelength approximation breaks down and the two channels are no longer uncorrelated, hence an extra correction term is needed in Equation 2 [81, 82]. Fortunately, EMRI signals we are interested in, are located mostly within the low-frequency range. This is because the highest orbital frequency of CO is determined by the last stable orbit (LSO), where $\nu_{\text{LSO}} = 1/(2\pi M) \times 6^{3/2}$. Given a typical MBH with $10^5 M_\odot$, $\nu_{\text{LSO}} = 0.04 < f_* = 0.28$ Hz. Therefore, we stick with the long-wavelength approximation throughout the work.

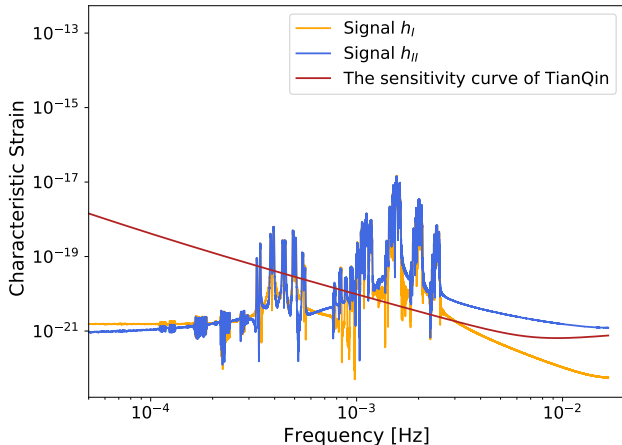


FIG. 1. An example EMRI signals compared with the sensitivity curve of TianQin. A total length of 3 months observation time is assumed.

As an illustration, we show in Figure 1 an example EMRI signal (on both channels) on the TianQin sensitivity curve. This EMRI system consists of a $10 M_\odot$ BH and MBH with $4.5 \times 10^6 M_\odot$ and a spin of 0.97. The Signal-to-Noise Ratio (SNR) of this signal equals 50.

We define the overall SNR as the root sum square of inner products of waveform with itself across two channels of TianQin,

$$\begin{aligned} \rho &= \sqrt{\sum_{i=I,II} \langle h_i | h_i \rangle} \\ &= \sqrt{4\Re \sum_{i=I,II} \int_0^\infty \frac{\tilde{h}_i(f) \tilde{h}_i^*(f)}{S_n(f)} df} \\ &\approx \sqrt{4 \sum_{i=I,II} \int_{f_{\text{low}}}^{f_{\text{high}}} \left(\frac{\tilde{h}_i(f)}{\sqrt{S_n(f)}} \right) \left(\frac{\tilde{h}_i(f)}{\sqrt{S_n(f)}} \right)^* df}. \end{aligned} \quad (6)$$

Here we use “ $\langle g|h \rangle$ ” to represent the inner product between the two frequency domain waveforms $\tilde{g}(f)$ and $\tilde{h}(f)$. Notice that in the last line, the SNR in a channel is simply the integration of squared modulus of whitened signal over the frequency range between $f_{\text{low}} = 10^{-4}$ Hz and $f_{\text{high}} = 1$ Hz, which are sensitivity frequency band of TianQin.

III. CONVOLUTIONAL NEURAL NETWORKS FOR DETECTION

A. Data preparation

For the detection of the signal in noise we are going to use CNNs. In order to train and verify CNN properly, one needs to divide the data into three groups, namely *training data*, *validation data*, and *testing data*. In these datasets, half of them are signal+noise samples, and the rest are noise-only samples. Using training data, the CNN learns how to classify samples. Validation data is incorporated in the training process to verify that the CNN is learning correctly. Finally, testing data can test the efficiency of the trained CNN.

In our case, we aim to accomplish the task of classifying the data into two categories: with signal, or without signal. One can express the data d as the addition of random Gaussian noise n and the GW signal h (in the case it is present in the data):

$$d(t) = \begin{cases} h(t) + n(t), & \text{if signal is present} \\ n(t), & \text{if there is no signal.} \end{cases} \quad (7)$$

The raw output of TianQin is three correlated time domain data channels. After some preprocessing they can be reduced to two channels with uncorrelated noise, plus a channel with a response to astrophysical signals highly suppressed. This process is implemented by the Time Delay Interferometry (TDI) [84, 85]. Out of simplicity, we do not implement TDI, but rather treat TianQin as two orthogonal Michelson interferometers. In Figure 2, we illustrate a data sample that contains an EMRI signal, using the parameters demonstrated in Figure 1.

We prepare the generation of noise and signal separately. Due to the “3-month on 3-month off” work scheme of TianQin, the longest chunk of continuous data lasts only three months. To avoid the complexity introduced by considering the data gap issue, we focus our analysis on data length of three months (or 7864320 seconds). By adopting the long-wavelength approximation, it is also sufficient to assume the sample rate of 1/30 Hz, which leads to the data size of 262144 samples. The parameters are summarised in Table I.

TABLE I. The setting of data simulation.

Symbol	Physical Meaning	Range
T	observation time	7864320 seconds
f_s	sample rate	1/30 Hz
N_t	a data size	262144

The detector noise was assumed to be Gaussian and was generated by sampling from the one-sided PSD given by Equation 1. For the simulation of an EMRI waveform signal, we used both AK and AAK models. For different purposes that we will detail in the following sections,

