Machine-learning Love: classifying the equation of state of neutron stars with transformers

Gonçalo Gonçalves, Márcio Ferreira, João Aveiro, Antonio Onofre, Felipe F. Freitas, Constança Providência and José A. Font

aCFisUC, Department of Physics, University of Coimbra, P-3004 - 516 Coimbra, Portugal
bDepartamento de Engenharia Física, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
cCentro de Física das Universidades do Minho e do Porto (CF-UM-UP), Universidade do Minho, 4710-057 Braga, Portugal
dDepartamento de Física da Universidade de Aveiro and, Centre for Research and Development in Mathematics and Applications (CIDMA), Campus de Santiago, 3810-183 Aveiro, Portugal
eDepartamento de Astronomía y Astrofísica, Universitat de València, Dr. Moliner 50, 46100, Burjassot (València), Spain
fObservatori Astronòmic, Universitat de València, Catedrático José Beltrán 2, 46980, Paterna (València), Spain

E-mail: goncalo.mota.goncalves@gmail.com, marcio.ferreira@uc.pt, javeiro@fe.up.pt, antonio.onofre@cern.ch, felipefreitas@ua.pt, cp@fis.uc.pt, j.antonio.font@uv.es

Received June 8, 2023
Revised October 27, 2023
Accepted November 13, 2023
Published December 1, 2023

Abstract. The use of the Audio Spectrogram Transformer (AST) model for gravitational-wave data analysis is investigated. The AST machine-learning model is a convolution-free classifier that captures long-range global dependencies through a purely attention-based mechanism. In this paper a model is applied to a simulated dataset of inspiral gravitational wave signals from binary neutron star coalescences, built from five distinct, cold equations of state (EOS) of nuclear matter. From the analysis of the mass dependence of the tidal deformability parameter for each EOS class it is shown that the AST model achieves a promising performance in correctly classifying the EOS purely from the gravitational wave signals, especially when the component masses of the binary system are in the range $[1, 1.5]M_{\odot}$. Furthermore, the
generalization ability of the model is investigated by using gravitational-wave signals from a new EOS not used during the training of the model, achieving fairly satisfactory results. Overall, the results, obtained using the simplified setup of noise-free waveforms, show that the AST model, once trained, might allow for the instantaneous inference of the cold nuclear matter EOS directly from the inspiral gravitational-wave signals produced in binary neutron star coalescences.

**Keywords:** Machine learning, neutron stars

**ArXiv ePrint:** 2210.08382
1 Introduction

Gravitational waves (GWs) are a crucial source of information in understanding the properties of astrophysical compact objects like black holes and neutron stars (NS). In particular, the GWs emitted during the coalescence of binary NS (BNS) carry important information on the high-density behavior of cold NS matter. The tidal deformability of NS quantifies the quadrupolar deformation of the star when subject to an external tidal field, such as the one generated by the companion star in BNS systems. The tidal deformability parameter is a decreasing function of the NS mass, spanning many orders of magnitude, that is sensitive to the equation of state (EOS) of NS matter. The stars’ tidal response affects the binding energy of the BNS system and the rate of emission of gravitational waves. The EOS of NS matter has a direct imprint on the GW signals through the tidal deformability of each binary component [15, 22, 56].

The landmark GW observation of the BNS merger event GW170817 by the Advanced LIGO and Advanced Virgo detectors [3], followed by dozens of observations of its electromagnetic counterpart GRB170817A/AT2017gfo [2, 4], gave rise to the new field of multi-messenger astronomy with gravitational waves. The observation of the GW170817 event allowed for the inference of the effective tidal deformability [28] of the binary system \( \tilde{\Lambda} \leq 800 \) (90% confidence, using a low-spin prior) [3]. A follow up reanalysis tightened the constraints to \( \tilde{\Lambda} = 300^{+420}_{-230} \) (90% confidence), under minimal assumptions about the nature of the compact objects [6]. These constraints on the tidal effects allowed to set constraints on neutron star radii and on the EOS. The radii of the binary components were estimated as to be \( R_1 = 11.9^{+1.4}_{-1.4} \) km (heavier star) and \( R_2 = 11.9^{+1.4}_{-1.4} \) km (lighter star) [5]. Moreover, the electromagnetic
counterparts of GW170817 have set additional constraints on the lower limit of the tidal deformability: \( A_{1.37 M_\odot} > 210 \) \([12]\), 300 \([55]\), 279 \([17]\), and 309 \([65]\).

The expected increase in the number of detections of BNS coalescences in the upcoming observing runs of second-generation detectors \([1]\) and future experiments \([27, 38, 49]\), will most likely help tighten current constraints. Not surprisingly, the prospects to further improve our knowledge on the NS EOS using upcoming tidal deformability information from inspiral BNS waveforms has received considerable attention \([8, 20, 40, 45, 47, 64]\). Existing approaches are based on Bayesian inference procedures and, to the best of our knowledge, attempts considering alternatives routes, such as machine learning, are generally lacking, although steps have already been taken in that direction \([51]\). Here, we take this path and assess the performance of a machine learning model in classifying the NS EOS from the analysis of the tidal deformability parameter of a simulated dataset of inspiral GW signals from BNS coalescences. Machine Learning (ML) is becoming a powerful tool in the analysis of GW data and the list of applications is rapidly growing \([18, 42]\). Deep learning methods based on convolutional neural networks (CNN) have been used for binary black hole (BBH) signal detection and multiple-parameter estimation, reaching accuracies comparable to matched-filtering methods with higher computational efficiency \([33]\). CNNs also show similar performance as matched-filtering for BNS mergers but they are more resilient to non-stationary, non-Gaussian noise (glitches) in the GW data \([32]\). Deep learning models have been used for instantaneous Bayesian posterior approximation in GW parameter estimation in \([16]\). The use of masked autoregressive flows as a way of increasing the flexibility of deep neural networks in modeling GW posterior distributions was studied in \([37]\). Detection of GW events with deep learning and their comparison with standard inference codes was carried out in \([10, 19]\), showing a fast and accurate inference of physical parameters. The development of fast ML alternatives to the computationally demanding Bayesian inference approaches has been remarkably established by \([30]\) applying conditional variational autoencoders to BBH signals. ML approaches based on object detection have also been applied successfully in the task of detecting BNS coalescence events from the GW data of current detectors \([11]\). Motivated by this growing body of work, we study here the potential to infer the NS EOS directly from simulated BNS inspiral signals using a ML approach.

Across all fields of ML, convolutional-based networks (such as CNN \([48]\), arguably the most common architecture) have seen wide adoption. CNNs were explored as multi-class image classification in \([26]\), using 1824 rotating core collapse GW simulations from 8 different nuclear EOSs. The percentage of correct classifications in the test data set was found to be above 70%, a number that can be improved to better than 90%, if limiting the labels. However, despite their success and popularity these networks have biases inherent to them (such as spatial locality and translation equivariance), which hinder their performance in certain tasks. Conversely, models based on self-attention do not share these inductive bias \([35]\) and, additionally, they do not have non-degenerated convolutions \([39]\). Due to this, they have seen increasing adoption in recent years and have been able to outperform convolutional networks in different tasks. For the Imagenet-1k Dataset \([21]\), for instance, the model with the best performance, PeCo \([24]\), is based on the transformer architecture. The same is seen for the CIFAR-10 Dataset \([44]\), where the model ViT-H/14 \([25]\) shows the highest correct percentage.

