Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Efficient self-attention-based emulation of gravitational-wave spectrum with full parameter space exploration

Bohua Li

bohuali@gxu.edu.cn

Guangxi University https://orcid.org/0000-0002-3600-0358

Fan Zhang

Zhejiang University https://orcid.org/0009-0004-3509-7495

Yifang Luo

Joel Meyers

Southern Methodist University

Paul Shapiro

The University of Texas at Austin https://orcid.org/0000-0002-0410-3045

Erik Katsavounidis

Massachusetts Institute of Technology

Article

Keywords:

Posted Date: March 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8253597/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Efficient self-attention-based emulation of gravitational-wave spectrum with full parameter space exploration

Fan Zhang[1,2], Yifang Luo[1], Bohua Li[3*], Joel Meyers[4], Paul R. Shapiro[5], Erik Katsavounidis[6,7]

[1]State Key Laboratory of Ocean Sensing & Ocean College, Zhejiang University, Zhoushan, 316021, Zhejiang, China.
[2]Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, 02139, MA, USA.
[3]Guangxi Key Laboratory for Relativistic Astrophysics, School of Physical Science and Technology, Guangxi University, Nanning, 530004, Guangxi, China.
[4]Department of Physics, Southern Methodist University, Dallas, 75275, TX, USA.
[5]Department of Astronomy, The University of Texas at Austin, Austin, 78712, TX, USA.
[6]Department of Physics, Massachusetts Institute of Technology, Cambridge, 02139, MA, USA.
[7]LIGO Laboratory, Massachusetts Institute of Technology, Cambridge, 02139, MA, USA.

*Corresponding author(s). E-mail(s): bohuali@gxu.edu.cn;

Abstract

Fast emulation of sequential data generated by numerical computations is vital for large-scale Bayesian analysis in cosmology and astrophysics. Despite the success of neural network models in such tasks, benchmarks typically fall short of the required sub-percent accuracy and lack generalizability to broader data sets of different features and physical origins. Here we introduce SageNet+, a physics-informed framework for emulating gravitational-wave spectrum, while our self-attention-based design is adaptable to generic sequential data. First-principles modeling of the stochastic gravitational-wave background from cosmic inflation takes tens of seconds per evaluation to solve differential integrations. Instead, our SageNet+ emulator can predict the present-day gravitational-wave spectrum in less than 10 ms for every set of the nine cosmological parameters with wide prior ranges, achieving a thousand-fold speed-up over numerical methods. Based on the Transformer architecture, SageNet+ combines learnable positional embeddings and adaptive sequence standardization to preserve spectral features. We further employ refined data preprocessing to maximize physical information and devise novel metrics to assess model performance. Trained on 180,000 spectra that explore the full parameter space, SageNet+ attains sub-percent accuracy for all test data. This enables real-time parameter inferences and joint analyses with external data. SageNet+ presents an optimal, systematic and generalizable framework for emulation tasks in physical sciences.

Introduction

Neural network emulation has become an important approach to the forward modeling of physical systems governed by first principles, a method that dates back to NeuroAnimator[1]. In cosmology and astrophysics, deep learning emulators have proved to be successful alternatives to expensive numerical computations of physics-based models in many cases[2–7]. However, while most of these models are developed for one-dimensional regression tasks (i.e., 1D curves as output), it remains challenging for them to achieve subpercent level accuracy, which is required in the era of precision cosmology[8]. Even those that meet the accuracy requirement are typically limited in the dimension of the parameter space or the dynamical range of the problem. In general, there is a lack of systematic frameworks to

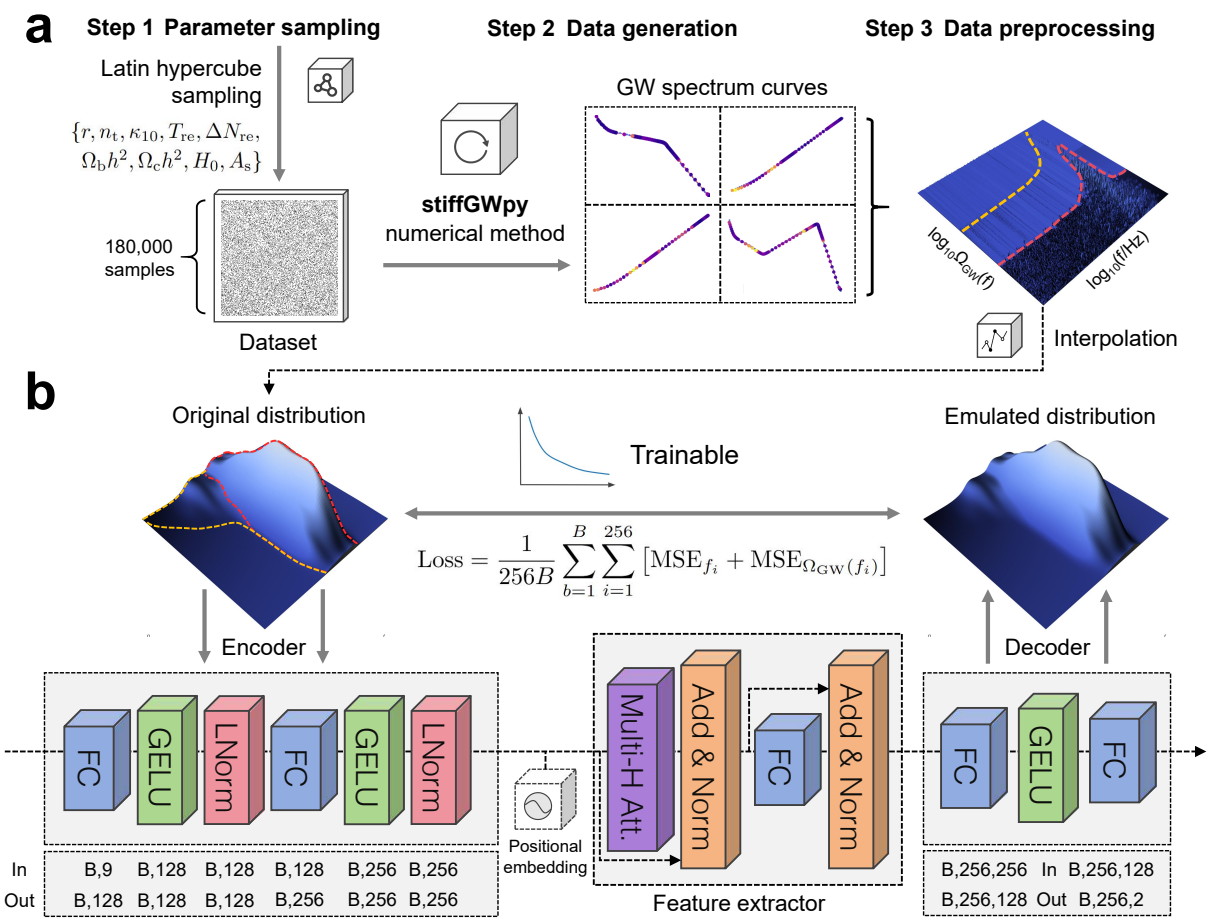


Fig. 1: **Schematic framework of SageNet+. a**, Data generation phase. First, in **Steps 1 and 2**, we perform a nine-dimensional Latin hypercube sampling on the prior distributions of the nine physical parameters, $\{r, n_t, \kappa_{10}, T_{re}, \Delta N_{re}, \Omega_b h^2, \Omega_c h^2, H_0, A_s\}$. The resultant sample points are densely distributed in the parameter space. In **Step 3**, these sampled parameters are fed into the stiffGWpy numerical computation method, generating the energy density spectra of the inflationary SGWB, $\log_{10}\Omega_{GW}$. The distribution of points along each spectral curve is denser in regions where abrupt slope changes occur, which reflects the artificial design of stiffGWpy. **b**, Training phase. In **Steps 4 and 5**, by learning the positions of f_i and $\log_{10}\Omega_{GW}(f_i)$, the network simultaneously assimilates the minimum and maximum frequencies, amplitudes, evolutionary trends and local variational features in the GW spectra for different cosmological parameters.

construct accurate, robust, explainable neural network models that generalize to all 1D emulation tasks in cosmology and astrophysics.

Here we introduce a self-attention-based framework[9] for emulating 1D sequences that result from first-principles calculations. Our model is implemented in the context of the energy spectrum of the inflationary stochastic gravitational-wave background (SGWB), but can generalize to similar problems, e.g., galactic or stellar spectra. The SGWB composed of primordial tensor fluctuations is a key prediction of inflationary cosmology[10–12]. This relic from the early Universe offers a unique probe of cosmic inflation, high energy physics and gravity. A blue-tilted primordial tensor power spectrum with a positive spectral index can produce detectable gravitational-wave (GW) signals across a wide frequency range, from nanohertz bands probed by Pulsar Timing Arrays (PTAs) to tens of hertz accessible to laser interferometers[13, 14]. Recent PTA observations have reported strong evidence for an SGWB signal[15–18], which shows consistency with a blue-tilted inflationary SGWB[19, 20]. Furthermore, a stiff phase or kination (with equation of state $w = 1$) in the early Universe provides additional interpretations for this tentative signal because of the kination-amplification effect on the primordial tensor modes[21–25]. This extended cosmological model can also be tested by current and next-generation laser interferometer detectors such as LIGO-Virgo-KAGRA[26].

The main observable of the inflationary SGWB is its present-day energy density spectrum, $\Omega_{\rm GW}(f)$. Its theoretical modeling involves considerable computational costs due to the backreaction of the GW energy density on the expansion rate[21]. For blue-tilted spectra (either of primordial origin or kination-amplified), the contribution of the inflationary SGWB to the effective number of relativistic species, $\Delta N_{\rm eff,GW}$, becomes observationally significant[22]. This extra radiation component can alter the Hubble constant inferred from cosmological data by a few percent[24], an unacceptable error in view of today's serious Hubble tension[27]. As a result, the backreaction effect must be taken into account self-consistently, and therefore accurate evaluations of $\Omega_{\rm GW}(f)$ require iteratively solving a coupled system of integro-differential equations[25]. Traditional numerical solvers such as the LSODA algorithm[28] cost tens of seconds per evaluation and lack hardware acceleration in the meantime, thus hindering large-scale Bayesian analysis or real-time applications.