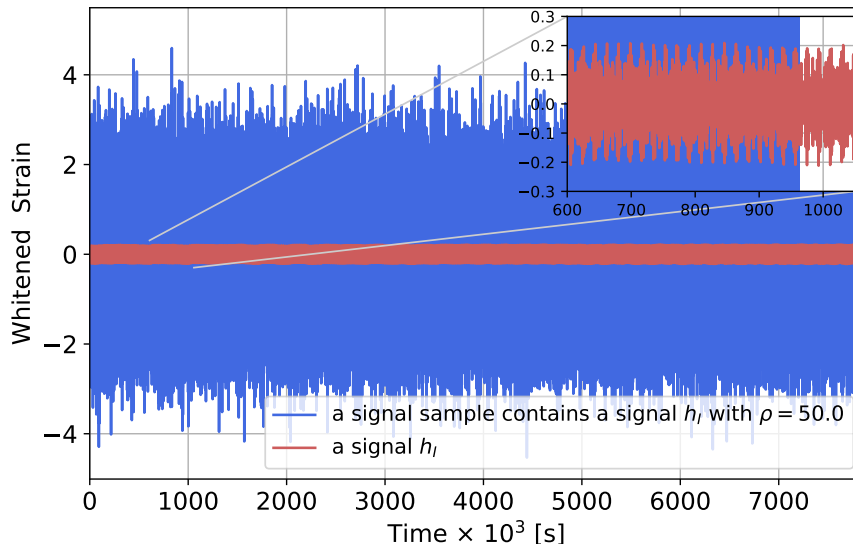


FIG. 2. An example of whitened data in channel I in comparison with signal h_I alone. For this event, the SNR is set to be 50. We draw the reader's attention to the difference in scale for the noise and the signal.

TABLE II. Summary of physical parameters and their meaning under the analytic kludge model.

Symbol	Physical Meaning	distribution
M	MBH mass	uniform in $\ln M$ over $[10^4 M_\odot, 10^7 M_\odot]$
μ	CO mass	$10 M_\odot$
a	MBH spin magnitude, defined as $ \vec{S} /M^2$, with \vec{S} being the MBH spin angular momentum	uniform $[0,1]$
e_{iso}	orbital eccentricity at plunge	uniform $[0,0.2]$
Φ_0	mean anomaly at plunge	uniform $[0, 2\pi]$
α_0	azimuth angle for orbital angular momentum \vec{L} at plunge	uniform $[0, 2\pi]$
λ	angle between \vec{L} and \vec{S}	$\cos(\lambda) = \text{uniform}[-1,1]$
γ_0	angle between $\vec{L} \times \vec{S}$ and pericenter at plunge	uniform $[0, 2\pi]$
ν_{iso}	orbital frequency of the last stable orbit	depend on e_{iso}
θ_S	source direction's polar angle	uniform $[0, 2\pi]$
ϕ_S	azimuthal direction of source	$\cos(\phi_S) = \text{uniform}[-1, 1]$
θ_K	the polar angle of MBH's spin	uniform $[0, 2\pi]$
ϕ_K	azimuthal direction of MBH's spin	$\cos(\phi_K) = \text{uniform}[-1, 1]$
t_c	merger time of a CO plunges into MBH	fixed to 7864320 seconds
D_L	luminosity distance to source	irrelevant as will be rescaled to SNR with uniform $[50,120]$

we generate 7 groups of signals, whose properties are explained in Table IV, and the shared default parameter distributions are listed in Table II. The data are whitened in the frequency domain. When data are converted back to the time domain, we apply a Tukey window function with the coefficient $\alpha = 1/4$ to avoid the spectrum leakage.

Notice that Table II does not reflect an astrophysical model, and one might argue that an astrophysically motivated distribution for EMRI parameters could be more realistic. However, an astrophysically motivated distribution is not necessarily the best choice for training the CNN. The reason for that is that the astrophysically motivated models are subject to huge uncertainty. Therefore

sticking with any specific model runs a risk of biasing the network.

SNR of a signal is inversely proportional to the distance. When assessing the performance of a CNN, it is convenient to compare signals with similar SNRs than with similar distances. Therefore, we do not fix distribution of distance D_L in Table II. The choice of SNR range is determined by balancing the hardware requirement between data storage, waveform generation speed, and I/O speed in training.

B. Mathematical model

CNN is a specific realization of Neural Network (NN) which features the inclusion of convolution operations [63, 86]. In NN, each “neuron” represents a non-linear mapping or an affine transformation function. The non-linearity is achieved by an activation function. A typical realization of a CNN includes convolution layers, pooling layers, and fully-connected layers (or dense layers) [87]. It can be treated as a classifier. Such a CNN composed of multiple neurons can be understood as a non-linear function $f_w(\cdot)$ that maps the input space of the data to the output space:

$$y_i = f_w(d_i), \quad (8)$$

where y_i is the output probability for the i -th data d_i to belong to a certain class, w are weights of the CNN.

Based on training data a CNN can learn features of the input data by updating the weights associated with different neurons by minimizing the difference between the output prediction and training data labels. The size of our training data is large, therefore we cannot optimize weights with all the data at once and need to adopt mini-batching. The training process is divided into multiple epochs. For each epoch, the weights are updated by the optimizer algorithm. When the training process converges, the CNN can classify new data into different categories.

To achieve the reliable ability to classify, one needs, first, to define the loss function $L(y_{\text{true}}, y_{\text{pred}})$, which quantifies the difference between the CNN output and labels in the training process. This could be done by adopting the binary cross-entropy function:

$$L = \sum_{j=1}^N \left[-y_{\text{true}}^j \log(y_{\text{pred}}^j) - (1 - y_{\text{true}}^j) \log(1 - y_{\text{pred}}^j) \right], \quad (9)$$

where N is the total number of training data samples, for the j -th data instance. The noise-only samples are labeled as $y_{\text{true}}^j = 0$, and samples containing signals are labeled as $y_{\text{true}}^j = 1$. The output of CNN y_{pred}^j represents the assigned probability for each category. The loss function is minimized when the predicted value for the label class matches the one from the training data with the highest probability. In the training process, the weights are updated through the *Nadam* algorithm, such that the loss function can be minimized by following the adaptive gradient descent with Nesterov momentum [88]. During training, validation data is used to test the trained CNN after each epoch and use those result to evaluate the performance of the CNN and its stability towards overfitting.