In the present work the Audio Spectrogram Transformer (AST) model \([34]\), which is a purely attention-based and convolution-free network, is applied to the analysis of GW data and its performance is explored. In \([35]\) it was shown that this model’s peak performance (on
the AudioSet [31] and ImageNet [21] datasets) was attained when initialized with the weights of a Destiled Data-efficient image Transformer (DeiT) with a resolution of 384. Although previously in [61] it was shown that the DeiT model with this resolution (DeiT-384) could outperform SOTA models in the ImageNet [21], ImageNet Real [13] and Imagenet V2 matched frequency [57] datasets, the AST model was not only able to perform well in the AudioSet [31] dataset but, more importantly, it was able to outperform SOTA models in the ESC-50 [31] and Speech Commands [66] datasets. A proof-of-concept approach is presented for (likelihood-free) rapid parameter estimation. The AST is applied as a classification model on a dataset of simulated inspiral signals from BNS mergers using different NS EOS. The model estimation is orders of magnitude faster than the widely used Bayesian based inference approaches. While the use of DL methods for posterior estimates in gravitational-wave data analysis is only presently beginning, it seems likely that it will soon become mainstream to rapidly analyze an increasing number of transient gravitational-wave events, hopefully reaching the accuracy of the widely used sampling algorithms based methods at a much lower computational cost.

The paper is organized as follows: the generation of all datasets used is described in section 2 and the model and training procedure are presented in section 3. Results of the test applied to the trained models are discussed in section 4 and a possible application of the model is explored in section 5. Lastly, the conclusions are drawn in section 6.

2 Dataset

In this section, we present the neutron star matter EOS that will be used in this study and we will discuss how the dataset used to train the model is generated.

LIGO-Virgo Collaboration uses an agnostic description of the EOS in their analysis of the GW170817 [3], in particular, a spectral representation of the EOS, and their analysis assesses indirectly the validity of different nuclear models through confidence intervals on thermodynamic quantities (e.g., pressure vs density) determined using specific EOS parametrizations. In the present study we will adopt a different methodology considering EOS based on microscopic nuclear models, characterized by given nuclear matter properties, and will assess directly the probability of each one from a given waveform. Our main objective is to learn directly from the waveform properties of nuclear/hadronic matter properties, also on its composition and to show it allows a successful analysis of the waveform that directly assesses the validity of microscopic nuclear models.

2.1 Neutron star matter EOS

For the AST model to be sensitive to a wide range of possible EOS in the mass-radius, \( M(R) \), diagram, five representative EOS based on a microscopic model have been selected. These are plotted in solid lines in figure 1 and they correspond to NS matter covering the ranges \( 2.16 < M_{\text{max}}/M_\odot < 2.77 \), \( 11.74 < R_{1.4M_\odot}/\text{km} < 14.63 \), and \( 11.18 < R_{2.0M_\odot}/\text{km} < 14.67 \). All five EOS are unified EOS that have been built from either a relativistic mean field (RMF) approach or a non-relativistic Skyrme interaction, and satisfy the constraint on the maximum NS mass \( M_{\text{max}} \geq 2M_\odot \) (the details can be found in [29]). The RMF EOS are divided into two types: i) nonlinear Walecka models with constant coupling, NL3 [46] and TM1 [59]; and ii) density-dependent coupling parameters, DD2 [62]. Two Skyrme EOS are used: BSk20 [36] and Sly9 [14]. The pressure dependencies on baryonic density are shown in figure 2 for all five EOS, together with the posterior credible level extracted from the GW170817 event [5].
Before discussing the properties of the EOS, we will introduce the NS properties predicted by these models, in particular, mass radius curves.

The NS properties are determined by solving the Tolmann-Oppenheimer-Volkoff (TOV) equations [53, 60] together with the differential equations that determine the tidal deformability [41]. Figure 1 shows the $M(R)$ diagram and figure 3 displays the tidal deformability dependence on the NS mass for each EOS. The NS tidal deformability is given by $\Lambda = \frac{2}{3}k_2C^{-5}$, where $k_2$ is the quadrupole tidal Love number and $C = M/R$ is the star’s compactness [41]. The properties of the GW emission from a BNS coalescence during the inspiral phase are sensitive to the EOS via the relations $\Lambda_{\text{EOS}}(M)$. Figures 1 and 3, also show the BSR6 model [9] (dashed line), a nonlinear Walecka model with constant coupling, selected to test the ML model in the last section.

Four of the EOS (TM1, DD2, SLy9 and BSk20) predicted by the models chosen lie inside the pressure baryonic density 90% confidence band obtained by the LIGO-Virgo Collaboration from the analysis of GW170817, see figure 2. In the set of chosen EOS we also include NL3.
because this is a model frequently used as reference. The model has been very successfully fitted to nuclear properties of both stable and unstable nuclei [46]. Therefore, it is expected that it describes correctly not very asymmetric nuclear matter at saturation density or below. It is clearly seen that the models above saturation density present different high density behaviors. In particular, the high density behavior of the two non-relativistic models, SLy9 and BSk20, is quite similar, however, BSk20 shows a softer behavior in all density range, originating, as a consequence, more compact stars as seen in figure 1. It is also interesting to notice the behavior of TM1 that goes from a quite hard EOS for densities below three times saturation density (in fact at saturation density it has properties similar to NL3) to the softest one above four times saturation density. This behavior is reflected in the mass-radius curve in figure 1: 1.0 $M_\odot$ stars described by TM1 have a radius similar to the prediction from NL3, but the maximum mass configuration is similar to the one of SLy9 and BSk20. In the three figures 1, 2 and 3, a sixth curve is present corresponding to the model BSR6 [9]. This model will be later on discussed. It is a model with a behavior similar to DD2, however, being harder at saturation density and below.

2.2 Dataset generation

2.2.1 Waveform generation

All datasets used for training, validating and testing were obtained with the PyCBC package [52], employing the IMRPhenomPv2-NRTidalv2 waveform approximant [23]. To generate the inspiral waveform of a compact binary, both the intrinsic parameters that describe the components of the binary and the extrinsic parameters that specify the location and orientation of the binary with respect to the observer are required. The intrinsic parameters are the component masses $M_1$ and $M_2$, their spins $S_1$ and $S_2$, and tidal deformabilities $\Lambda_1$ and $\Lambda_2$.