Deep learning offers a transformative approach to accelerate such first-principles computations by means of emulation. Given their outstanding capability of feature extraction, neural networks have already excelled in a range of applications in GW astronomy, including GW signal detection, instrumental glitch recognition, and rapid parameter estimation[29–35]. For our purpose here, recurrent[36], convolutional[37], and networks with attention mechanisms[9] are promising architectures for modeling inflationary SGWB spectra of various shapes, aiming at capturing both local features and long-range dependencies over decades in the frequency domain.

In this work, we present SageNet+, a neural network emulator for the energy density spectrum of the inflationary SGWB. The emulator explores the full nine-dimensional parameter space of an extended cosmological model with a possible early stiff phase. This physical model expands both the dimension of the parameter space and the prior ranges in contrast to the original SageNet[7]; see Extended Data Table 1. Based on the Transformer architecture, SageNet+ combines learnable positional embeddings optimized for frequency-domain predictions and adaptive sequence standardization to preserve spectral features. We also introduce new metrics to assess model performance, generally applicable to 1D data of logarithmic representations. Trained on 180,000 numerically computed spectra, SageNet+ attains a subpercent accuracy level with inference times in milliseconds—thousands of times faster than traditional methods. This speed-up enables previously infeasible analysis, including population-level studies based on Monte Carlo sampling, real-time parameter estimation using GW data, and joint analysis with other cosmological data sets. SageNet+ is able to reproduce all the patterns in the 1D spectral sequences imprinted by the first-principles solver, capturing both physical relationships and artificial features. Its physics-informed design hence offers a novel, optimal, and generalizable framework for scientific emulation, applicable to a diversity of problems in cosmology and astrophysics with high-dimensional parameter spaces.

Results

SageNet+ implements a Transformer-based architecture to emulate the present-day energy spectrum, $\Omega_{\rm GW}(f)$, of the stiff-amplified inflationary SGWB, achieving a sub-percent error for almost all combinations of the nine cosmological parameters. The emulator successfully captures all critical features of the spectrum curves, such as characteristic spectral indices and key transitions in the expansion history of the Universe. Unlike traditional numerical methods that take tens of seconds per execution, SageNet+ can deliver the predicted $\Omega_{\rm GW}(f)$ in less than 10 ms on an Intel Core i7-11800H CPU, enabling efficient Bayesian inference of the full set of cosmological parameters. We present multiple analyses on the performance of SageNet+ to assess its accuracy and efficiency.

SageNet+ framework

The two-stage schematic framework of SageNet+ is illustrated in Fig. 1. The primary goal of the emulator is to efficiently output the GW spectrum curve for any set of parameters of the cosmological model. Fig. 1a describes the pipeline for generating the data set. In Step 1, we apply the Latin hypercube sampling (LHS) method to the nine-dimensional parameter space, while the prior probability distributions of the parameters are assumed to be uniform; see Extended Data Table 1. The LHS method ensures a dense and thorough exploration of the parameter space. In Step 2, each set of sampled parameters is passed to the first-principles numerical method, stiffGWpy[24, 25], to calculate accurate $\Omega_{\rm GW}(f)$ curves as the ground truth. All data curves are then processed with a refined interpolation strategy in Step 3 such that each spectrum has a uniform length of 256 points; see the 'Data interpolation' section in the Methods.

The parameters of the training samples and their corresponding GW spectrum curves are subsequently fed into the network module for training and validation, as illustrated in Fig. 1b. We adopt an encoder-decoder architecture for SageNet+. The

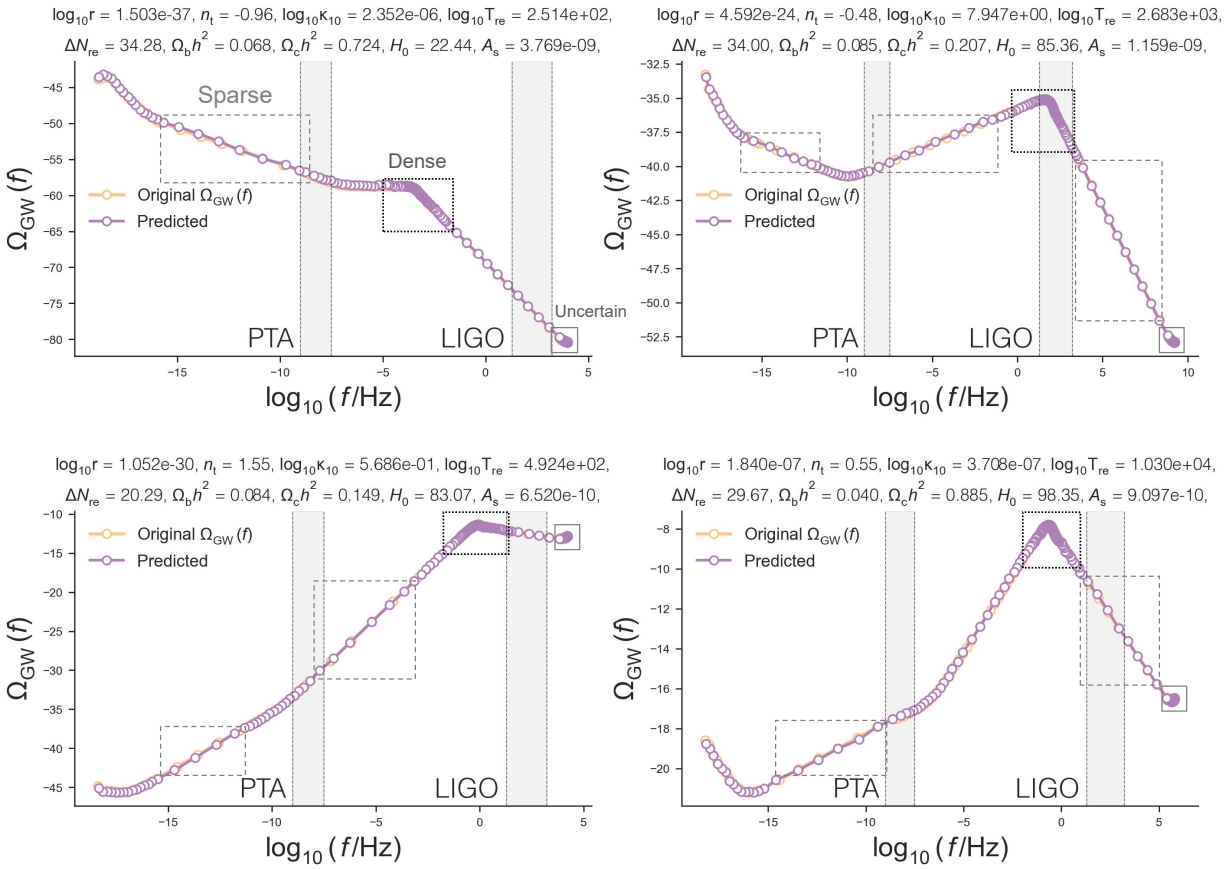


Fig. 2: **Example emulation results of SageNet+ for different physical parameters.** In each case, the predicted and original GW spectra reach perfect agreement, even if the spectral shapes vary significantly across different parameter combinations. We identify three types of frequency intervals in the original spectral curves: sparse, dense, and uncertain. The sparse interval corresponds to the stage in the expansion of the Universe when the equation of state is time-independent, and thus the curve has a constant slope. The dense interval with varying slopes represents the evolutionary stage when the equation of state changes. The uncertain interval at the maximum frequency corresponds to the very early Universe with high energy scales where the physics is uncertain. This part of the spectrum has relatively low reference value for most SGWB searches, but generates significant noise that affects model accuracy. Encouragingly, SageNet+ achieves excellent reconstruction results in all three categories, accurately capturing both the overall evolutionary trends (in the frequency domain) and local artificial patterns.

core feature extractor of this architecture is a Transformer encoder that captures all possible positional relationships between two points in a GW spectrum. The spectrum curves resulting from our first-principles calculations, $\{f_i, \Omega_{\rm GW}(f_i)\}$, have varying lengths and nonuniform densities of points. Some of these features are of physical origin, while others are artificial. For example, physical information including the equation of state of the Universe, the duration of each cosmological era, and the energy scale of inflation is encapsulated in global features such as the spectral indices and the length of every power-law segment. On the other hand, the local feature of density/sparsity of points is mostly an artifact arising from the numerical implementation of the first-principles code. In view of this, our Transformer-based model is designed to extract both physical and artificial information simultaneously using the self-attention mechanism, maximizing the accuracy of prediction.