C. The CNN architecture

In a CNN different layers play different roles. In the *convolution layer*, features can be extracted by apply-

ing convolution between the input and the convolution kernel. After multiple convolution layers with various filters, a CNN can learn different features of the input data from lower-level to higher-level. The purpose of the *pooling layer* is to reduce the size of the convolved features while remaining the key features so that one can reduce computation complexity and avoid over-fitting. In the *fully-connected layer*, neurons are connected with all activation values to the previous layer, with the final output indicating the predicted class probability. In this work, *rectified linear unit (relu)* is selected as the activation function to all but the output layer, for which the *softmax* function is adopted. Relu function is convenient and efficient to avoid vanishing gradient problems when training a CNN [41], and softmax function at the last layer achieves a normalization so that the output is constrained in the range of (0,1). In this way, the output can appropriately represent the probability that the input sample belongs to one of the classes. The performance of a CNN depends on many hyper-parameters, like CNN depth, convolutional layer number, number of kernels, and their respective size.

IV. SEARCH PROCEDURE

In this section, we will explain the details about the CNN application to the EMRI signal search ².

A. Training phase

In order for CNN to gain the classification ability, a certain amount of input data is necessary for the training. More samples are needed to train CNN to detect weaker signals. In the actual training, it turns out that the I/O of training data becomes the bottleneck. Therefore, we are limited in the size of total data, which in turn determines the SNR lower limit. After some trial and error, we converged at 50 as the SNR lower limit in our training sample. On the other hand, stronger signals can be detected more easily. We concentrate on signals with SNR comparable with the lower limit, therefore an upper limit of SNR 120 is adopted. The total amount of 2.5×10^5 *signal plus noise* samples are used to train the final CNN.

With the data ready, we need to train the network to maximize the classification ability without over-fitting. In general, a more complicated network is expected to perform better on the training data. However, a too-complicated neural network runs a risk of over-fitting or being too acquainted with the training data so that the

² In our implementation of the CNN, we use Keras 2.2.4, and Tensorflow 1.1.18. We implement our software on top of a GPU Tesla V100 PCIe 16 GB, CPU Intel Xeon Gold 6140 (72) @ 2.3, with a memory of 256GB.

performance on training data is not generalizable towards unfamiliar data. Therefore one needs to check the consistency of the CNN performance on both training data and validation data. One can identify the over-fitting if the performance on the training data is significantly better than the validation data. This can be relieved by adopting a less complicated neural network, adding dropout layers, or increasing the size of the training set.

Finally, we need to fix the training architecture. We start our trial of convolution kernel size of (1,5) as suggested by previous studies [53, 58], and finally settled with a size of (1,34) through trial and error. We apply the max-pooling in the pooling layer, which can retain the maximum value of each patch of the input. The mini-batch size is set to 56, and we train the CNN with 300 epochs. For each mini-batch, 56 simulated datasets will be randomly selected from samples, half of which are *signal plus noise* samples and the other half are *noise-only* samples. The wall clock time for each epoch of training is about 2 hours. Strategies like early stopping [89] are also employed to increase efficiency. In our experiments, the training process will be stopped when either 300 epochs are reached, or the accuracy of validation data no longer increases for 50 epochs. Finally, it takes about 10.5 days to train the final CNN. This trained model can afterwards be used for testing. We summarise the architecture of the CNN in Table III.

TABLE III. the architecture of the CNN

Layers	kernel number	kernel size	Activation function
1 Input	/	matrix(size: 2×262144)	/
2 Convolution	32	matrix(size: 1×34)	relu
3 Pooling	16	matrix(size: 1×8)	relu
4 Convolution	16	matrix(size: 1×8)	relu
5 Pooling	16	matrix(size: 1×6)	relu
6 Convolution	16	matrix(size: 1×6)	relu
7 Pooling	16	matrix(size: 1×4)	relu
8 Flatten	/	/	/
9 Dense	/	vector(size: 128)	relu
10 Dense	/	vector(size: 32)	relu
11 Output	/	vector(size: 2)	softmax

B. Testing phase

Once a CNN is trained, we test the performance with 7 different groups of testing data, the properties of which are introduced in Table IV. The setup of these groups is motivated by the testing of two main points, namely the validity of the CNN, and the sensitivity of the CNN. For testing the validity, we prepare three groups of signal+noise samples: signals that follow the same distribution as the training data, signals that follow an astrophysically motivated distribution, and signals produced by AAK but with the same distribution as the training data. For testing the sensitivity, we apply the CNN on top of the remaining four groups, in which many intrinsic

parameters are fixed, but either mass, redshift, or spin is enumerated over a certain list. Taking group 4 as an example, it replaces the SNR distribution in group 1 with a set of fixed SNRs, taking values 10, 20, 30, ..., 130. Similarly, in group 5, it enumerates the mass of the MBH, taking values of 10^4 , $10^{4.5}$, ..., $10^7 M_\odot$. In group 6, it enumerates the spin of the MBH in a similar way, taking values of 0.0, 0.2, 0.4, 0.6, 0.8, 0.98 with fixed MBH mass is $10^6 M_\odot$. In group 7, it focuses on EMRI systems with three kinds of MBHs: 1. $M = 10^{5.5} M_\odot$ and $a = 0.98$; 2. $M = 10^6 M_\odot$ and $a = 0.98$; 3. $M = 10^6 M_\odot$ and $a = 0$. And they enumerates the redshift, taking the value of 0.1, 0.15, 0.2, 0.25, 0.3, 0.35.

The trained CNN takes the testing data and outputs a probability y_{pred} , which predicts the probability that the data contains a signal. Conversely, $1 - y_{\text{pred}}$ predicts the probability that it is purely Gaussian noise.

V. RESULTS

A. Validity

To measure the effectiveness of the trained CNN model, we used the testing data of groups 1-3 in Table IV. One can compare the efficiencies among detection methods through the Receiver Operator Characteristics (ROC) curve, where the True Alarm Probability (TAP) is plotted against the False Alarm Probability (FAP).

In the calculation of the ROC, one first chooses a value of y^* as a detection threshold. Any value of $y_{\text{pred}} > y^*$ is regarded as an alarm. At a given y^* , the FAP and the TAP is constructed as the fraction of *noise-only* and signal plus noise samples, respectively, that are reported as an alarm. By varying y^* , one obtain pairs of TAP-FAPs, and a line that links all these pairs is the ROC curve. By definition, the TAP is a monotonic function of the FAP.