The sampling procedure consists in randomly choosing values for $M_{1,2}$ from a uniform distribution in the range $[1.0, 2.0]M_\odot$ and imposing $M_1 \geq M_2$ (adhering to the convention $M_1 \geq M_2$ for the binary mass components). The corresponding tidal deformability parameters, $\Lambda_1$ and $\Lambda_2$, were obtained from the $\Lambda(M)$ curve for each EOS. The spins of the binary NS components were set to zero in all generated waveforms. The extrinsic parameters that specify the localization and orientation of the binary system were fixed as follows. In all instances, the inclination of the binary was set to zero and its distance was set to 39 Mpc. After generating
Having the waveform in the desired shape, all that is left is converting it to a spectrogram. High-frequency emissions from inspiral events are primarily observable during the late stages of coalescence, while the majority of the inspiral occurs at frequencies below 100 Hz. Using a linear frequency scale poses a challenge, as it leads to reduced spectral resolution in
lower-frequency bins. This results in the confinement of the entire GW signal, including the pre-coalescence phase, to a single frequency bin for extended time periods, subsequently amplifying the quantization error within these frequency ranges. To circumvent this, the Constant-Q Transform [58] supplied by the librosa package [50] is used and the absolute magnitude of its output was taken. For this transform, the following parameters were employed: 128 frequency bins, a minimum frequency of 32 Hz, a hop length\(^1\) of 64, 28 bins per octave and a tuning parameter set to zero. The resulting spectrogram has 128 frequency bins and 129 time bins. Figure 4 shows one example of a generated spectrogram.

3 Model

The AST model takes the input spectrogram, splits it into several \(16 \times 16\) patches and proceeds to flatten each patch to a 1D patch embedding of size 768, using a linear projection layer. An additional trainable positional embedding is added to each of the previous embeddings and the resulting sequence is used as input for the Transformer [63]. This allows for the option of transfer-learning with other models that make use of the standard Transformer architecture.

Although the AST model can be trained from scratch, it has been shown that increased performance can be achieved when the model is initialized with pre-trained weights [61]. This hypothesis was corroborated by preliminary tests. Due to this, the weights from a Data-efficient Image Transformer (DeiT) [61], trained on the ImageNet dataset [21], were used. This model — referred to as DeiT-base distilled 384 — was chosen because it showed very good Top-1 Accuracy\(^2\) with fewer parameters, when compared to the state of the art model PeCo [24], as is presented in table 2.

\(^1\)This approach is similar to the usual discrete-time Short-time Fourier Transform — albeit with a different transform — wherein the time-series is split into chunks using a specific window size (cf. hop size), overlapped by half the window width to reduce time boundary artifacts between chunks. For each chunk, the transformation is computed, thus resulting in a discrete, time-variant spectrum — i.e. the spectrogram we use for feeding the model.

\(^2\)This metric is calculated by dividing the total number of correct predictions, \(p_c\), by the total number of predictions, \(p_t\), and multiplying by 100, or \(p_c/p_t \times 100\).
### Table 2

Number of parameters and Top-1 Accuracy (achieved on the ImageNet-1k dataset [21]) comparison of the model used for pre-training, DeiT-base distilled 384 [61], and the state of the art model PeCo [24].

| Model Name          | # Params | Top-1 Accuracy |
|---------------------|----------|----------------|
| DeiT-base distilled | 384      | 85.2%          |
| PeCo                | 656M     | 88.3%          |

3.1 Training configuration
This section gives an overview of the training process for the models.

3.1.1 Hardware and resources
A NVIDIA A100 Tensor Core GPU with 40GB of memory was used for training, which allowed for a batch size of 480 samples. The training was carried out for 50 epochs\(^3\) which took around 11 hours and 40 minutes (approximately 14 min/epoch).

3.1.2 Learning rate and loss function
The optimizer used for training was Adam [43]. The parameters used for this optimizer were set as follows: the starting learning rate to \(lr = 10^{-5}\), the weight decay to \(\alpha = 5 \times 10^{-7}\), and \((\beta_1, \beta_2) = (0.95, 0.999)\). The weight decay parameter controls the regularization amount and the \(\beta_{1,2}\) control the running averages of gradient and its square on the Adam algorithm. The scheduler MultiStepLR was also employed to reduce the learning rate by half at epochs 5, 10, 20 and 30.

The Loss Function used was BCEWithLogitsLoss [68] (which is the standard Cross Entropy Loss inside a generalized sigmoid function\(^4\)).

3.1.3 Data augmentation
The purpose of Data Augmentation is to leverage suitable data transformations to achieve an increase in flexibility, variability and generalization ability of the model. In the case of image classification, many transformations commonly used (such as rotation, translation, resizing, flipping, . . .) were designed to help classify ordinary objects like fruits or vehicles, not spectrograms. Due to this, in this work only two methods of data augmentation were employed, SpecAugment [54] and mixup [67].

The SpecAugment method consists in applying masking\(^5\) to the spectrogram, in both the frequency and time domains. Frequency masking is applied so that \(f\) consecutive bins \([f_0, f_0 + f]\) are masked, where \(f\) is sampled from a uniform distribution from 0 to the frequency mask parameter \(F\). The initial frequency \(f_0\) is chosen from \([0, \nu - f]\) where \(\nu\) is the total

---

\(^3\)The training was stopped after 50 epochs because by this stage of the training, the variations of the model performance became negligible.

\(^4\)The generalization procedure used to enable the sigmoid function to be applied to a multi-class problem is known as one-vs-rest method.

\(^5\)The term “masking” means to remove/omit the values of certain bins from the spectrogram.
The SpecAugment method. The top left image corresponds to the original, un-augmented image and the other 24 correspond to possible augmentations.

number of frequency bins. Similarly, time masking is applied so that \( t \) consecutive time steps \([t_0, t_0 + t]\) are masked, where \( t \) is sampled from an uniform distribution from 0 to the time mask parameter \( T \). The initial time \( t_0 \) is chosen from \([0, \tau - t]\), where \( \tau \) is the total number of time bins. Since the number of steps masked is re-calculated at every epoch, it was found that setting the mask parameters to \( F = \frac{\tau}{2} \) and \( T = \frac{\tau}{2} \), allowed for a high degree of variability while still preserving sufficient information in the resulting spectrogram, as can be seen in figure 5. This method of augmentation helps the network to be robust against partial loss of frequency information and partial loss of small segments of the input. Some of these configurations might simulate the possibility of a sudden glitch or some detection failure in a specific time segment of a GW event.

The mixup is a simple data-agnostic data augmentation routine that constructs virtual training samples as \( \tilde{x} = \lambda x_i + (1 - \lambda) x_j \) and \( \tilde{y} = \lambda y_i + (1 - \lambda) y_j \), where \( x_i \) and \( x_j \) are two distinct spectrograms and \( y_i \) and \( y_j \) are their respective one-hot labels. So, from two samples, \((x_i, y_i)\) and \((x_j, y_j)\), a virtual sample is created \((\tilde{x}, \tilde{y})\) and passed on to the model. For this method, \( \lambda \) was sampled from a \( \beta \) distribution, using Numpy’s function \texttt{numpy.random.beta()} with \( \alpha = \beta = 10 \). The impact of this sampling can be seen in figure 6. This augmentation method had a crucial effect in stabilising the learning process and on the model’s generalisation, by suppressing oscillations when predicting outside the training set.

4 Tests

In this section, the success of three models independently trained is explored. It should be noted that the same configuration (detailed in section 3.1) was used for all models, and the only difference comes from the dataset used for training (table 1).