Qualitative evaluation

SageNet+ achieves accurate emulations of the inflationary SGWB spectrum, as indicated by the comparison between the predicted and original $\Omega_{\rm GW}(f)$ curves in Fig. 2 for a few examples. The predicted curves reach fine-grained agreement with the respective ground truth over the entire frequency ranges that encompass the key frequency bands probed by PTA and LIGO. These results show that the model can reconstruct complex GW spectra with a variety of shapes for the full range of cosmological parameters. Furthermore, variations in the

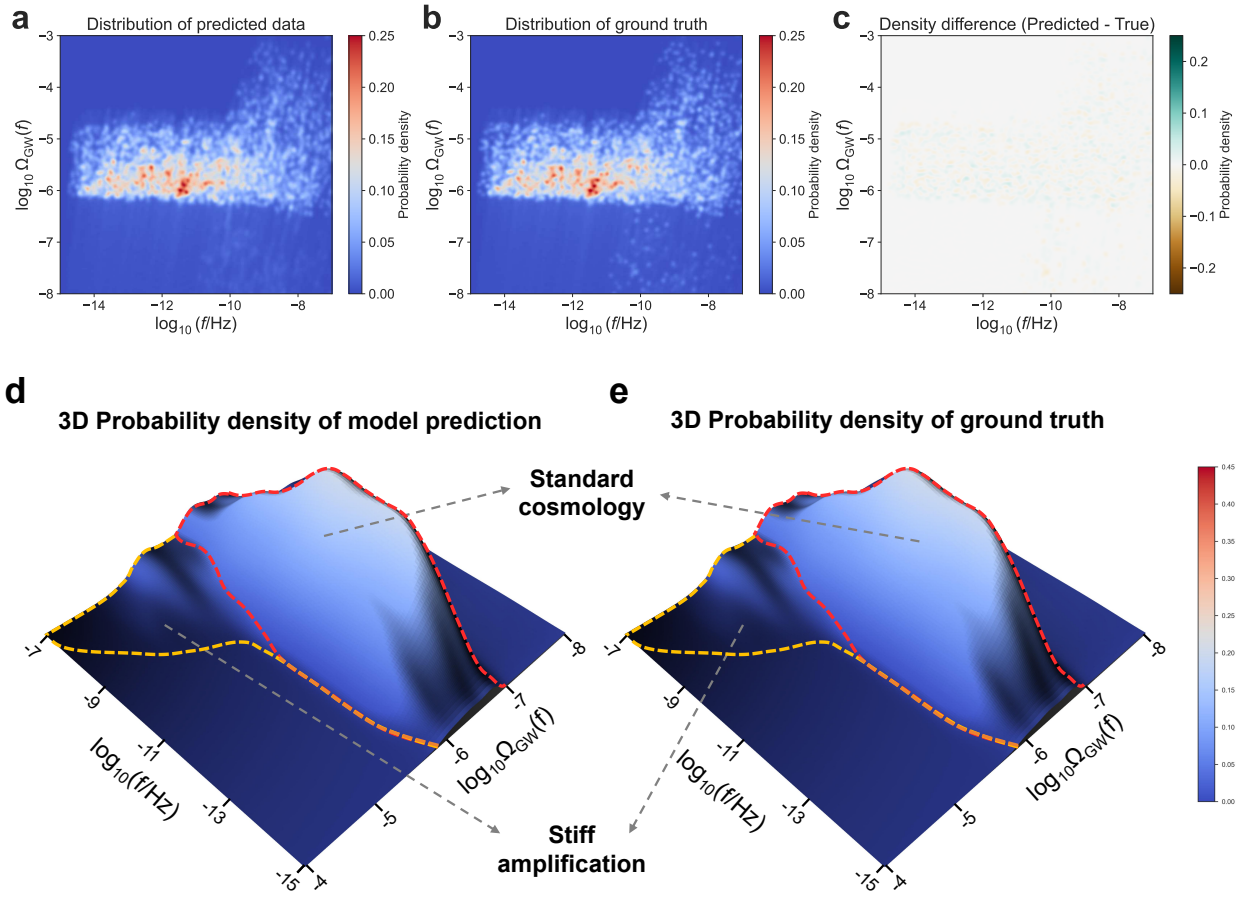


Fig. 3: **Distribution density of GW curve points.** We present the distribution density of 10,000 spectral curves generated through our LHS parameter sampling and first-principles method, alongside the corresponding predicted distribution by SageNet+. **a** and **b**, The two-dimensional form of the density distribution is displayed. **c**, The relative difference in probability density between the two is negligible, within 0.3% in most regions. **d** and **e**, The red dashed lines delineate the primary distribution region of the curves due to standard ΛCDM cosmology, where the vast majority of $\{\log_{10}(f_i/\mathrm{Hz}), \log_{10}\Omega_{\mathrm{GW}}(f_i)\}$ points are located. Within this region, SageNet+ achieves excellent emulation of the ground truth. The area between the red and the yellow dashed lines represents the region due to stiff/kination amplification of the inflationary SGWB. The probability density in this high-frequency region is relatively lower, yet the emulator is still capable of reproducing this physical feature.

density of points in each spectrum curve are reliably captured and reproduced by SageNet+. As mentioned above, such variations are of artificial origin, ascribed to the adaptive frequency sampling strategy employed in the numerical algorithm (StiffGWpy). Remarkably, even in sparsely sampled parts of the curves, the Transformer-based model is able to predict the locations of data points with high precision.

Fig. 3 provides another perspective for comparing the predicted GW spectra with the ground truth in terms of the distribution of all data points, $\{f_i, \Omega_{\mathrm{GW}}(f_i)\}$, in 10,000 curves. The distribution of the spectrum points reconstructed by SageNet+ (Fig. 3a) is perfectly consistent with the original distribution (Fig. 3b), reproducing features such as the stiff/kination amplification of the inflationary SGWB at high frequencies and the horizontal hill across lower frequencies attributed to the standard cosmology without kination (Fig. 3d,e).

Quantitative evaluation by metrics

To evaluate the performance of SageNet+ quantitatively, we devise multiple new metrics to assess the accuracy of the predicted GW spectra against the ground truth. In this work, we introduce the mean logarithmic area error (MLAE), defined as the logarithmic average difference between the predicted

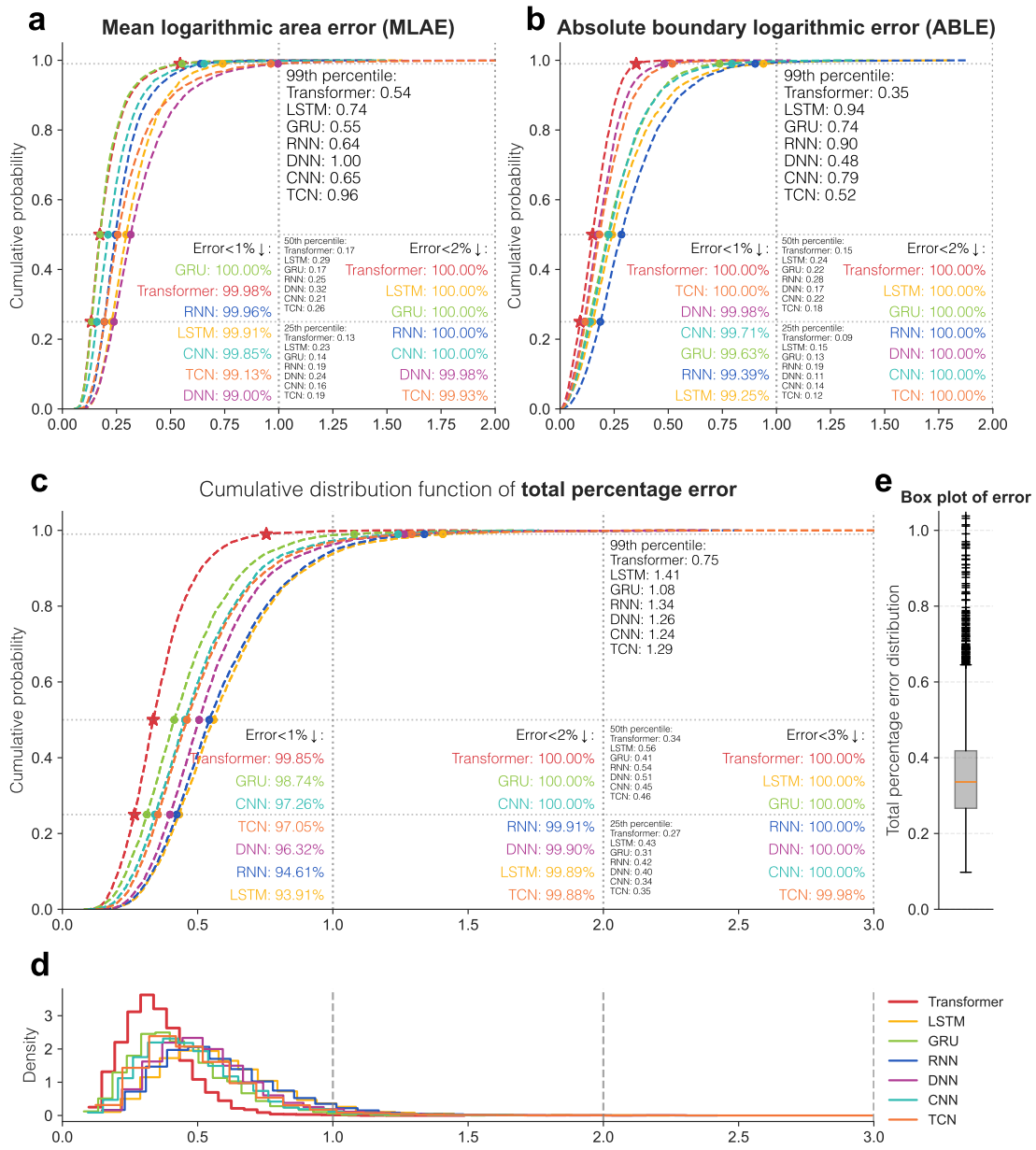


Fig. 4: **Quantitative model performance by metrics. a** and **b**, We report the CDFs of MLAE and ABLE for SageNet+ (Transformer) and other feature extractors on the same test set. **c** and **d**, We report the CDF and the trapezoidal diagram of total percentage errors. The attention layer of Transformer results in the best performance, with 99.8% of samples (red curve) achieving a relative error below 1%. **e**, We show the box plot of the error distribution for SageNet+. These results highlight variations in the learning capabilities of different neural network architectures. Recurrent networks excel at learning long-distance information in the middle segment of GW spectra, leading to high MLAE scores. Convolutional networks, on the other hand, are proficient at recognizing local features including those in the head and tail segments (near $f_{\rm min}$ and $f_{\rm max}$), thus achieving good ABLE performance. Overall, all models demonstrate robust emulations of GW spectra under our unified framework, while the Transformer-based model, due to its self-attention mechanism, exhibits the optimal performance in all metrics.