The blue line in Figure 3 represents a ROC curve for group 1. The dotted line is the worst possible performance which indicates a total inability of distinguishing signal from noise. One can observe that for FAP equal to 1%, the corresponding TAP is approximately 92%. This fact indicates that when the CNN is applied to data constructed in a similar way to training data, it has a very good ability to distinguish the signal from noise. We remark that the results from group 1 are expected to be the upper limit among all groups and could be used as a benchmark.

The purple line in Figure 3 corresponds to the ROC curve for group 2. Compared with the blue line, the purple line is lower: for a FAP of 1%, the TAP is about 75%. Since group 2 is an astrophysically motivated population, the majority of signals are expected to be associated with relatively low SNRs. Due to the choice of SNR threshold of 50 for the training data, we observe clustering of SNRs just above 50. This significant decrease in TAP can be explained by such clustering. On the other hand, even

TABLE IV. Different waveform setups used as testing data. The first column indicates the index; the second column is the waveform model; the third column describes the parameter distribution, where to some astrophysical parameters the range shown in the table will be applied, but other unspecified parameters will remain in the same range as the training data (See table II).

number	waveform model	physical parameters distribution	signal samples number
1	AK	$\rho \in \text{uniform } [50, 120]$	500
2	AK	$\rho > 50$, astrophysical model M12	500
3	AAK	$\rho \in \text{uniform } [50, 120]$	500
4	AK	ρ enumerates 10, 20, ..., 130	1000×13
5	AK	M enumerates $10^4, 10^{4.5}, \dots, 10^7 M_\odot$, $a = 0.98$	$z = 0.1$ $z = 0.2$ $z = 0.3$ $1000 \times 7 \times 3$
6	AK	$M = 10^6 M_\odot$, a enumerates 0.0, 0.2, 0.4, 0.6, 0.8, 0.98	$z = 0.1$ $z = 0.2$ $z = 0.3$ $1000 \times 6 \times 3$
7	AK	$M = 10^{5.5} M_\odot$, $a = 0.98$, z enumerates 0.1, 0.2, 0.3 $M = 10^6 M_\odot$, $a = 0.0$, z enumerates 0.1, 0.2, 0.3 $M = 10^6 M_\odot$, $a = 0.98$, z enumerates 0.1, 0.2, 0.3	$1000 \times 3 \times 3$

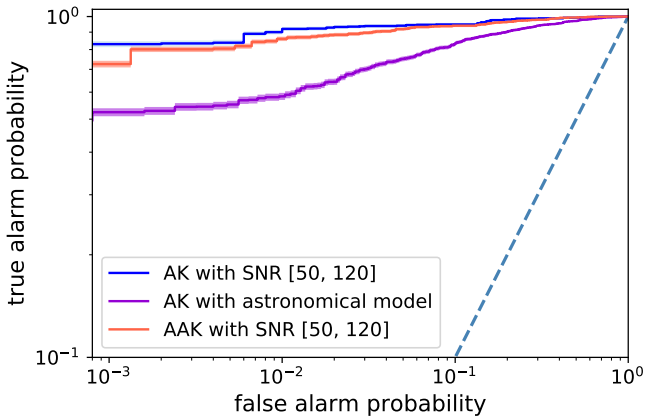


FIG. 3. The ROC curve of the signals from testing groups 1-3 is shown with the blue, purple, and red lines, respectively. The blue line indicates the expected effectiveness for group 1, the parameters have identical distribution to the training data; for group 2, the distribution is drawn from an astrophysical model; for group 3, the distribution is the same as group 1 and the training data, but switched to the AAK waveform model. The $1-\sigma$ confidence intervals are indicated by the shaded regions.

though the training data differ in the parameter distribution to the testing data, the trained CNN still has a good ability to detect signals, showing a good generalization ability.

The red line in Figure 3 corresponds to the ROC curve of group 3. In this group, the parameter distribution is the same as in group 1, but a different set of waveform models, namely the AAK waveform, is adopted. This group could be used to study the robustness of the

CNN against choices of EMRI waveforms. The similarity between the blue line and the red line indicates that a CNN trained on AK waveforms can achieve good detection ability on AAK waveforms. One should be cautious not to take this conclusion for granted. Although both AK and AAK adopt similar equations, the setup of the waveform calculation makes them dephase quickly after a timescale of around an hour. This observation gives us confidence that the CNN method could be robust, and even if a fully realistic waveform would not be available when space-borne GW detectors are operating, one can still rely on CNN trained on approximate waveforms to achieve good detection potential up to certain SNR values.

B. Sensitivity

One can present the sensitivity of the CNN over parameters like SNR, MBH mass, MBH spin, or redshift, using the distribution of signal parameters group 4-7 in Table IV. We have adopted fixed FAPs of 0.1 and 0.01. The results are presented in Figure 4 in the form of efficiency curves.

Figure 4(a) displays the TAP versus the SNR for signals from group 4. It shows that two efficiency curves are consistent with the expectation that the CNN exhibits higher sensitivity towards stronger signal, and for SNR of higher than about 100, the CNN both can guarantee a certain detection at false alarm probabilities of 0.01 and 0.1.

We study the effectiveness of the CNN over MBH masses with group 5. The MBH spin is fixed as $a = 0.98$, and redshifts are varied through $z = 0.1, 0.2, 0.3$. It can be seen from Figure 4(b) that EMRI systems con-