Firstly, Model A, with each star’s mass limited to \([1.0, 1.5] M_\odot\) (train set 1) was trained. The reason for this mass range is that this is the region where the EOS curves used as benchmarks in this study are mostly distinguishable from one another (cf. figure 3). Hence, it
is the region where the model should most easily converge. To contrast with the first model, a second model (Model B) was trained on $M_{1,2}/M_\odot \in [1.5, 2.0]$ (train set 2). Additionally, a third model, Model C, was also trained but on the full parameter range $[1.0, 2.0]M_\odot$ (train set 3). Our goal is to compare the performances between full and shorter range models and define the best approach when dealing with detections. After the training phase, each model is tested on new datasets, generated exclusively for testing, with the mass constraints of the stars equal to the ones used for the respective model’s training. Additionally, model C is also tested on test sets 1 and 2, in order to enable a performance comparison with the other two models.

To evaluate the performance of the model in the test datasets, a confusion matrix together with a table are presented, for each case. Along the horizontal lines, each number in the confusion matrix represents the number of samples predicted with the corresponding class. The metrics Distance Score, Top-$1^6$ and Top-$2$ Accuracy, reported on the tables, are described as follows:

**Top-n accuracy.** Given a sample, the model outputs a vector, $\vec{v}$, with size equal to the number of classes. In this vector each element $v_i$ is interpreted as the predicted probability that the given sample belongs to the $i$-th class. The top-n accuracy metric expresses the average percentage of instances where the target class was within the n predicted classes with the highest probabilities.

**Distance score.** This metric, calculated by averaging across samples, gives the difference between the predicted probability that a given sample belongs to the correct class (i.e. the label) $p_{\text{label}}$ and the predicted probability that the same sample belongs to a given class, $p_{\text{class}}$.

$$ds = p_{\text{label}} - p_{\text{class}}.$$  \hspace{1cm} (4.1)

---

6This metric can also be obtained from the confusion matrix. To do this, choose a EOS and divide the value of the corresponding diagonal square by the sum of all the values in the squares of the matching horizontal line.
Table 3. Results for the top-1 and top-2 accuracy and for the distance score \( ds \) obtained by Model A (see table 1) on the test set 1, \( M_{1,2}/M_\odot \in [1.0, 1.5] \). The last column indicates the average values of the metrics for all EOS.

| metric [%] | BSk20  | DD2   | NL3  | SLy9  | TM1  | Avg     |
|------------|--------|-------|------|-------|------|---------|
| top-1      | 96.5   | 74.6  | 92.1 | 72.8  | 71.9 | 81.58   |
| top-2      | 99.4   | 81.5  | 100.0| 99.6  | 99.2 | 95.94   |
| \( ds \)   | -7.39  | -8.89 | -9.06| -15.37| -13.16| -10.77  |

As an example assume, for a given sample, the label is BSk20 and it is interesting to know the distance score of EOS SLy9. If, in this case, the \( p_{\text{BSk20}} = 60\% \) and \( p_{\text{SLy9}} = 20\% \) then the \( ds \) is 40\%, according to eq. (4.1).

Although this metric might seem simple, it provides very useful information. In the extremes, \( ds = 0 \) means that, on average, the predicted probability that the given sample belongs to the correct class is equal to another class, which can be interpreted as the model being very indecisive between at least those two classes (this behaviour is expected when the sample belongs to a region with big overlap between the two corresponding EOS). Conversely, \( ds = 100\% \) means that, on average, the probability that a sample belongs to the correct class was predicted as 100\%, which implies that the model is able to distinguish between classes with complete certainty. On the other side, \( ds = -100\% \) indicates that the model completely failed its predictions.

In this work, this metric will only be calculated when the model misclassifies a sample and the class chosen for the calculation will be the one predicted with the highest probability. Due to this, the value of \( ds \) can be interpreted as how close — percentage-wise — the model is to making the correct prediction. The consequence of this choice is that \( ds \in [-100, 0] \) and the closer its value is to 0, the better.

4.1 Model A (\( M_{1,2}/M_\odot \in [1.0, 1.5] \))

The performance of the model on the train set 1, \( M_{1,2}/M_\odot \in [1.0, 1.5] \), is reported in table 3. It is found that the model best classifies samples from BSk20 with a top-1 accuracy of 96.5\%. On the other hand, the model has the worst performance for samples from TM1, but still reaches a top-1 accuracy of 71.9\%. While there is, approximately, a 25\% difference in these scores, values above 70\% are already promising.

Although the model is unlikely to misclassify a sample from the classes corresponding to the upper and lower bounds of the \( \Lambda(M) \) curves (i.e. BSk20 and NL3), when it does \( ds \) is higher than -10\%. More interestingly, even for the classes in between, where misclassification is more common, it only decreases to, at most, -15.37\% indicating that, even in failure, the model was close to making the correct prediction. This is corroborated when looking at the top-2 accuracy where, for the DD2 class, it has a value of 81.5\%, which is already quite good, but is surpassed by the excellent values (greater than 99\%) seen in the other classes.

The results in table 3 indicate that the model’s certainty appears to increase with the uniqueness of the class, i.e. with the unique behavior of the model that defines the class. The model misclassification pattern is shown in the confusion matrix displayed in figure 7. To understand this pattern figure 8 shows, as an illustrative example, the difference for \( \Lambda(M) \) between the SLy9 EOS and all other four EOS. It can be seen that the most similar EOS to SLy9 (in terms of \( \Lambda \)) are both DD2 and BSk20. This fact explains why the model
Figure 7. Confusion matrix obtained for Model A (see table 1) on test set 1, now with the constraint that $M_{1,2}/M_\odot \in [1.0, 1.5]$. There are 2491 samples in this subset and each number in the matrix represents the amount of samples classified as the respective class.

Figure 8. The prediction differences for the tidal deformability $\Lambda(M)$ between the SLy9 EOS and the remaining EOS of our sample.

misclassifications of SLy9 were attributed to BSk20 (with 16 mismatches) and DD2 (with 21 mismatches). A similar argument explains the misclassifications shown for the other EOS in figure 7.

It is interesting to notice that there is no symmetry between the mismatches of the models SLy9 and BSk20 shown on the confusion matrices just discussed and the ones that will be discussed in the following sections: although SLy9 may be mismatched as BSk20 the contrary is not true, or at least not as strong. This is possibly because the model learns better the upper/lower EOS bounds: BSk20 is the model with the lowest tidal deformability. SLy9 is the next, and lies between BSk20 and DD2, having also some mixture with DD2. Likewise with NL3 and TM1: NL3 is the upper bound model and the confusion matrix does not show a mixture with other models. However TM1, lying between NL3 and DD2 shows some overlap with NL3.
Figure 9. Confusion matrix obtained for Model B (see table 1) on the test set 2, now with the constraint that $M_{1,2}/M_\odot \in [1.5, 2.0]$. There are 2496 samples in this subset and each number in the matrix represents the amount of samples classified as the respective class.

| metric [%] | BSk20 | DD2 | NL3 | SLy9 | TM1 | Avg |
|------------|-------|-----|-----|------|-----|-----|
| top-1      | 84.0  | 37.1| 97.5| 22.1 | 44.7| **57.08** |
| top-2      | 91.7  | 73.6| 98.8| 90.6 | 96.2| **90.18** |
| $ds$       | $-9.54$ | $-8.52$ | $-6.96$ | $-12.06$ | $-14.38$ | **$-10.29$** |

Table 4. Results for the top-1 and top-2 accuracy and for the distance score $ds$ obtained by Model B (see table 1) on the test set 2, $M_{1,2}/M_\odot \in [1.5, 2.0]$. The last column indicates the average values of the metrics for all EOS.