and true spectra over the entire frequency range,

$$\mathrm{MLAE} \equiv \frac{1}{\ln(f_{\rm max}/f_{\rm min})} \int_{f_{\rm min}}^{f_{\rm max}} \left| \ln \frac{\Omega_{\rm GW}^{\rm pred}(f)}{\Omega_{\rm GW}^{\rm true}(f)} \right| \mathrm{d}\ln f \times 100\%, \quad (1)$$

This metric is mathematically proportional to the area between the two $\log_{10}\Omega_{\rm GW}(f)$ curves, as illustrated by Extended Data Fig. 1. Meanwhile, it represents the (logarithmic) mean absolute percentage error in the predicted value of $\Omega_{\rm GW}(f)$ across its frequency range, since $\ln(1+x) \approx x$ for

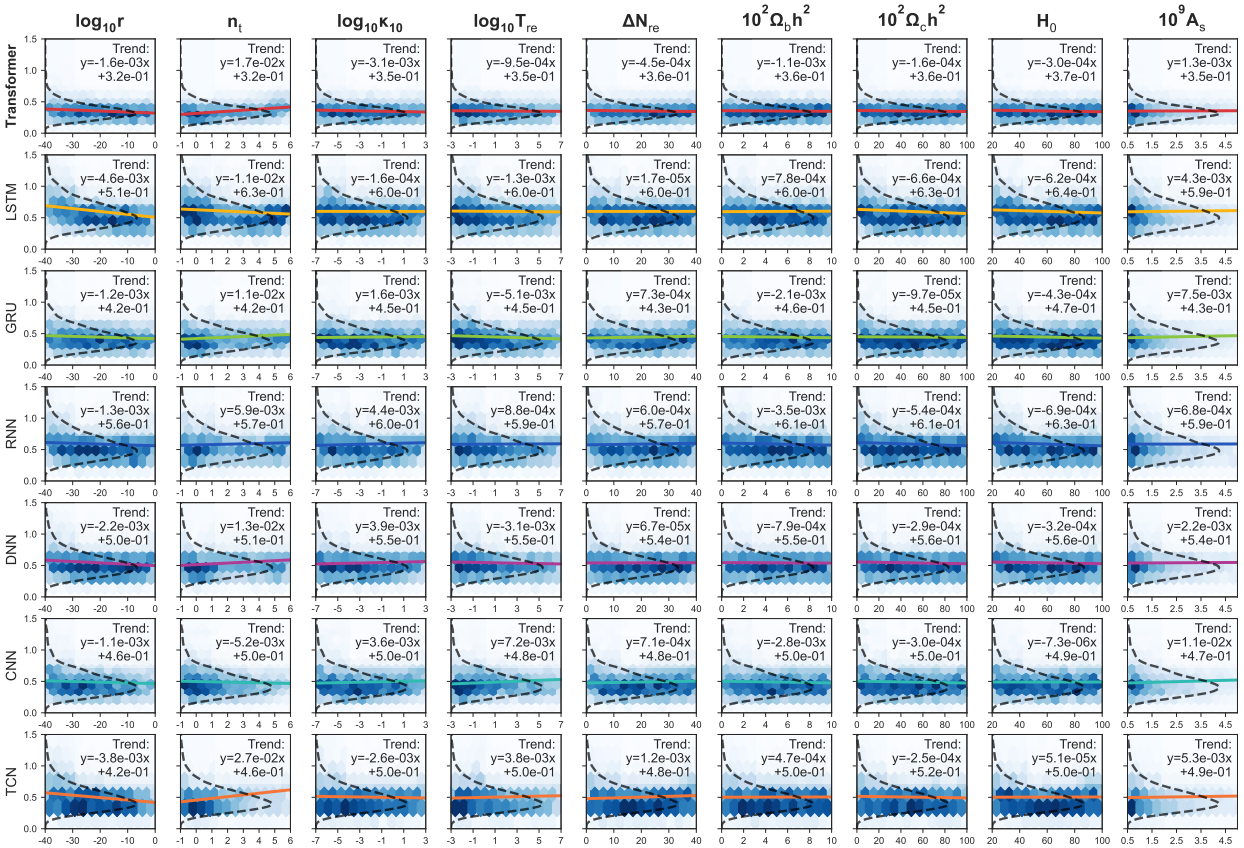


Fig. 5: **Dependence of model performance on physical parameters.** For the test set with 5,400 samples, the 2D distribution of the total percentage error and the value of one parameter is illustrated by color in each subplot, where the solid line and the equation describe the linear fit to the distribution. Each column represents a cosmological parameter and each row a neural network architecture. The black dashed curve in each subplot illustrates the 1D probability density function. The linear fits show that an increase in $n_{\rm t}$ introduces additional errors systematically, while the impact of other parameters is minimal. Considering that a bluer spectral index (larger $n_{\rm t}$) compresses the frequency range of the primordial spectrum of linear tensor fluctuations by a cutoff at lower frequencies[25], this effect is expected.

small x. Hence, MLAE embodies a deep connection between the logarithmic quantity output by the model ($\log_{10}\Omega_{\rm GW}$) and the original quantity ($\Omega_{\rm GW}$); the former manifests intrinsic physical features (e.g., spectral indices) but the latter is what GW experiments directly measure. The predictions of SageNet+ achieve a 99th percentile (P99) MLAE of 0.54% for test samples with various combinations of parameters (Fig. 4a), demonstrating excellent agreement with the ground truth.

Technically, the above MLAE applies only to the *overlapped* part of the predicted and true GW spectra in the frequency domain. To take into account the mismatch between the predictions and the ground truth at the two ends of the GW spectra, we consider the following absolute logarithmic error (ALE) for both the $f_{\rm min}$ and $f_{\rm max}$ predictions:

$$\mathrm{ALE}_{\rm min} \equiv \left|\ln \frac{f^{\rm pred}_{\rm min}}{f^{\rm true}_{\rm min}}\right|, \quad \mathrm{ALE}_{\rm max} \equiv \left|\ln \frac{f^{\rm pred}_{\rm max}}{f^{\rm true}_{\rm max}}\right|. \tag{2}$$

We call their sum the absolute boundary logarithmic error (ABLE), i.e., $\mathrm{ABLE} \equiv \mathrm{ALE}_{\rm min}+\mathrm{ALE}_{\rm max}$. It describes the error in (the boundaries of) the predicted frequency range and serves as a complementary metric to MLAE. SageNet+ achieves a P99 ABLE of 0.35% for the test set (Fig. 4b).

Finally, the total percentage error defined as the sum of MLAE and ABLE provides a comprehensive and balanced assessment of the ability of an emulator to recover the true GW spectrum. This metric reaches zero only when the predicted $\Omega_{\rm GW}(f)$ perfectly coincides with the true one. For all three metrics, we present the results of our Transformer-based SageNet+ model as well as other baselines for comparison [36–40]. In particular, the resultant

cumulative distribution function (CDF) illustrated in Fig. 4c shows that the overall performance of Transformer is the best among all these other architectures considered in the literature, indicated by the steepest ascent in its CDF. The P99 total percentage error of SageNet+ is 0.75%, and 99.85% of the test samples exhibit errors below 1%. Detailed error statistics are listed in Extended Data Table 2.

Dependence of SageNet+ performance on cosmological parameters

We further investigate the dependence of the emulator performance on each parameter of the full cosmological model for SageNet+ as well as existing benchmarks (Fig. 5). For most parameters, the distribution of the total percentage error (defined above) does not exhibit a pronounced dependence, as evidenced by the nearly flat trend of the linear fits. For each neural network model, the intercepts of these fits are also nearly equal across the nine cosmological parameters, e.g., $\sim 0.35\%$ for the Transformer-based SageNet+. However, the tensor spectral index, $n_{\rm t}$, has a notable impact on the error distribution. The positive-slope trend shown in the column '$n_{\rm t}$' of Fig. 5 and Extended Data Fig. 2 is consistent with the following physical interpretation. The steeper the inflationary tensor spectrum (with higher value of $n_{\rm t}$), the lower the maximum frequency ($f_{\rm max}$), as a result of the ultraviolet cutoff due to the linearity requirement on the inflationary tensor power (Supplementary Note)[25]. It leads to a lower value of the denominator in MLAE (equation 1), $\ln(f_{\rm max}/f_{\rm min})$, thus increasing the prediction error. This physical interpretability supports the usage of MLAE as a proper metric for such 1D emulation tasks, in general.

Optimized data preprocessing with respect to physical information

To implement the Transformer architecture, we preprocess all numerically-computed GW spectra by interpolation to reach a uniform number of points for each data sequence, $\{f_i, \Omega_{\rm GW}(f_i)\}$ (Step 3 in Fig. 1a). In contrast to conventional data preprocessing methods such as padding (commonly used in natural language processing)[9, 41–43], an interpolation scheme relieves the model of the task of predicting the curve lengths and, in the meantime, helps achieve faster convergence in training (Extended Data Fig. 3).

In addition, SageNet+ adopts an improved interpolation method compared with the original SageNet approach, which applied univariate interpolation in the *entire* frequency domain[7]. To be more specific, the previous approach inserted points between the closest pair of points in every GW spectrum sequence with the intention of preserving the artificial frequency sampling patterns from the numerical code (StiffGWpy). This algorithm often resulted in interpolation near $f_{\rm max}$. However, the high-frequency end of the spectrum is determined by uncertain high-energy physics in the early Universe (Fig. 2). Hence, an overly dense insertion of points near $f_{\rm max}$ actually compromises the true physical information carried by the data sequence. To fix this flaw, our optimized algorithm in SageNet+ restricts the extent of interpolation to the lower 60% of the logarithmic frequency range of each spectrum. This new approach preserves all the physical features of $\Omega_{\rm GW}(f)$ and does not introduce artificial noise for deep learning.

The effects of the optimized interpolation strategy on emulator performance are illustrated in Fig. 6. As we weaken the weight of the high-frequency end of the GW spectra by the optimized prescription for interpolation (Fig. 6a), the Transformer-based model pays more attention to parts of the curves that are physically more robust, thereby improving the accuracy of emulations (Fig. 6b). In fact, Fig. 6c shows that with the new method, not only SageNet+ but most of the models attain better performance as their MLAEs and ABLEs are reduced, demonstrating the ability of the optimized data preprocessing scheme to capture genuine physical information in GW spectrum sequences.