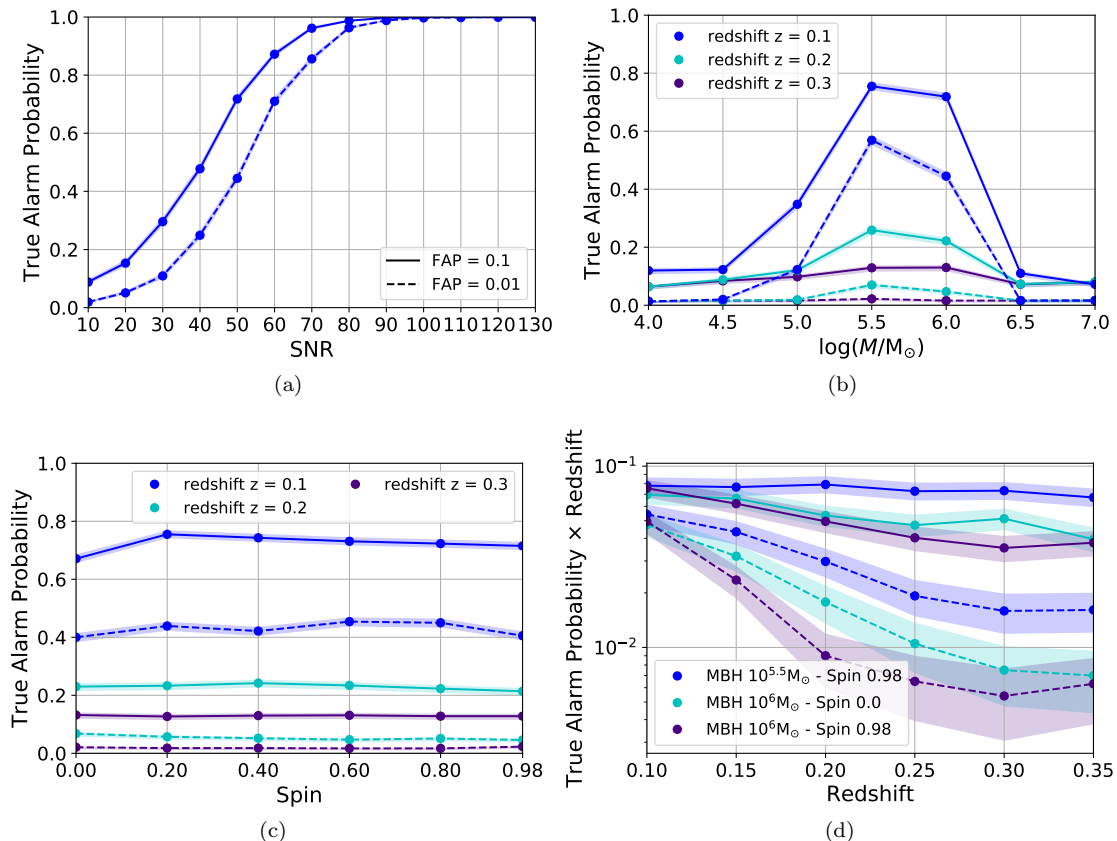


FIG. 4. The comparison of the CNN sensitivity over EMRIs with different parameters. The vertical axis is the TAP, while the horizontal axis is the single varying parameter. The $1\text{-}\sigma$ confidence intervals are indicated by the shaded regions.

taining a $10^{5.5}M_{\odot}$ MBH are more easily detected by the trained CNN. For such an EMRI with a redshift of 0.1, the CNN can achieve a TAP of 78% conditioned on a FAP of 0.1. The feature of a sensitivity peak with regards to MBH masses stems from two factors: the sensitivity curve, and the GW signal frequency distribution. Larger MBH masses lead to larger amplitudes and lower frequencies, while the detector is most sensitive only for a limited frequency band. This result is also consistent with horizon distance analysis in a previous study [12].

EMRI signals contained in group 6 have a fixed MBH mass ($M = 10^6M_{\odot}$) and fixed redshifts ($z = 0.1, 0.2, 0.3$). From Figure 4(c) one can observe that there is almost no apparent dependency of detection ability on MBH spin. This is a piece of indirect evidence that the detection ability is most sensitive to the SNR. Further investigation shows that EMRI systems with different spins share similar SNRs. This can be roughly explained as follows: the EMRI systems have a smooth evolution of overall GW amplitude before the final merger. According to our setup, the merger is fixed at the end of the three months observation period. Therefore, the spin parameter has a minor impact on the overall SNR.

We present the efficiency curve for group 7 in Figure 4(d). The SNR of a signal is roughly inversely propor-

tional to luminosity distance, which for the nearby universe is approximately proportional to redshift. Therefore, on the vertical axis we plot the TAP multiplied by the redshift. If the TAP is linearly correlated with the SNR, then such multiplication should remain constant. This is exactly what we observed in the blue solid line ($M = 10^{5.5}M_{\odot}$ and spin = 0.98, FAP equals 0.1). The distribution of SNR within this redshift range (0.1-0.35) roughly spans the range of 20-60, which shows good linearity between SNR and TAP in Figure 4(a). The prediction is violated to the biggest degree for the purple dashed line ($M = 10^6M_{\odot}$ and $a = 0.98$, FAP equals 0.01). Closer events are gaining more detection efficiency, which can be expected from the non-linearity of the dashed line in the similar SNR range in Figure 4(a).

From the above analyses, we can conclude that among all factors, the SNR has the strongest influence on the expected detection efficiency. Changes in other parameters can also lead to a different performance in TAP, but such differences can be mostly explained by the different SNRs.

VI. CONCLUSIONS AND FUTURE WORKS

The GW detection of EMRI signals is challenging from multiple perspectives. The accurate modeling of the EMRI waveform is a challenging task for theorists. At the same time a template bank with $\sim 10^{40}$ templates is needed, if one is to adopt the traditional matched filtering method [29]. In this work, we demonstrate a proof-of-principle application of a CNN on the detections of EMRI signals. Specifically, an EMRI source with a signal-to-noise ratio uniformly drawn from 50 to 120 can be detected with a detection rate of 91% at a search false alarm probability of 1%. Meanwhile, the effectiveness of the CNN varies for different physical parameters, however, most of the difference can be explained by the influence on SNRs. EMRI systems associated with higher SNR can be expected to be detected with a higher chance.

Our work indicates that detecting EMRI signals with CNN is appealing, also for multiple reasons. For example, after being properly trained, the CNN can be applied to three months' data, and finish the classification within one second. On the other hand, CNN shows a good generalization ability against a change of waveform models. While training on the less physical AK waveform, the network can successfully recognize signals injected with AAK waveforms. This implies that the CNN method has the potential to achieve actual detections even without accurately modeled waveforms.