4.2 Model B ($M_{1,2}/M_\odot \in [1.5, 2.0]$)

Since this range corresponds to the region with the strongest overlap between classes (cf. figure 3), a high degree of confusion is expected. This is shown in figure 9. Although the model is able to classify well samples from BSk20 and NL3 (which, unsurprisingly, correspond to the upper and lower bounds for the $\Lambda(M)$ curves shown in figure 3), it is unable to do so for the classes in-between which, as previously mentioned, is anticipated in this mass range. Looking at table 4, remarkably the model surpasses 90% top-2 accuracy on all classes but DD2, where it reaches 73.6%. The high values for the top-2 accuracy together with the high values of $ds$ (around -10%), validate the hypothesis that the model’s ability to classify the sample is correlated to the uniqueness of the sample’s EOS. Even when the model is trained on a harsher region (when compared to the $[1.0, 1.5]M_\odot$ case) and looking only at the least ‘unique’ class, DD2, we find that although the model misclassifies samples in 62.9% of the cases, it is close to making the correct prediction (which is supported by the high top-2 accuracy of 73.6% and the high $ds$ of -8.52%). A similar discussion is applicable to the model’s performance across all classes, since the model achieves, on average, a top-2 accuracy of 90.18% and a $ds$ of $-10.29\%$, see last column of table 4.
Figure 10. Confusion matrix obtained for Model C (see table 1) on test set 3, now with the constraint that $M_{1,2}/M_{\odot} \in [1.0, 2.0]$. There are 10000 samples in this subset and each number in the matrix represents the amount of samples classified as the respective class.

| metric [%] | BSk20 | DD2 | NL3 | SLy9 | TM1 | Avg   |
|------------|-------|-----|-----|------|-----|-------|
| top-1      | 87.6  | 58.0| 97.4| 31.5 | 41.6| **63.22** |
| top-2      | 95.8  | 78.2| 100 | 95.6 | 98.6| **93.64** |
| $ds$       | -4.39 | -5.88| -4.44| -11.09| -15.32| **-8.22** |

Table 5. Results for the top-1 and top-2 accuracy and for the distance score $ds$ obtained by Model C (see table 1) on the test set 3, $M_{1,2}/M_{\odot} \in [1.0, 2.0]$. The last column indicates the average values of the metrics for all EOS.

4.3 Model C ($M_{1,2}/M_{\odot} \in [1.0, 2.0]$).

For the case of the full mass range, the results presented in table 5 show that the top-1 accuracy from either BSk20 or NL3 are high (87.6% and 97.4%, respectively). The performance for samples from the other EOS classes is not as impressive, only reaching a top-1 accuracy of 43.7% (averaged across the values seen for DD2, SLy9 and TM1). Since this range includes regions with overlap between classes, the values achieved for the top-2 accuracy should also be closely analysed. This metric exhibits promising results reaching over 95% on all classes but DD2, where 78.2% is obtained. These results, together with the high values of $ds$ are encouraging and the success of the model is highlighted when observing its average behaviour in this range, where it accomplishes a top-1 accuracy of 63.22%, top-2 accuracy of 93.64% and $ds$ of -8.22%, last column of table 5.

Since this model was trained on the full mass range, it is expected to have a good performance across all the test datasets used. To enable a fair comparison between this model and the ones discussed in sections 4.1 (Model A) and 4.2 (Model B), it was additionally tested on the datasets used in those sections (test sets 1 and 2). It is worth noting that the average of the values of the metrics, on both sets, is not expected to coincide with those reported in
Figure 11. Confusion matrices obtained for Model C on the test set 1 (left), $M_{1,2}/M_\odot \in [1, 1.5]$, and on the test set 2 (right), $M_{1,2}/M_\odot \in [1.5, 2]$. The datasets have 2491 and 2496 samples, respectively, and each number in the matrix represents the amount of samples classified as the corresponding class.

| Dataset | Metric | BSk20 | DD2 | NL3 | SLy9 | TM1 | Avg |
|---------|--------|-------|-----|-----|------|-----|-----|
| Set 1   | top-1  | 99.6  | 81.0 | 91.9 | 47.3 | 67.6 | 77.44 |
|         | top-2  | 99.8  | 87.3 | 99.8 | 99.4 | 97.2 | 96.7  |
|         | $ds$   | -12.0 | -5.3 | -4.6 | -16.2 | -5.4 | -8.7  |
| Set 2   | top-1  | 98.1  | 54.9 | 99.8 | 24.6 | 52.5 | 65.98 |
|         | top-2  | 97.2  | 88.4 | 100  | 97.2 | 98.8 | 96.32 |
|         | $ds$   | -4.43 | -4.58 | -1.34 | -12.84 | -8.58 | -6.35 |

Table 6. Results for the top-1 and top-2 accuracy and for the distance score $ds$ obtained by Model C (see Table 5. Since the mass restriction is applied to each individual star, samples corresponding to systems with, for example, $M_1/M_\odot = 1.1$ and $M_2/M_\odot = 1.8$ will not be present in either test set 1 or 2, due to one of the stars being outside the range, but will be included in test set 3 since the associated constraint is $M_{1,2}/M_\odot \in [1.0, 2.0]$. Hence, the test set 3 contains samples which are not available in either of the two other test sets.

The left panel of figure 11 displays the performance of this model on the test set 1, $M_{1,2}/M_\odot \in [1.0, 1.5]$, which can be compared to that of model of section 4.1, on the same test set. The corresponding values of the metrics are reported in the top half of Table 6. Although both models show good values of the top-1 accuracy, it is seen that for Model C there is a slight decrease in the score (4.14% averaged across all EOS). Despite this decline being undesired, it can be seen that there is a slight increase in the top-2 accuracy (0.76% on average) and in the $ds$ (2.1% on average), which compensate the decline in the top-1 accuracy. Therefore, in this region, this model’s performance appears to be very similar to that of Model A.
Correspondingly, the right panel of figure 11 shows the performance of Model C on the test set 2, \( M_{1.2}/M_\odot \in [1.5, 2.0] \), and the bottom half of table 6 reports the values of the associated metrics. While Model C classifies quite well the BSk20 and NL3 EOS (with top-1 accuracy over 90%), for TM1 and DD2 the numbers go down to 52% and 55% respectively and, for SLy9, the model seems to mostly misclassify its samples as BSk20, only achieving a top-1 accuracy of 24%. Although these values might seem low, they are a clear improvement over the ones obtained by Model B, with an average rise of 8.9%. A similar increase is also seen in the top-2 accuracy, going up to an average of 96.32% (6.14% more than what is seen for Model B). An additional improvement seen when comparing table 6 with table 4 is that the \( ds \) increases by an average of 3.94%.