Computational Efficiency

SageNet+ enables an average inference time of approximately 6.55 ms on an Intel Core i7-11800H CPU (see Extended Data Fig. 4), the same order of magnitude compared to the original SageNet[7], while extending the dimension of the parameter space by almost a factor of two (9D versus 5D). The Transformer-based model thus maintains a significant speed advantage over traditional numerical methods (e.g., stiffGWpy) that require iterative numerical integration of differential equations and typically cost $\mathcal{O}(10)$ seconds per run. As a result, SageNet+ allows for rapid parameter space exploration and subsequent Bayesian inferences. The framework can be further generalized to a wide range of high-dimensional inference tasks with extended cosmological models, where efficient and accurate theoretical predictions are vital for analyzing large GW and cosmological data sets.

Discussion

We introduce a Transformer-based emulator, SageNet+, for generating fast and accurate predictions of the energy spectrum of inflationary

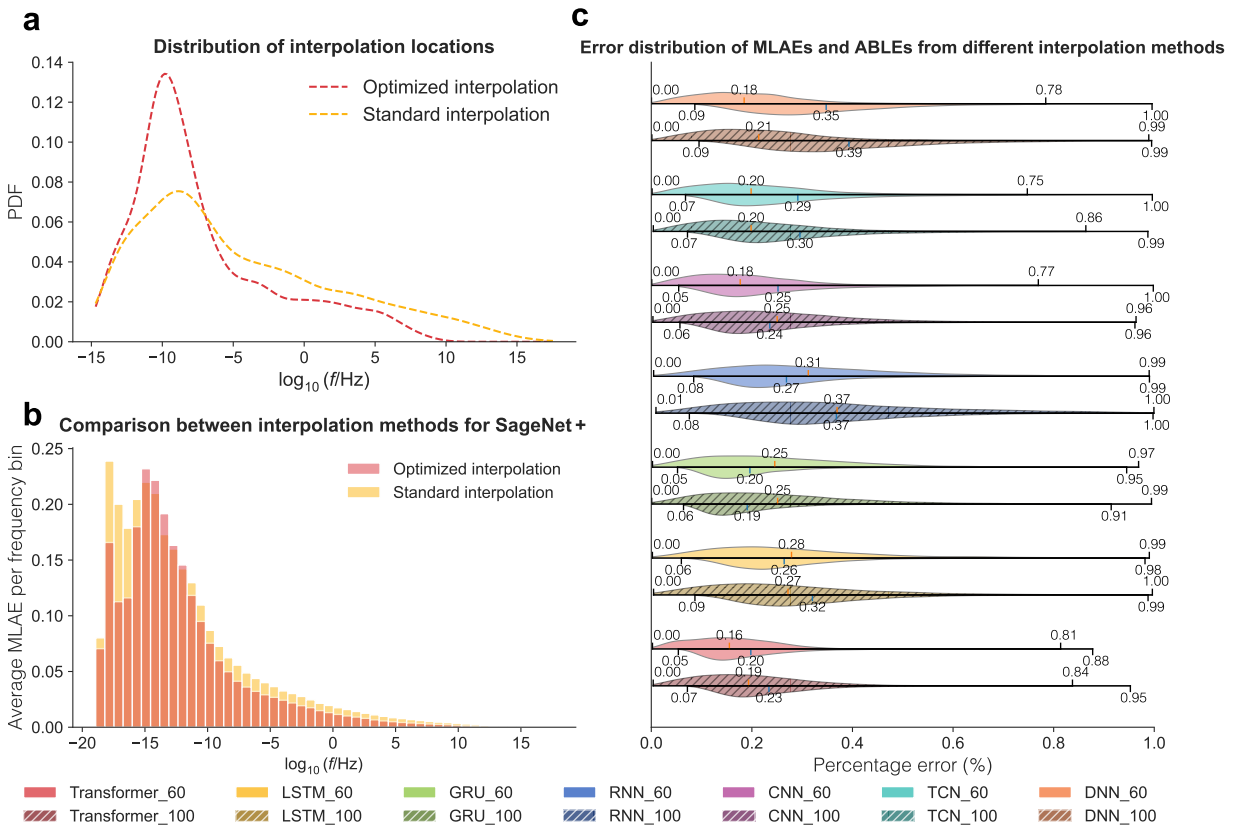


Fig. 6: **Effects of optimized data preprocessing on model performance.** The optimized interpolation strategy implemented in SageNet+ results in lower average errors. **a**, We show the probability density functions (PDFs) of the interpolated frequencies for two different methods. Our optimized interpolation method places more weight on the low-to-mid frequency range to ensure greater model attention in this band. **b**, Correspondingly, we report the histograms of the mean logarithmic area errors over frequency bins for the two strategies. When the GW spectrum sequences are preprocessed with the optimized interpolation method, the trained emulator achieves higher accuracy in the low-to-mid frequency range while retaining low errors at high frequencies. **c**, The performance of nearly all neural network models is enhanced by the optimized interpolation method. The unshaded violin in each set of experiments represents the error distribution resulting from the optimized method.

gravitational waves with possible stiff amplification. Not only does it retain the efficiency level of $\mathcal{O}(1)$ ms per emulation on CPUs, SageNet+ also achieves enhanced accuracy, proving itself as a robust alternative to conventional first-principles numerical solvers for sequential data.

Demands for efficient emulation of spectra or 1D sequences are now ubiquitous in many branches of cosmology and astrophysics [2–6, 44–46]. The results in this work demonstrate the advantage of Transformer-based models over other benchmarks in such tasks[47, 48]. Self-attention is the key here, given its competence in learning both global and local patterns of a sequence. In our unified SageNet+ framework, there are primarily two sources of errors: (i) the difference between the predicted and true GW spectra in the overlapped part of their frequency ranges, and (ii) the mismatch between the boundaries of the frequency ranges, $f_{\rm min}$ and $f_{\rm max}$. Neural networks with recurrent architectures, e.g., RNN, LSTM and GRU, can effectively capture long-distance information in the middle part of a sequence, leading to good performance in the MLAE metric (Fig. 4a). However, they have difficulty extracting precise local patterns near the boundaries of frequency domains because the two ends of a GW spectrum lack sufficient contextual information and also suffer from intrinsic uncertainties of physical origin. These reasons lead to relatively lower ABLE scores for networks with recurrent architectures (Fig. 4b). Neural networks with convolutional architectures such as CNN and TCN, in contrast, exhibit stronger performance in the ABLE metric, since these models excel at capturing local features and can therefore keep track of points near the boundaries of a spectrum curve. Nevertheless, their limited receptive fields hinder the reconstruction of global functional

dependencies in $\Omega_{\rm GW}(f)$, resulting in lower MLAE scores. Unlike the above two classes of benchmarks, Transformer exploits its self-attention mechanism to directly learn the correlations between different points in a sequence. By dynamically adjusting the scope of attention, our Transformer-based model has proved a great success in extracting both long- and short-distance features in $\Omega_{\rm GW}(f)$, bringing about superior performance in both metrics, MLAE and ABLE. By the same token, all these merits of self-attention are generally applicable to modeling other sequential data in similar studies.

Furthermore, the above two types of metrics can also be generalized to other 1D emulation tasks. In fact, MLAE and ABLE together provide an optimized framework for evaluating model performance in such tasks. Because 1D spectra/sequences in cosmology and astrophysics are essentially continuous functions for which the specific locations of points are unimportant, a sound characterization of emulation accuracy is presented by the two aforementioned sources of errors rather than common metrics such as the mean absolute percentage error (MAPE). The latter is typically dependent on all the particular points in a sequence and is thus suboptimal for assessing the errors in modeling continuous functions (Extended Data Fig. 1). Unlike these standard metrics, our approach here correctly captures the difference between a predicted curve and its ground truth. In doing so, we take advantage of the power-law nature of the GW spectrum segments to construct the MLAE based on the logarithmic quantity, $\ln \Omega_{\rm GW}(f)$. The MLAE can be interpreted as the average log accuracy ratio over the frequency span of each spectrum (equation 1)[49, 50], while geometrically equal to the area between the two log curves. Similarly, the design of ABLE also utilizes the logarithmic attribute of the frequency range. Hence, as long as the data sequences are suitable for logarithmic representations by nature, the new metrics devised in this work offer a favorable measure of model performance.

SageNet+ employs an interpolation scheme for data preprocessing. It is a necessary step in emulation tasks, since the sequential data computed by first-principles methods do not always have a uniform length for varying parameters. Our interpolation approach is again supported by the fact that 1D sequences are continuous functions for which the total number of points in a curve is of no significance as long as there are sufficient points to capture all the features. Compared with common approaches such as padding or truncation, interpolation-based data preprocessing can maximally preserve physical features in a sequence. In addition, our optimized method for SageNet+ with a selective frequency range for interpolation results in a considerable improvement on the previous standard method (Fig. 6). By paying less attention to the physically uncertain, high-frequency end of GW spectra, the Transformer-based model achieves better performance in learning genuine cosmological information. Therefore, it is important to optimize the data preprocessing method in synergy with the optimal neural network architecture. The optimization here based on physical information is generalizable to other 1D data in related fields.

The SageNet+ framework provides a systematic, explainable solution to generic 1D emulation tasks or forward modeling problems in cosmology and astrophysics, along with novel metrics for robust model evaluation. Our results show that with the aid of optimized data preprocessing, state-of-the-art Transformer-based emulators currently yield the optimal predictions of 1D spectra/sequences.