It is important to benchmark the performance of the CNN to other methods. However, to the best of the authors' knowledge, there is no other study that uses physical EMRI waveforms to search the signal and also provides a FAP estimate [30]. For example, some studies use physical waveforms to search for the signal, but no FAP is available [29]. On the other hand, other studies can provide FAP estimates, but under a very different context, and the authors adopt phenomenological wave-

forms for the detection [30].

Although this proof-of-principle study is encouraging, we recognize that there are still lots of challenges to implement a reliable CNN to detect EMRI signals. For example, one needs to push the SNR threshold to the values lower than 50. This translates to a much larger amount of training data, the data generation, data storage, and training procedure would all become more hardware demanding. We leave the issue for future studies, with the potential of applying more sophisticated methods or constructing a hierarchical search and examining strategy for the detection pipeline.

At present, there are many Machine Learning (ML) algorithms that have wider applications in GW data analysis. For different problems in the GW field, developers use different methods or different architectures of the same method like CNN. Similar challenges might also apply to the EMRI search, and we could borrow the existing wisdom in the future.

VII. ACKNOWLEDGMENTS

The authors thank Hui-Min Fan, Lianggui Zhu, En-Kun Li, Jianwei Mei, and Micheal Williams for helpful discussions. This work has been supported by Guangdong Major Project of Basic and Applied Basic Research (Grant No. 2019B030302001), the Natural Science Foundation of China (Grants No. 12173104, No. 11805286, and No. 11690022). CM is supported by the Science and Technology Facilities Council grant ST/V005634/1. The authors would like to thank the Gravitational-wave Excellence through Alliance Training (GrEAT) network for facilitating this collaborative project. The GrEAT network is funded by the Science and Technology Facilities Council UK grant no. ST/R002770/1. NK would like to thank the support of the CNES fellowship. The authors acknowledge the uses of the calculating utilities of `numpy` [90], `scipy` [91], and the plotting utilities of `matplotlib` [92].

-
- [1] B. P. Abbott *et al.* (LIGO Scientific, Virgo), *Phys. Rev. X* **9**, 031040 (2019), [arXiv:1811.12907 \[astro-ph.HE\]](#).
- [2] R. Abbott *et al.* (LIGO Scientific, Virgo), (2020), [arXiv:2010.14527 \[gr-qc\]](#).
- [3] R. Abbott *et al.* (LIGO Scientific, VIRGO), (2021), [arXiv:2108.01045 \[gr-qc\]](#).
- [4] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA), (2021), [arXiv:2111.03606 \[gr-qc\]](#).
- [5] S.-J. Huang, Y.-M. Hu, V. Korol, P.-C. Li, Z.-C. Liang, Y. Lu, H.-T. Wang, S. Yu, J. Mei, *et al.*, *Physical Review D* **102**, 063021 (2020).
- [6] V. Korol, E. M. Rossi, P. J. Groot, G. Nelemans, S. Toonen, and A. G. A. Brown, *Mon. Not. Roy. Astron. Soc.* **470**, 1894 (2017), [arXiv:1703.02555 \[astro-ph.HE\]](#).
- [7] H.-T. Wang, Z. Jiang, A. Sesana, E. Barausse, S.-J. Huang, Y.-F. Wang, W.-F. Feng, Y. Wang, Y.-M. Hu, J. Mei, *et al.*, *Physical Review D* **100**, 043003 (2019).
- [8] A. Klein *et al.*, *Phys. Rev. D* **93**, 024003 (2016), [arXiv:1511.05581 \[gr-qc\]](#).
- [9] S. Liu, Y.-M. Hu, J.-d. Zhang, J. Mei, *et al.*, *Physical Review D* **101**, 103027 (2020).
- [10] A. Sesana, *Phys. Rev. Lett.* **116**, 231102 (2016), [arXiv:1602.06951 \[gr-qc\]](#).
- [11] S. Babak, J. Gair, A. Sesana, E. Barausse, C. F. Sopuerta, C. P. Berry, E. Berti, P. Amaro-Seoane, A. Petiteau, and A. Klein, *Phys. Rev. D* **95**, 103012 (2017), [arXiv:1703.09722 \[gr-qc\]](#).
- [12] H.-M. Fan, Y.-M. Hu, E. Barausse, A. Sesana, J.-d. Zhang, X. Zhang, T.-G. Zi, and J. Mei, *Phys. Rev. D* **102**, 063016 (2020), [arXiv:2005.08212 \[astro-ph.HE\]](#).
- [13] Z.-C. Liang, Y.-M. Hu, Y. Jiang, J. Cheng, J.-d. Zhang, and J. Mei, (2021), [arXiv:2107.08643 \[astro-ph.CO\]](#).
- [14] Z.-C. Chen, F. Huang, and Q.-G. Huang, *Astrophys. J.* **871**, 97 (2019), [arXiv:1809.10360 \[gr-qc\]](#).
- [15] C. Shi, J. Bao, H.-T. Wang, J.-d. Zhang, Y.-M. Hu, A. Sesana, E. Barausse, J. Mei, J. Luo, *et al.*, *Physical Review D* **100**, 044036 (2019).