Although the behavior in the lower mass region is very similar to the one displayed by Model A, in the upper mass range \([1.5, 2.0]M_\odot\) there is a substantial improvement in its performance, over the one exhibited by Model B. This result, together with the high metric values presented in table 5, shows that training with the full dataset not only leads to a model with good performance across the full parameter range, but it also shows that the resulting model is more capable than its smaller counterparts (when comparing in their respective ranges).

5 Additional model

The full-range model presented in section 4.3 (Model C) has shown great performance in classifying correctly the EOS from which the GW signal was generated. Herein, a step forward is given in testing the generalization capacity of this model. This is made by introducing a new EOS, not used in the training stage and using it as input for the model, with the purpose of analysing whether the model classifies the new GW samples as belonging to the nearest EOS in the \( \Lambda(M) \) diagram. The EOS added is BSR6 and its \( \Lambda(M) \) curve is shown, with a dashed line, in figure 3. It can be seen that its neighbours \( \Lambda(M) \) curves are those of the DD2 and TM1 EOS. Therefore, it is expected that the samples corresponding to BSR6 should be labeled mostly as either one of these two classes.

The results are shown in figure 12, where it can be seen that they are in agreement with the expectation. A clear pattern arises in all three ranges. TM1 is the class predicted in more than half the cases, DD2 follows and, surprisingly, NL3 also has a significant number of predictions. Although, ideally, the predictions would be split only among the two neighboring EOS in the \( \Lambda(M) \) diagram (i.e. DD2 and TM1), the fact that NL3 is commonly predicted is not a sign of failure. The reasoning behind this statement relates to the fact that the \( \Lambda(M) \) curve for NL3 is very similar to that of TM1 (see figure 3) which hinders the model’s ability to discern between samples from these classes. The confusion between these classes is not new and can be seen in figure 11 where in almost all (95.8%) instances that the model’s prediction fails for samples from TM1, they are misclassified as NL3.

All in all this test shows that even in the instance of a BNS coalescence event of unknown EOS, the ML model employed in this study is still able to successfully locate it in relation to the EOS employed in its training.

6 Conclusion

In this work we have employed the Audio Spectrogram Transformer (AST) model, a ML model that uses purely an attention-based mechanism, to capture long-range global dependencies
for analysing simulated GW data corresponding to BNS coalescences. This is a supervised problem where the goal is to correctly predict the class that each BNS merger event belongs to. For that purpose, a dataset has been generated composed of five classes of BNS mergers that correspond to five distinct equations of state of nuclear matter. The imprint of each EOS on the BNS merger GW inspiral signal is encoded on the tidal deformability dependence on the NS mass, $\Lambda(M)$. The tidal deformability parameter $\Lambda$ is a steeply decreasing function of the NS mass, covering many orders of magnitude: for a given EOS of our set, $\Lambda \sim O(10^3)$ for $1M_\odot$ and $O(10)$ for $2M_\odot$. However, in the present classification problem, it is the variability of $\Lambda$ among the different EOS for fixed $M$ that most affects the model performance. The difference in $\Lambda$ prediction for the set spans $624 < \Delta \Lambda(1M_\odot) < 5648$ and $74 < \Delta \Lambda(2M_\odot) < 128$, and thus the capacity to distinguish the different EOS strongly decreases with increasing NS mass. Considering this fact, three models were trained on three distinct ranges of NS masses ($M_\odot$), namely $[1.0, 1.5]$, $[1.5, 2.0]$, and $[1.0, 2.0]$. Model A showed a very good accuracy, ranging from a top-1 accuracy of 96.5% for BSk20 to 71.9% for TM1. Furthermore, the top-2 accuracy showed that even in the case where misclassifications happened, the true EOS was still the second EOS with highest probability. Although Model B presented lower values for the top-1 accuracy (between 84% and 22%), the results seen for the top-2 accuracy were still high, with values ranging from 96.2% to 73.6%, depending on the EOS. Finally, regarding Model C, when tested on the full dataset it showed that, despite having a significant number of misclassifications, the values of the top-2 accuracy are still very high. More interestingly, when this same model is tested in the range $M_{1.2}/M_\odot \in [1.0, 1.5]$ it showed similar performance to Model A (see section 4.1) but, when tested in the range $M_{1.2}/M_\odot \in [1.5, 2.0]$ it is seen that its results yield a clear improvement over the ones achieved by Model B (see section 4.2). Therefore, it follows that training in the range $M_{1.2}/M_\odot \in [1.0, 2.0]$ leads to a robust model that not only preforms well over that range, as expected, but also performs equally well or better in smaller ranges, when compared to models trained specifically for those ranges. We have checked that the model’s performance is mainly affected by the intrinsic degeneracies of the waveforms and not from limitations arising from the model’s architecture.

**Figure 12.** Confusion matrices of Model C on the BSR6 set 1 (top), $M_{1.2}/M_\odot \in [1.0, 1.5]$, the BSR6 set 2 (middle), $M_{1.2}/M_\odot \in [1.5, 2.0]$, and BSR6 set 3 (bottom), $M_{1.2}/M_\odot \in [1.0, 2.0]$. Since these datasets are subsets of those used in previous sections, the total number of samples is much smaller having only 251 (top), 246 (middle) and 1000 (bottom). The numbers in the confusion matrices represent the amount of samples classified as the respective class.
Despite the fact that this ML approach has been only tested with simulated, noise-free GW data, it already shows great promise. As a result, applications with simulated signals injected into Gaussian noise and real detector noise as well as BNS signals corresponding to real events will be explored in the future. The preliminary investigations in these directions clearly indicate that a two-model approach may likely be required, i.e. the present classification model must be supplemented with an already denoised GW signal, and thus a denoising model is required. However, an alternative possibility is exploring different architectures for the AST model that may enable both tasks, denoising and classification, to be carried out simultaneously. This will be the subject of a future work.

To end, we point out a potential benefit the methodology reported in this work might bring to current approaches regarding searches and parameter inferences of signals from compact binary coalescences (CBCs). Waveforms from such sources are well modelled, especially in the inspiral part of the signal, using a combination of analytical and numerical relativity techniques. This has allowed the confident detection of 90 CBC events [7]. The pipelines for CBC searches used by the LIGO-Virgo-KAGRA collaboration rely on matched filtering techniques that require huge template banks as complete as possible across the parameter space of binary systems. Template banks, however, may not be dense enough in some regions of the parameter space. For instance, in the case of BNS mergers, numerical relativity waveforms including the effects of spins in the individual stars are fairly scarce. Other sub-optimal situations include the consideration of effects such as precession of the orbital plane or eccentricity in the case on non-quasi-circular orbits. Accounting for these effects will be increasingly more important as the detection rate increases in upcoming runs. In addition, the inference of the physical parameters of the binaries, usually accomplished through Bayesian statistical analysis, also relies fundamentally on the availability of the “correct” waveform approximant. Therefore, a general-purpose, agnostic model as the one we propose in this study might be a useful supplement to the standard procedures to conduct GW searches and parameter inference in the case of compact binaries and, in particular BNS systems. The application of our ML model may allow to identify in a rapid way the EOS that best fits an observation, and this information could be used as a guide to build specific, fine-tuned waveform models to perform Bayesian inference.