Methods

Data generation

Parameter sampling. To construct a comprehensive training dataset for modeling GW spectra, we generated 180,000 samples spanning a nine-dimensional parameter space using Latin hypercube sampling, a stratified sampling technique that ensures uniform coverage of the parameter space by dividing each dimension into equally probable intervals and sampling once from each interval [51]. We used the `pyDOE` Python package with the 'maximin' criterion to maximize the minimum distance between samples, thereby improving the diversity of the generated dataset. The parameter space encompasses nine cosmological and inflationary parameters critical to GW spectra modeling: the tensor-to-scalar ratio (r), the tensor spectral index ($n_{\rm t}$), the curvature perturbation amplitude at $k = 10\,{\rm Mpc}^{-1}$ (κ_{10}), the reheating temperature ($T_{\rm re}$), the number of e-folds during reheating ($\Delta N_{\rm re}$), the baryon density ($\Omega_{\rm b}h^2$), the cold dark matter density ($\Omega_{\rm c}h^2$), the Hubble constant (H_0), and the amplitude of scalar perturbations ($A_{\rm s}$). Each parameter is sampled according to physically motivated prior distributions, which are either linearly or logarithmically scaled depending on their expected range and physical constraints.

The LHS-generated parameters were mapped to their physical ranges using appropriate scaling functions. For logarithmically sampled parameters, the transformation follows $\theta = 10^{a+x(b-a)}$, where $x \in [0, 1)$ is the LHS sample, and $[10^a, 10^b]$ is the prior range. For linearly sampled parameters, the transformation is $\theta = a+x(b-a)$, where $[a, b]$ is the prior range. The resulting parameter sets were stored

in a database for subsequent GW spectra computation. The Python implementation leverages the `pyDOE.lhs` function, with the code adapted to handle both linear and logarithmic scalings efficiently. This approach ensures that the training dataset captures the full variability of GW spectra across the parameter space, enabling robust modeling of cosmological and inflationary effects. The generated samples serve as inputs to a numerical pipeline that computes the corresponding GW spectra, which are used to train our predictive model.

Data interpolation. In the dataset, most curves have a variable unit length, between 100 and 250. The variable length sequences introduce additional challenges for network training. In the context of SGWB, the nature of the curves is continuous, and the number of points does not affect their physical accuracy. Therefore, we attempted to fix the curve length to a specific value. In our network, this value was chosen as $N = 256$, with all curves having a unit length less than or equal to this value. For a curve with original length $M < 256$, represented as a sequence of points $\mathbf{c} = \{\mathbf{c}_i = (x_i, y_i) \mid i = 0, 1, \ldots, M-1\}$, we performed dense linear interpolation between the two most sparsely spaced points to extend the curve to $N = 256$ points. Specifically, let $\mathbf{c}_k$ and $\mathbf{c}_{k+1}$ be the pair of adjacent points with the maximum Euclidean distance:

$$d_{k,k+1} = \max_{i=0,\ldots,M-2} \|\mathbf{c}_{i+1} - \mathbf{c}_i\|_2. \tag{3}$$

We insert $N - M$ new points between $\mathbf{c}_k$ and $\mathbf{c}_{k+1}$ using linear interpolation. For a new point $\mathbf{c}'_j$ at interpolation step $t_j = \frac{j}{N-M+1}$ (for $j = 0, 1, \ldots, N-M$), the coordinates are computed as:

$$\mathbf{c}'_j = (1 - t_j)\mathbf{c}_k + t_j \mathbf{c}_{k+1}. \tag{4}$$

This ensures the interpolated segment preserves the slope between $\mathbf{c}_k$ and $\mathbf{c}_{k+1}$, as the points lie on the straight line connecting them. The resulting sequence $\mathbf{c}' = \{\mathbf{c}'_0, \mathbf{c}'_1, \ldots, \mathbf{c}'_{N-1}\}$ has exactly $N = 256$ points, with the original curve's shape and slope information retained as much as possible. Compared to variable-length input sequences, this standardization to a fixed length of $N = 256$ allows us to retain the maximum amount of curve information while reducing the model's additional computational burden and errors in the attention-enhanced neural network.

To maximize the representation of the physical information of the curve while reducing the number of sampling points, the StiffGWpy algorithm (see Algorithm 1 in the Supplementary Notes) appropriately increases the interpolation frequency when $\log_{10} \Omega_{\rm GW}(f_i)$ undergoes significant changes. However, when the evolution process is in the high-frequency boundary stage, the curve inevitably carries strong uncertainty. This uncertainty information is accurately captured by the StiffGWpy algorithm but lacks practical physical significance for the study of the evolution of SGWB.

Therefore, interpolating noisy high-frequency regions which carries high uncertainties will lead to deviations from the ground truth. To address this, we restrict interpolation to the first 60% of the curve's frequency range, prioritizing denser sampling in the middle portion to enhance the representation of physically relevant regions. The interpolation method in SageNet+ is illustrated in Extended Data Fig. 5. This approach increases the weight of reliable data during training, prevents the model from learning incorrect information due to computational uncertainties, and avoids oversampling noisy regions that could diminish the contribution of other curve segments.

Neural network architecture

To emulate the stiff-amplified inflationary GW spectrum, we developed a Transformer-based neural network model to handle the expanded nine-dimensional parameter space. This model replaces traditional recurrent architectures, such as LSTM networks, with a Transformer framework to enhance accuracy and computational efficiency in predicting the present-day energy density spectrum, $\Omega_{\rm GW}(f)$.

Parameter encoder. The input is a nine-dimensional tensor:

$$\mathbf{p} = \{r, n_{\rm t}, \kappa_{10}, T_{\rm re}, \Delta N_{\rm re}, \Omega_{\rm b}h^2, \Omega_{\rm c}h^2, H_0, A_{\rm s}\}, \tag{5}$$

where r is the tensor-to-scalar ratio, $n_{\rm t}$ is the tensor spectral index, κ_{10} parameterizes the kination-to-radiation transition, $T_{\rm re}$ is the reheating temperature, $\Delta N_{\rm re}$ is the number of e-folds during reheating, and $\Omega_{\rm b}h^2$, $\Omega_{\rm c}h^2$, H_0, and $A_{\rm s}$ are standard cosmological parameters. The encoder transforms this input via linear layers with GELU activations and layer normalization:

$$\mathbf{h}_0 = \text{LayerNorm}\left(\text{GELU}\left(\mathbf{W}_2 \cdot \mathbf{p_0}\right)\right), \tag{6}$$

$$\mathbf{p_0} = \text{LayerNorm}\left(\text{GELU}\left(\mathbf{W}_1 \cdot \mathbf{p}\right)\right) \tag{7}$$

where $\mathbf{W}_1 \in \mathbb{R}^{128\times 9}$ and $\mathbf{W}_2 \in \mathbb{R}^{256\times 128}$. The output $\mathbf{h}_0 \in \mathbb{R}^{256}$ is a compact representation.

Feature extractor. The feature extractor is designed to capture the physical dependencies within the SGWB spectrum, such as frequency-dependent spectral indices and transitions across cosmological epochs. In SageNet+, we use the attention layer as the primary feature extractor.

The input sequence is prepared by repeating the initial hidden state $\mathbf{h}_0$ across 256 positions and adding positional embeddings:

$$\mathbf{pos} = \text{Embedding}(\text{positions}), \tag{8}$$

$$\mathbf{seq} = \mathbf{h}_0 + \mathbf{pos}. \tag{9}$$

This sequence is processed through a Transformer encoder with three layers, each featuring multi-head self-attention and feed-forward networks. To reduce network complexity while ensuring the accuracy, in SageNet+, each attention layer is configured with a feature dimension of 256, 8 attention heads, a feed-forward dimension of 512, a dropout rate of 0.1, and uses the GELU activation function.

The multi-head self-attention mechanism effectively models long-range correlations in the SGWB spectrum, such as those arising from kination epochs or reheating transitions, producing a contextualized sequence, $\mathbf{Embedded} \in \mathbb{R}^{\text{batch}\times 256\times 256}$, which encodes these physical relationships across the nine-dimensional cosmological parameter space.

Decoder. The decoder projects the attention layer's output to predict frequency and spectrum value pairs, reflecting the physical characteristics of the SGWB spectrum:

$$\hat{\mathbf{y}} = \mathbf{W}_4 \cdot \text{GELU}\left(\mathbf{W}_3 \cdot \mathbf{Att.}\right), \tag{10}$$

where $\mathbf{W}_3 \in \mathbb{R}^{128\times 256}$ and $\mathbf{W}_4 \in \mathbb{R}^{2\times 128}$. The output $\hat{\mathbf{y}} \in \mathbb{R}^{\text{batch}\times 256\times 2}$ provides predicted $\{f_i, \log_{10}\Omega_{\text{GW}}(f_i)\}$ pairs, accurately capturing spectral features like power-law behavior and epoch transitions.

This architecture leverages the ability of the attention layer to extract physically relevant dependencies, allowing SageNet+ to achieve high accuracy and scalability in modeling the inflationary SGWB spectrum across a diverse cosmological parameter space.

Acknowledgements. FZ would like to thank the MIT LIGO Laboratory for its continuous support and advice. FZ is supported by Ministry of Science and Technology of the People's Republic of China (Grant No. 2023ZD0120704 of Project No. 2023ZD0120700) and the National Natural Science Foundation of China (Grant No. 62372409). BL is supported by Guangxi Key Research and Development Program (Guike FN2504240040), the Guangxi Natural Science Foundation (grant No. 2023GXNSFBA026114) and the National Natural Science Foundation of China (Grant Nos. 12203012, 12494575). This work is also supported by the Guangxi Talent Program ("Highland of Innovation Talents"). JM is supported by the US Department of Energy under Grant DE-SC0010129 and by NASA through Grant 80NSSC24K0665. PRS acknowledges support from NASA under Grant No. 80NSSC22K175.

Data Availability. We provide the data set used in writing this paper, including a training set with 180,000 samples and a test set with 5,400 samples[52]. The data in our study were generated using the `stiffGWpy` numerical computation algorithm (https://github.com/bohuarolandli/stiffGWpy), and we have uploaded the code, including the steps to generate the samples on GitHub (https://github.com/YifangLuo/SageNet).