- [16] T.-G. Zi, J.-D. Zhang, H.-M. Fan, X.-T. Zhang, Y.-M. Hu, C. Shi, and J. Mei, *Phys. Rev. D* **104**, 064008 (2021), [arXiv:2104.06047 \[gr-qc\]](#).
- [17] L.-G. Zhu, Y.-M. Hu, H.-T. Wang, J.-D. Zhang, X.-D. Li, M. Hendry, and J. Mei, (2021), [arXiv:2104.11956 \[astro-ph.CO\]](#).
- [18] L.-G. Zhu, L.-H. Xie, Y.-M. Hu, S. Liu, E.-K. Li, N. R. Napolitano, B.-T. Tang, J.-d. Zhang, and J. Mei, (2021), [arXiv:2110.05224 \[astro-ph.CO\]](#).
- [19] J. Luo, L.-S. Chen, H.-Z. Duan, Y.-G. Gong, S. Hu, J. Ji, Q. Liu, J. Mei, V. Milyukov, M. Sazhin, *et al.*, *Classical and Quantum Gravity* **33**, 035010 (2016).
- [20] J. Mei, Y.-Z. Bai, J. Bao, E. Barausse, L. Cai, E. Canuto, B. Cao, W.-M. Chen, Y. Chen, Y.-W. Ding, *et al.*, *Progress of Theoretical and Experimental Physics* (2020).
- [21] J. Baker *et al.*, (2019), [arXiv:1907.06482 \[astro-ph.IM\]](#).
- [22] P. Amaro-Seoane, J. R. Gair, M. Freitag, M. Coleman Miller, I. Mandel, C. J. Cutler, and S. Babak, *Class. Quant. Grav.* **24**, R113 (2007), [arXiv:astro-ph/0703495](#).
- [23] C. J. Moore, A. J. K. Chua, and J. R. Gair, *Class. Quant. Grav.* **34**, 195009 (2017), [arXiv:1707.00712 \[gr-qc\]](#).
- [24] N. Yunes, P. Pani, and V. Cardoso, *Phys. Rev. D* **85**, 102003 (2012), [arXiv:1112.3351 \[gr-qc\]](#).
- [25] P. Canizares, J. R. Gair, and C. F. Sopuerta, *Phys. Rev. D* **86**, 044010 (2012), [arXiv:1205.1253 \[gr-qc\]](#).
- [26] L. Barack and C. Cutler, *Phys. Rev. D* **75**, 042003 (2007), [arXiv:gr-qc/0612029](#).
- [27] P. Lynch, M. van de Meent, and N. Warburton, (2021), [arXiv:2112.05651 \[gr-qc\]](#).
- [28] S. Isoyama, R. Fujita, A. J. K. Chua, H. Nakano, A. Pound, and N. Sago, (2021), [arXiv:2111.05288 \[gr-qc\]](#).
- [29] J. R. Gair, L. Barack, T. Creighton, C. Cutler, S. L. Larson, E. S. Phinney, and M. Vallisneri, *Classical and Quantum Gravity* **21**, S1595 (2004).
- [30] Y. Wang, Y. Shang, and S. Babak, *Physical Review D* **86**, 104050 (2012).
- [31] Y. Wang, G. Heinzl, and K. Danzmann, *Physical Review D* **92**, 044037 (2015).
- [32] J. Gair and L. Wen, *Classical and Quantum Gravity* **22**, S1359 (2005).
- [33] L. Wen and J. R. Gair, *Classical and Quantum Gravity* **22**, S445 (2005).
- [34] J. Gair and G. Jones, *Classical and Quantum Gravity* **24**, 1145 (2007).
- [35] J. R. Gair, I. Mandel, and L. Wen, *Class. Quant. Grav.* **25**, 184031 (2008), [arXiv:0804.1084 \[gr-qc\]](#).
- [36] J. R. Gair, E. Porter, S. Babak, and L. Barack, *Classical and Quantum Gravity* **25**, 184030 (2008).
- [37] S. Babak, J. R. Gair, and E. K. Porter, *Classical and quantum gravity* **26**, 135004 (2009).
- [38] A. Ali, N. Christensen, R. Meyer, and C. Rover, *Class. Quant. Grav.* **29**, 145014 (2012), [arXiv:1301.0455 \[gr-qc\]](#).
- [39] A. J. K. Chua, *J. Phys. Conf. Ser.* **716**, 012028 (2016), [arXiv:1602.00620 \[gr-qc\]](#).
- [40] A. J. K. Chua, N. Korsakova, C. J. Moore, J. R. Gair, and S. Babak, *Phys. Rev. D* **101**, 044027 (2020), [arXiv:1912.11543 \[astro-ph.IM\]](#).
- [41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Advances in neural information processing systems* **25**, 1097 (2012).
- [42] K. Simonyan and A. Zisserman, *arXiv e-prints*, [arXiv:1409.1556](#) (2014), [arXiv:1409.1556 \[cs.CV\]](#).
- [43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, *arXiv e-prints*, [arXiv:1512.00567](#) (2015), [arXiv:1512.00567 \[cs.CV\]](#).
- [44] K. He, X. Zhang, S. Ren, and J. Sun, *arXiv e-prints*, [arXiv:1512.03385](#) (2015), [arXiv:1512.03385 \[cs.CV\]](#).
- [45] F. Chollet, *arXiv e-prints*, [arXiv:1610.02357](#) (2016), [arXiv:1610.02357 \[cs.CV\]](#).
- [46] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, *arXiv e-prints*, [arXiv:1602.07261](#) (2016), [arXiv:1602.07261 \[cs.CV\]](#).
- [47] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero (IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013).
- [48] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, *arXiv e-prints*, [arXiv:1103.0398](#) (2011), [arXiv:1103.0398 \[cs.LG\]](#).
- [49] D. George and E. A. Huerta, *Phys. Rev. D* **97**, 044039 (2018), [arXiv:1701.00008 \[astro-ph.IM\]](#).
- [50] D. George and E. A. Huerta, *Phys. Lett. B* **778**, 64 (2018), [arXiv:1711.03121 \[gr-qc\]](#).
- [51] E. Huerta, G. Allen, I. Andreoni, J. Antelis, E. Bachelet, G. Berriman, F. Bianco, R. Biswas, M. Carrasco?Kind, K. Chard, M. Cho, P. Cowperthwaite, Z. Etienne, M. Fishbach, F. Forster, D. George, T. Gibbs, M. Graham, W. Gropp, and Z. Zhao, *Nature Reviews Physics* **1**, 1 (2019).
- [52] H. Xia, L. Shao, J. Zhao, and Z. Cao, (2020), [arXiv:2011.04418 \[astro-ph.HE\]](#).
- [53] H. Gabbard, M. Williams, F. Hayes, and C. Messenger, *Phys. Rev. Lett.* **120**, 141103 (2018), [arXiv:1712.06041 \[astro-ph.IM\]](#).
- [54] M. B. Schäfer, F. Ohme, and A. H. Nitz, *Physical Review D* **102**, 063015 (2020).
- [55] P. G. Krastev, K. Gill, V. A. Villar, and E. Berger, *Phys. Lett. B* **815**, 136161 (2021), [arXiv:2012.13101 \[astro-ph.IM\]](#).
- [56] M. L. Chan, I. S. Heng, and C. Messenger, *Physical Review D* **102**, 043022 (2020).
- [57] J. Bayley, C. Messenger, and G. Woan, *Phys. Rev. D* **100**, 023006 (2019).
- [58] J. Bayley, C. Messenger, and G. Woan, *Phys. Rev. D* **102**, 083024 (2020).
- [59] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, (2019), [arXiv:1909.06296 \[astro-ph.IM\]](#).
- [60] M. J. Williams, J. Veitch, and C. Messenger, (2021), [arXiv:2102.11056 \[gr-qc\]](#).
- [61] D. George, H. Shen, and E. A. Huerta, in *NiPS Summer School 2017* (2017) [arXiv:1711.07468 \[astro-ph.IM\]](#).
- [62] H. Shen, D. George, E. A. Huerta, and Z. Zhao, (2019), [10.1109/ICASSP.2019.8683061](#), [arXiv:1903.03105 \[astro-ph.CO\]](#).
- [63] J. Heaton, *Genetic Programming and Evolvable Machines* **19** (2017), [10.1007/s10710-017-9314-z](#).
- [64] P. Amaro-Seoane, *Living Rev. Rel.* **21**, 4 (2018), [arXiv:1205.5240 \[astro-ph.CO\]](#).
- [65] Z. Pan and H. Yang, *Phys. Rev. D* **103**, 103018 (2021), [arXiv:2101.09146 \[astro-ph.HE\]](#).
- [66] Z. Pan, Z. Lyu, and H. Yang, *Phys. Rev. D* **104**, 063007 (2021), [arXiv:2104.01208 \[astro-ph.HE\]](#).
- [67] R. Arcodia *et al.*, *Nature* **592**, 704 (2021), [arXiv:2104.13388 \[astro-ph.HE\]](#).