Acknowledgments

JA acknowledges support from the Portuguese Foundation for Science and Technology (FCT) in the framework of the Strategic Funding UIDB/00099/2020 and from the project PTDC/FISAST/7002/2020. FFF is supported by the FCT project PTDC/FIS-PAR/31000/2017 and by the Center for Research and Development in Mathematics and Applications (CIDMA) through FCT, references UIDB/04106/2020 and UIDP/04106/2020. MF and CP acknowledges partial support by national funds from FCT (Fundação para a Ciência e a Tecnologia, I.P, Portugal) under the Projects No. UID/FIS/04564/2019, No. UIDP/04564/2020, No. UIDB/04564/2020, and 2022-06400-PTDC. AO acknowledges support from national funds from FCT, under the projects CERN/FIS-PAR/0029/2019 and CERN/FIS-PAR/0037/2021. JAF acknowledges support from the Spanish Agencia Estatal de Investigación (PGC2018-095984-B-I00 and PID2021-125485NB-C21) and from the Generalitat Valenciana (PROMETEO/2019/071). The authors acknowledge the Laboratory for Advanced Computing at the University of Coimbra (http://www.uc.pt/lca) for providing access to the HPC computing resource Navigator, Minho Advanced Computing Center (MACC) for
providing HPC resources that have contributed to the research results reported within this paper, the Portuguese National Network for Advanced Computing for the grant CPCA/A1-428291-2021. Finally, the authors gratefully acknowledge the computer resources at Artemisa, funded by the European Union ERDF and Comunitat Valenciana as well as the technical support provided by the Instituto de Física Corpuscular, IFIC (CSIC-UV).

References

[1] KAGRA et al. collaborations, Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA, Living Rev. Rel. 21 (2018) 3 [arXiv:1304.0670] [inSPIRE].

[2] LIGO Scientific et al. collaborations, Gravitational Waves and Gamma-rays from a Binary Neutron Star Merger: GW170817 and GRB 170817A, Astrophys. J. Lett. 848 (2017) L13 [arXiv:1710.05834] [inSPIRE].

[3] LIGO Scientific and Virgo collaborations, GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral, Phys. Rev. Lett. 119 (2017) 161101 [arXiv:1710.05832] [inSPIRE].

[4] LIGO Scientific et al. collaborations, Multi-messenger Observations of a Binary Neutron Star Merger, Astrophys. J. 848 (2017) L12 [arXiv:1710.05833].

[5] LIGO Scientific and Virgo collaborations, GW170817: Measurements of neutron star radii and equation of state, Phys. Rev. Lett. 121 (2018) 161101 [arXiv:1805.11581] [inSPIRE].

[6] LIGO Scientific and Virgo collaborations, Properties of the binary neutron star merger GW170817, Phys. Rev. X 9 (2019) 011001 [arXiv:1805.11579] [inSPIRE].

[7] T. Abdelsalhin, A. Maselli and V. Ferrari, Solving the relativistic inverse stellar problem through gravitational waves observation of binary neutron stars, Phys. Rev. D 97 (2018) 084014 [arXiv:1712.01303] [inSPIRE].

[8] B.K. Agrawal, Asymmetric nuclear matter and neutron-skin in extended relativistic mean field model, Phys. Rev. C 81 (2010) 034323 [arXiv:1003.3295] [inSPIRE].

[9] J.D. Álvares et al., Exploring gravitational-wave detection and parameter inference using deep learning methods, Class. Quant. Grav. 38 (2021) 155010 [arXiv:2011.10425].

[10] J. Aveiro et al., Identification of binary neutron star mergers in gravitational-wave data using object-detection machine learning models, Phys. Rev. D 106 (2022) 084059 [arXiv:2207.00591] [inSPIRE].

[11] A. Bauswein et al., Equation-of-state Constraints and the QCD Phase Transition in the Era of Gravitational-Wave Astronomy, AIP Conf. Proc. 2127 (2019) 020013 [arXiv:1904.01306].

[12] L. Beyer et al., Are we done with ImageNet?, arXiv:2006.07159.

[13] E. Chabanat, Interactions effectives pour des conditions extrêmes d’isospin, Ph.D. thesis, Université Claude Bernard-Lyon I, France (1995).

[14] K. Chatziioannou, Neutron star tidal deformability and equation of state constraints, Gen. Rel. Grav. 52 (2020) 109 [arXiv:2006.03168] [inSPIRE].

[15] A.J.K. Chua and M. Vallisneri, Learning Bayesian posteriors with neural networks for gravitational-wave inference, Phys. Rev. Lett. 124 (2020) 041102 [arXiv:1909.05966] [inSPIRE].
[17] M.W. Coughlin et al., Constraints on the neutron star equation of state from AT2017gfo using radiative transfer simulations, Mon. Not. Roy. Astron. Soc. 480 (2018) 3871 [arXiv:1805.09371] [inSPIRE].

[18] E. Cuoco et al., Enhancing Gravitational-Wave Science with Machine Learning, Mach. Learn. Sci. Tech. 2 (2021) 011002 [arXiv:2005.03745] [inSPIRE].

[19] M. Dax et al., Real-Time Gravitational Wave Science with Neural Posterior Estimation, Phys. Rev. Lett. 127 (2021) 241103 [arXiv:2106.12594] [inSPIRE].

[20] W. Del Pozzo et al., Demonstrating the Feasibility of Probing the Neutron-Star Equation of State with Second-Generation Gravitational-Wave Detectors, Phys. Rev. Lett. 111 (2013) 071101 [arXiv:1307.8338].

[21] J. Deng et al., Imagenet: A large-scale hierarchical image database, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, in the proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, (2009), p. 248–255 [DOI:10.1109/CVPR.2009.5206848].

[22] T. Dietrich, T. Hinderer and A. Samajdar, Interpreting Binary Neutron Star Mergers: Describing the Binary Neutron Star Dynamics, Modelling Gravitational Waveforms, and Analyzing Detections, Gen. Rel. Grav. 53 (2021) 27 [arXiv:2004.02527] [inSPIRE].

[23] T. Dietrich et al., Improving the NRTidal model for binary neutron star systems, Phys. Rev. D 100 (2019) 044003 [arXiv:1905.06011] [inSPIRE].

[24] X. Dong et al., PeCo: Perceptual Codebook for BERT Pre-training of Vision Transformers, arXiv:2111.12710.

[25] A. Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv:2010.11929.

[26] M.C. Edwards, Classifying the Equation of State from Rotating Core Collapse Gravitational Waves with Deep Learning, Phys. Rev. D 103 (2021) 024025 [arXiv:2009.07367] [inSPIRE].