Code Availability. To facilitate community access, we have released our emulator as an open-source package, `sagenetgw`, available on PyPI at https://pypi.org/project/sagenetgw, enabling seamless integration into cosmological and gravitational wave analyses. The original code has been open sourced to Github: https://github.com/YifangLuo/SageNet.

References

[1] Grzeszczuk R, Terzopoulos D, Hinton G. NeuroAnimator: Fast Neural Network Emulation and Control of Physics-Based Models. In: Cohen M, editor. Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '98. ACM; 1998. p. 9–20.

[2] Ting YS, Conroy C, Rix HW, Cargile P. The Payne: self-consistent ab initio fitting of stellar spectra. The Astrophysical Journal. 2019;879(2):69. Publisher: The American Astronomical Society. https://doi.org/10.3847/1538-4357/ab2331.

[3] Knabenhans M, Stadel J, Potter D, Dakin J, Hannestad S, Tram T, et al. Euclid preparation : IX. EuclidEmulator2 - Power spectrum emulation with massive neutrinos and self-consistent dark energy perturbations. Monthly Notices of the Royal Astronomical Society. 2021;505(2):2840–2869. Publisher: Oxford University Press. https://doi.org/10.1093/mnras/stab1366.

[4] Günther S, Lesgourgues J, Samaras G, Schöneberg N, Stadtmann F, Fidler C, et al. CosmicNet II: emulating extended cosmologies with efficient and accurate neural networks. J Cosmol Astropart Phys. 2022 nov;2022(11):035. Publisher: IOP Publishing. https://doi.org/10.1088/1475-7516/2022/11/035.

[5] Spurio Mancini A, Piras D, Alsing J, Joachimi B, Hobson MP. COSMOPOWER: emulating cosmological power spectra for accelerated Bayesian inference from next-generation surveys. Mon. Not. R. Astron. Soc. 2022 Apr;511(2):1771–1788. https://doi.org/10.1093/mnras/stac064. arXiv:2106.03846. [astro-ph.CO].

[6] Zhong K. Attention-based neural network emulators for multiprobe data vectors. I. Forecasting the growth-geometry split. Phys. Rev. D. 2025;111(12). https://doi.org/10.1103/PhysRevD.111.123519.

[7] Zhang F, Luo Y, Li B, Cao R, Peng W, Meyers J, et al. SageNet: Fast Neural Network Emulation of the Stiff-amplified Gravitational Waves from Inflation. Astrophys. J. Suppl. Ser. 2025 Aug;279(2):44. https://doi.org/10.3847/1538-4365/ade4c6. arXiv:2504.04054. [astro-ph.CO].

[8] Planck Collaboration, Aghanim N, Akrami Y, Ashdown M, Aumont J, Baccigalupi C, et al. Planck 2018 results. VI. Cosmological parameters. Astron. Astrophys. 2020 Sep;641:A6. https://doi.org/10.1051/0004-6361/201833910. arXiv:1807.06209. [astro-ph.CO].

[9] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al.. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al., editors.: Attention Is All You Need. arXiv. Available from: http://arxiv.org/abs/1706.03762.

[10] Starobinsky AA. A new type of isotropic cosmological models without singularity. Physics Letters B. 1980 Mar;91(1):99–102. https://doi.org/10.1016/0370-2693(80)90670-X.

[11] Guth AH. Inflationary universe: A possible solution to the horizon and flatness problems. Phys. Rev. D. 1981 Jan;23(2):347–356. https://doi.org/10.1103/PhysRevD.23.347.

[12] Linde AD. A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems. Physics Letters B. 1982 Feb;108(6):389–393. https://doi.org/10.1016/0370-2693(82)91219-9.

[13] Hellings RW, Downs GS. Upper limits on the isotropic gravitational radiation background from pulsar timing analysis. Astrophys. J. Lett. 1983 Feb;265:L39–L42. https://doi.org/10.1086/183954.

[14] Meerburg PD, Hložek R, Hadzhiyska B, Meyers J. Multiwavelength constraints on the inflationary consistency relation. Phys. Rev. D. 2015 May;91(10):103505. https://doi.org/10.1103/PhysRevD.91.103505. arXiv:1502.00302. [astro-ph.CO].

[15] Agazie G, Anumarlapudi A, Archibald AM, Arzoumanian Z, Baker PT, Bécsy B, et al. The NANOGrav 15 yr Data Set: Evidence for a Gravitational-wave Background. Astrophys. J. Lett. 2023 Jul;951(1):L8. https://doi.org/10.3847/2041-8213/acdac6. arXiv:2306.16213. [astro-ph.HE].

[16] EPTA Collaboration, InPTA Collaboration, Antoniadis J, Arumugam P, Arumugam S, Babak S, et al. The second data release from the European Pulsar Timing Array. III. Search for gravitational wave signals. Astron. Astrophys. 2023 Oct;678:A50. https:

//doi.org/10.1051/0004-6361/202346844. arXiv:2306.16214. [astro-ph.HE].

[17] Reardon DJ, Zic A, Shannon RM, Hobbs GB, Bailes M, Di Marco V, et al. Search for an Isotropic Gravitational-wave Background with the Parkes Pulsar Timing Array. Astrophys. J. Lett. 2023 Jul;951(1):L6. https://doi.org/10.3847/2041-8213/acdd02. arXiv:2306.16215. [astro-ph.HE].

[18] Xu H, Chen S, Guo Y, Jiang J, Wang B, Xu J, et al. Searching for the Nano-Hertz Stochastic Gravitational Wave Background with the Chinese Pulsar Timing Array Data Release I. Research in Astronomy and Astrophysics. 2023 Jul;23(7):075024. https://doi.org/10.1088/1674-4527/acdfa5. arXiv:2306.16216. [astro-ph.HE].

[19] Afzal A, Agazie G, Anumarlapudi A, Archibald AM, Arzoumanian Z, Baker PT, et al. The NANOGrav 15 yr Data Set: Search for Signals from New Physics. Astrophys. J. Lett. 2023 Jul;951(1):L11. https://doi.org/10.3847/2041-8213/acdc91. arXiv:2306.16219. [astro-ph.HE].

[20] EPTA Collaboration, InPTA Collaboration, Antoniadis J, Arumugam P, Arumugam S, Babak S, et al. The second data release from the European Pulsar Timing Array. IV. Implications for massive black holes, dark matter, and the early Universe. Astron. Astrophys. 2024 May;685:A94. https://doi.org/10.1051/0004-6361/202347433. arXiv:2306.16227. [astro-ph.CO].

[21] Giovannini M. Gravitational wave constraints on post-inflationary phases stiffer than radiation. Phys. Rev. D. 1998 Oct;58(8):083504. https://doi.org/10.1103/PhysRevD.58.083504. arXiv:hep-ph/9806329. [hep-ph].

[22] Li B, Shapiro PR, Rindler-Daller T. Bose-Einstein-condensed scalar field dark matter and the gravitational wave background from inflation: New cosmological constraints and its detectability by LIGO. Phys. Rev. D. 2017 Sep;96(6):063505. https://doi.org/10.1103/PhysRevD.96.063505. arXiv:1611.07961. [astro-ph.CO].

[23] Figueroa DG, Tanin EH. Ability of LIGO and LISA to probe the equation of state of the early Universe. J. Cosmol. Astropart. Phys. 2019 Aug;2019(8):011. https://doi.org/10.1088/1475-7516/2019/08/011. arXiv:1905.11960. [astro-ph.CO].

[24] Li B, Shapiro PR. Precision cosmology and the stiff-amplified gravitational-wave background from inflation: NANOGrav, Advanced LIGO-Virgo and the Hubble tension. J. Cosmol. Astropart. Phys. 2021 Oct;2021(10):024. https://doi.org/10.1088/1475-7516/2021/10/024. arXiv:2107.12229. [astro-ph.CO].

[25] Li B, Meyers J, Shapiro PR. Multimodality in the Search for New Physics in Pulsar Timing Data and the Case of Kination-amplified Gravitational-wave Background from Inflation. Astrophys. J. 2025 May;985(1):117. https://doi.org/10.3847/1538-4357/adcc14. arXiv:2503.18937. [astro-ph.CO].

[26] The LIGO Scientific Collaboration, the Virgo Collaboration, the KAGRA Collaboration, Abac AG, Abouelfettouh I, Acernese F, et al. Cosmological and High Energy Physics implications from gravitational-wave background searches in LIGO-Virgo-KAGRA's O1-O4a runs. arXiv e-prints. 2025 Oct;p. arXiv:2510.26848. https://doi.org/10.48550/arXiv.2510.26848. arXiv:2510.26848. [gr-qc].

[27] Riess AG, Yuan W, Macri LM, Scolnic D, Brout D, Casertano S, et al. A Comprehensive Measurement of the Local Value of the Hubble Constant with 1 km s^{-1} Mpc^{-1} Uncertainty from the Hubble Space Telescope and the SH0ES Team. Astrophys. J. Lett. 2022 Jul;934(1):L7. https://doi.org/10.3847/2041-8213/ac5c5b. arXiv:2112.04510. [astro-ph.CO].

[28] Petzold L. Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. SIAM Journal on Scientific and Statistical Computing. 1983;4(1):126–148.

[29] Marx E, Benoit W, Gunny A, Omer R, Chatterjee D, Venterea RC, et al. Machine-learning pipeline for real-time detection of gravitational waves from compact binary coalescences. Phys Rev D. 2025 feb;111(4):042010. Publisher: American Physical Society. https://doi.org/10.1103/PhysRevD.111.042010.

[30] Raikman R, Moreno EA, Govorkova E, Marx EJ, Gunny A, Benoit W, et al.

GWAK: gravitational-wave anomalous knowledge with recurrent autoencoders. Mach Learn: Sci Technol. 2024 apr;5(2):025020. Publisher: IOP Publishing. https://doi.org/10.1088/2632-2153/ad3a31.

[31] Chatterjee D, Marx E, Benoit W, Kumar R, Desai M, Govorkova E, et al. Rapid likelihood free inference of compact binary coalescences using accelerated hardware. Mach Learn: Sci Technol. 2025 Jul;5(4):045030. Publisher: IOP Publishing. https://doi.org/10.1088/2632-2153/ad8982.