- [68] S. Babak, J. Gair, A. Sesana, E. Barausse, C. F. Sopuerta, C. P. Berry, E. Berti, P. Amaro-Seoane, A. Petiteau, and A. Klein, *Physical Review D* **95**, 103012 (2017).
- [69] L. Barack, *Classical and Quantum Gravity* **26**, 213001 (2009).
- [70] M. Van De Meent, *Physical Review D* **97**, 104033 (2018).
- [71] J. Miller and A. Pound, *Phys. Rev. D* **103**, 064048 (2021), [arXiv:2006.11263 \[gr-qc\]](https://arxiv.org/abs/2006.11263).
- [72] S. A. Hughes, N. Warburton, G. Khanna, A. J. K. Chua, and M. L. Katz, *Phys. Rev. D* **103**, 104014 (2021), [arXiv:2102.02713 \[gr-qc\]](https://arxiv.org/abs/2102.02713).
- [73] J. McCart, T. Osburn, and J. Y. J. Burton, *Phys. Rev. D* **104**, 084050 (2021), [arXiv:2109.00056 \[gr-qc\]](https://arxiv.org/abs/2109.00056).
- [74] S. A. Hughes, S. Drasco, E. E. Flanagan, and J. Franklin, *Physical review letters* **94**, 221101 (2005).
- [75] S. Babak, H. Fang, J. R. Gair, K. Glampedakis, and S. A. Hughes, *Physical Review D* **75**, 024005 (2007).
- [76] L. Barack and C. Cutler, *Phys. Rev. D* **69**, 082005 (2004), [arXiv:gr-qc/0310125 \[gr-qc\]](https://arxiv.org/abs/gr-qc/0310125).
- [77] A. J. Chua and J. R. Gair, *Classical and Quantum Gravity* **32**, 232002 (2015).
- [78] M. L. Katz, A. J. K. Chua, L. Speri, N. Warburton, and S. A. Hughes, *Phys. Rev. D* **104**, 064047 (2021), [arXiv:2104.04582 \[gr-qc\]](https://arxiv.org/abs/2104.04582).
- [79] A. J. K. Chua and J. R. Gair, *Class. Quant. Grav.* **32**, 232002 (2015), [arXiv:1510.06245 \[gr-qc\]](https://arxiv.org/abs/1510.06245).
- [80] C. Cutler, *Physical Review D* **57**, 7089 (1998).
- [81] N. J. Cornish and S. L. Larson, *Classical and Quantum Gravity* **18**, 3473 (2001).
- [82] N. J. Cornish and L. J. Rubbo, *Physical Review D* **67**, 022001 (2003).
- [83] C. Zhang, Q. Gao, Y. Gong, B. Wang, A. J. Weinstein, C. Zhang, et al., *Physical Review D* **101**, 124027 (2020).
- [84] F. B. Estabrook, M. Tinto, and J. W. Armstrong, *Phys. Rev. D* **62**, 042002 (2000).
- [85] M. Tinto and J. W. Armstrong, *Phys. Rev. D* **59**, 102003 (1999).
- [86] A. Ng, J. Ngiam, C. Y. Foo, Y. Mai, C. Suen, A. Coates, A. Maas, A. Hannun, B. Huval, T. Wang, et al., “Unsupervised feature learning and deep learning,” (2013).
- [87] Y. LeCun et al., URL: <http://yann.lecun.com/exdb/lenet> **20**, 14 (2015).
- [88] T. Dozat, (2016).
- [89] Y. L. R. A. C. Yao, *Constructive Approximation*. **26**, 289 (2007).
- [90] S. van der Walt, S. C. Colbert, and G. Varoquaux, *Comput. Sci. Eng.* **13**, 22 (2011), [arXiv:1102.1523 \[cs.MS\]](https://arxiv.org/abs/1102.1523).
- [91] P. Virtanen et al., *Nature Meth.* **17**, 261 (2020), [arXiv:1907.10121 \[cs.MS\]](https://arxiv.org/abs/1907.10121).
- [92] J. D. Hunter, *Comput. Sci. Eng.* **9**, 90 (2007).