[27] É.É. Flanagan and T. Hinderer, Constraining neutron-star tidal Love numbers with gravitational-wave detectors, Phys. Rev. D 77 (2008) 021502 [arXiv:0709.1915].

[28] M. Fortin et al., Neutron star radii and crusts: uncertainties and unified equations of state, Phys. Rev. C 94 (2016) 035804 [arXiv:1604.01944] [inSPIRE].

[29] H. Gabbard et al., Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy, Nature Phys. 18 (2022) 112 [arXiv:1909.06296] [inSPIRE].

[30] J.F. Gemmeke et al., Audio set: An ontology and human-labeled dataset for audio events, in the proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), t (2017), p. 776–780.

[31] D. George and E.A. Huerta, Deep Learning for Real-time Gravitational Wave Detection and Parameter Estimation: Results with Advanced LIGO Data, Phys. Lett. B 778 (2018) 64 [arXiv:1711.03121] [inSPIRE].

[32] D. George and E.A. Huerta, Deep Neural Networks to Enable Real-time Multimessenger Astrophysics, Phys. Rev. D 97 (2018) 044039 [arXiv:1701.00008] [inSPIRE].

[33] Y. Gong, Y.-A. Chung and J. Glass, AST: Audio Spectrogram Transformer, arXiv:2104.01778.

[34] Y. Gong, C.-I.J. Lai, Y.-A. Chung and J. Glass, SSAST: Self-Supervised Audio Spectrogram Transformer, arXiv:2110.09784.

[35] S. Goriely, N. Chamel and J.M. Pearson, Further explorations of Skyrme-Hartree-Fock-Bogoliubov mass formulas. XII: Stiffness and stability of neutron-star matter, Phys. Rev. C 82 (2010) 035804 [arXiv:1009.3840] [inSPIRE].
[37] S.R. Green, C. Simpson and J. Gair, Gravitational-wave parameter estimation with autoregressive neural network flows, Phys. Rev. D 102 (2020) 104057 [arXiv:2002.07656] [nSPIRE].

[38] E.D. Hall, Cosmic Explorer: A Next-Generation Ground-Based Gravitational-Wave Observatory, Galaxies 10 (2022) 90 [nSPIRE].

[39] K. He et al., Masked Autoencoders Are Scalable Vision Learners, arXiv:2111.06377.

[40] F. Hernandez Vivanco et al., Measuring the neutron star equation of state with gravitational waves: The first forty binary neutron star merger observations, Phys. Rev. D 100 (2019) 103009 [arXiv:1909.02698].

[41] T. Hinderer, Tidal Love numbers of neutron stars, Astrophys. J. 677 (2008) 1216 [arXiv:0711.2420] [nSPIRE].

[42] E.A. Huerta et al., Enabling real-time multi-messenger astrophysics discoveries with deep learning, Nature Rev. Phys. 1 (2019) 600 [arXiv:1911.11779] [nSPIRE].

[43] D.P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, arXiv:1412.6980 [nSPIRE].

[44] A. Krizhevsky et al., Learning multiple layers of features from tiny images.

[45] B.D. Lackey and L. Wade, Reconstructing the neutron-star equation of state with gravitational-wave detectors from a realistic population of inspiralling binary neutron stars, Phys. Rev. D 91 (2015) 043002 [arXiv:1410.8866].

[46] G.A. Lalazissis, J. Konig and P. Ring, A new parametrization for the Lagrangian density of relativistic mean field theory, Phys. Rev. C 55 (1997) 540 [nucl-th/9607039] [nSPIRE].

[47] P. Landry, R. Essick and K. Chatziioannou, Nonparametric constraints on neutron star matter with existing and upcoming gravitational wave and pulsar observations, Phys. Rev. D 101 (2020) 123007 [arXiv:2003.04880].

[48] Y. LeCun et al., Convolutional networks for images, speech, and time series, The handbook of brain theory and neural networks 3361 (1995) 1995.

[49] M. Maggiore et al., Science case for the Einstein telescope, JCAP 2020 (2020) 050 [arXiv:1912.02622].

[50] B. McFee et al., librosa/librosa: 0.9.1, [DOI:10.5281/zenodo.6097378].

[51] F. Morawski and M. Bejger, Neural network reconstruction of the dense matter equation of state derived from the parameters of neutron stars, Astron. Astrophys. 642 (2020) A78 [arXiv:2006.07194] [nSPIRE].

[52] A. Nitz et al., gwastro/pycbc: v2.0.2 release of PyCBC, [DOI:10.5281/zenodo.6324278].

[53] J.R. Oppenheimer and G.M. Volkoff, On massive neutron cores, Phys. Rev. 55 (1939) 374 [nSPIRE].

[54] D.S. Park et al., SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition, arXiv:1904.08779 [DOI:10.21437/Interspeech.2019-2680].

[55] D. Radice, A. Perego, F. Zappa and S. Bernuzzi, GW170817: Joint Constraint on the Neutron Star Equation of State from Multimessenger Observations, Astrophys. J. 852 (2018) L29 [arXiv:1711.03647].

[56] J.S. Read et al., Matter effects on binary neutron star waveforms, Phys. Rev. D 88 (2013) 044042 [arXiv:1306.4065].

[57] B. Recht, R. Roelofs, L. Schmidt and V. Shankar, Do imagenet classifiers generalize to imagenet?, in the proceedings of the International conference on machine learning, (2019), p. 5389–5400.
[58] C. Schörkhuber and A. Klapuri, Constant-Q transform toolbox for music processing, Proc. 7th Sound and Music Computing Conf. (2010).

[59] Y. Sugahara and H. Toki, Relativistic mean field theory for unstable nuclei with nonlinear sigma and omega terms, Nucl. Phys. A 579 (1994) 557 [inSPIRE].

[60] R.C. Tolman, Static solutions of Einstein’s field equations for spheres of fluid, Phys. Rev. 55 (1939) 364 [inSPIRE].

[61] H. Touvron et al., Training data-efficient image transformers & distillation through attention, in the proceedings of the International Conference on Machine Learning, (2021), p. 10347–10357.

[62] S. Typel et al., Composition and thermodynamics of nuclear matter with light clusters, Phys. Rev. C 81 (2010) 015803 [arXiv:0908.2344] [inSPIRE].

[63] A. Vaswani et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[64] B. Wang, Z. Zhu, A. Li and W. Zhao, Comprehensive analysis of the tidal effect in gravitational waves and implication for cosmology, Astrophys. J. Suppl. 250 (2020) 6 [arXiv:2005.12875] [inSPIRE].

[65] Y.-Z. Wang et al., GW170817: The energy extraction process of the off-axis relativistic outflow and the constraint on the equation of state of neutron stars, Astrophys. J. 877 (2019) 2 [arXiv:1811.02558] [inSPIRE].

[66] P. Warden, Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition, arXiv:1804.03209.

[67] H. Zhang, M. Cisse, Y.N. Dauphin and D. Lopez-Paz, mixup: Beyond Empirical Risk Minimization, arXiv:1710.09412.

[68] https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html.