[32] Dax M, Green SR, Gair J, Deistler M, Schölkopf B, Macke JH. Group equivariant neural posterior estimation. In: Proceedings of International Conference on Learning Representations; 2021. p. 1. Available from: https://openreview.net/forum?id=u6s8dSporO8.

[33] Dax M. Real-Time Gravitational Wave Science with Neural Posterior Estimation. Phys Rev Lett. 2021 jan;127(24). https://doi.org/10.1103/PhysRevLett.127.241103.

[34] Wildberger J. Adapting to noise distribution shifts in flow-based gravitational-wave inference. Phys Rev D. 2023 jan;107(8). https://doi.org/10.1103/PhysRevD.107.084046.

[35] Dax M, Green SR, Gair J, Gupte N, Pürrer M, Raymond V, et al. Real-time inference for binary neutron star mergers using machine learning. Nature. 2025 mar;639(8053):49–53. Publisher: Nature Publishing Group. https://doi.org/10.1038/s41586-025-08593-z.

[36] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997 nov;9(8):1735 – 1780. Publisher: MIT PressPUB1010Cambridge, MA, USA. https://doi.org/10.1162/neco.1997.9.8.1735.

[37] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998;86(11):2278–2324. https://doi.org/10.1109/5.726791.

[38] Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences. 1982;79(8):2554–2558. Publisher: Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.79.8.2554.

[39] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. NIPS 2014 Workshop on Deep Learning, December 2014. 2014;.

[40] Lea C, Vidal R, Reiter A, Hager GD. Temporal Convolutional Networks: A Unified Approach to Action Segmentation. In: Hua G, Jégou H, editors. Computer Vision – ECCV 2016 Workshops. Springer International Publishing; 2016. p. 47–54.

[41] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North. Association for Computational Linguistics; 2019. p. 4171–4186. Available from: http://aclweb.org/anthology/N19-1423.

[42] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. In: Proceedings of the 2014 International Conference on Learning Representations; 2014. p. 1. Available from: https://doi.org/10.48550/arXiv.1409.0473.

[43] Sutskever I, View Profile, Vinyals O, View Profile, Le QV, View Profile. Sequence to sequence learning with neural networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. Guide Proceedings; 2014. p. 3104–3112. Available from: https://dl.acm.org/doi/10.5555/2969033.2969173.

[44] Lawrence E, Heitmann K, Kwan J, Upadhye A, Bingham D, Habib S, et al. The Mira-Titan Universe. II. Matter Power Spectrum Emulation. Astrophys. J. 2017 Sep;847(1):50. https://doi.org/10.3847/1538-4357/aa86a9. arXiv:1705.03388. [astro-ph.CO].

[45] Zhai Z, Tinker JL, Becker MR, DeRose J, Mao YY, McClintock T, et al. The Aemulus Project. III. Emulation of the Galaxy Correlation Function. Astrophys. J. 2019 Mar;874(1):95. https://doi.org/10.3847/1538-4357/ab0d7b. arXiv:1804.05867. [astro-ph.CO].

[46] Chen Z, Yu Y, Han J, Jing Y. CSST cosmological emulator I: Matter power spectrum emulation with one percent accuracy. Science China Physics, Mechanics & Astronomy. 2025 Jun;68(8):289512. https://doi.org/10.1007/s11433-025-2671-0.

[47] Różański T, Ting YS, Rix HW, Jabłońska M. TransformerPayne: Enhancing Spectral Emulation Accuracy and Data Efficiency by Capturing Long-range Correlations. The Astrophysical Journal. 2025;980:66. Publisher: The American Astronomical Society. https://doi.org/10.3847/1538-4357/ad9b99.

[48] Saraivanov E. Attention-based neural network emulators for multiprobe data vectors. II. Assessing tension metrics. Phys. Rev. D. 2025;111(12). https://doi.org/10.1103/PhysRevD.111.123520.

[49] Boyle LA, Buonanno A. Relating gravitational wave constraints from primordial nucleosynthesis, pulsar timing, laser interferometers, and the CMB: Implications for the early universe. Phys. Rev. D. 2008 Aug;78(4):043531. https://doi.org/10.1103/PhysRevD.78.043531. arXiv:0708.2279. [astro-ph].

[50] Morley SK, Brito TV, Welling DT. Measures of Model Performance Based On the Log Accuracy Ratio. Space Weather. 2018 Jan;16(1):69–88. https://doi.org/10.1002/2017SW001669.

[51] McKay MD, Beckman RJ, Conover WJ. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics. 1979 jan;21(2):239–245. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality]. https://doi.org/10.2307/1268522.

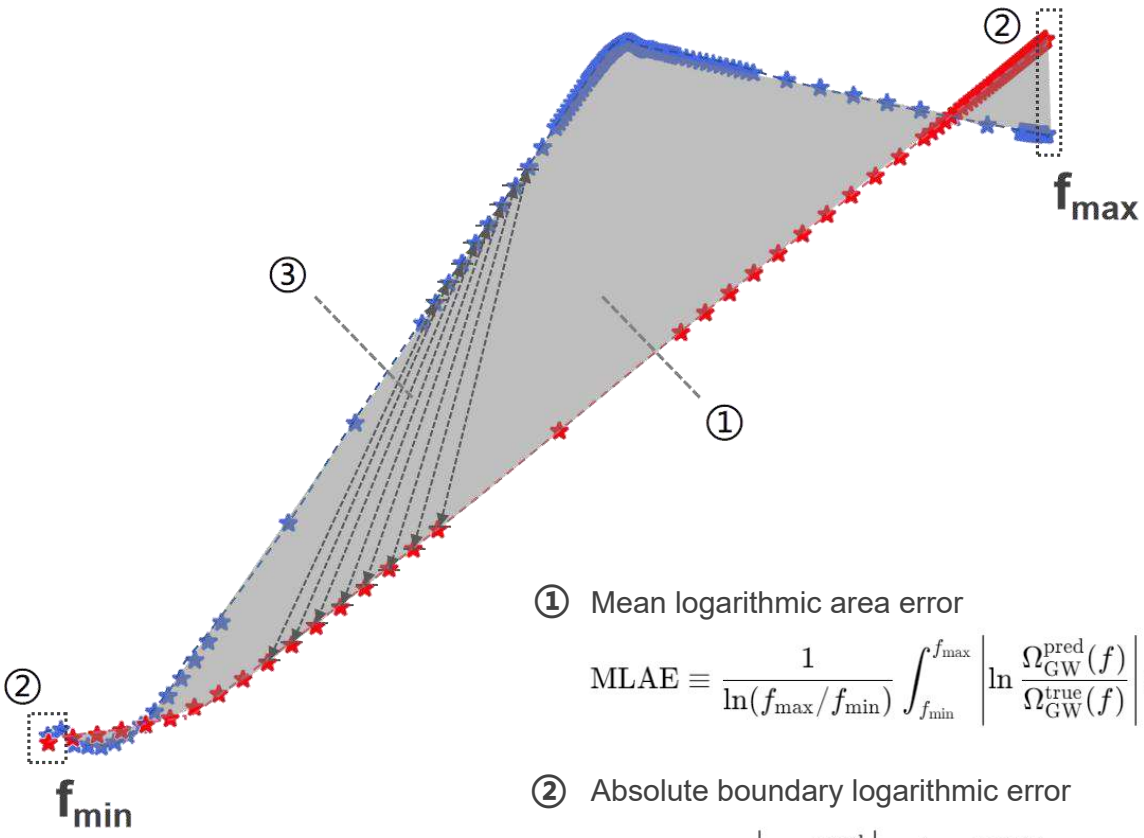
[52] Luo Y, Li B, Zhang F.: SageNet+ Dataset. Zenodo. Available from: https://doi.org/10.5281/zenodo.16723652.

Extended Data Table 1: Prior ranges for the nine-dimensional parameter space in LHS.

Parameter	**Prior Range**	**Definition**
$\log_{10}r$	$[-40, 0]$	Tensor-to-scalar ratio
$n_{\rm t}$	$[-1, 6]$	Tensor spectral index satisfies the consistency relation
$\log_{10}\kappa_{10}$	$[-7, 3]$	Ratio of the stiff-fluid density
$\log_{10}T_{\rm re}$	$[-3, 7]$	Reheating temperature
$\Delta N_{\rm re}$	$[0, 40]$	Number of e-folds at the end of reheating
$\Omega_{\rm b}h^2$	$[0.005, 0.1]$	Baryon density
$\Omega_{\rm c}h^2$	$[0.001, 0.99]$	Dark matter density
H_0	$[20, 100]$	Hubble constant today
$\ln(10^{10}A_{\rm s})$	$[1.61, 3.91]$	The scalar amplitude

Extended Data Table 2: Based on a test set of 5,400 samples, the percentage of samples with area error below specified thresholds for the recurrent architecture including LSTM, GRU and RNN, the convolutional architecture including CNN and TCN. We also tried Transformer, and DNN which is activated by LeakyReLU to capture GW curve features accurately [4].

Threshold	**Transformer**	**LSTM**	**GRU**	**RNN**	**TCN**	**CNN**	**DNN**
$< 0.25\%$	20.28	1.89	12.39	2.06	5.75	8.00	2.93
$< 0.5\%$	88.53	38.18	68.24	40.95	57.91	59.78	48.87
$< 0.75\%$	98.96	78.61	93.79	79.67	89.07	90.73	87.01
$< 1\%$	99.85	93.91	98.74	94.61	97.05	97.26	96.32



① Mean logarithmic area error

$$\mathrm{MLAE} \equiv \frac{1}{\ln(f_{\max}/f_{\min})} \int_{f_{\min}}^{f_{\max}} \left| \ln \frac{\Omega_{\mathrm{GW}}^{\mathrm{pred}}(f)}{\Omega_{\mathrm{GW}}^{\mathrm{true}}(f)} \right| \mathrm{d} \ln f$$

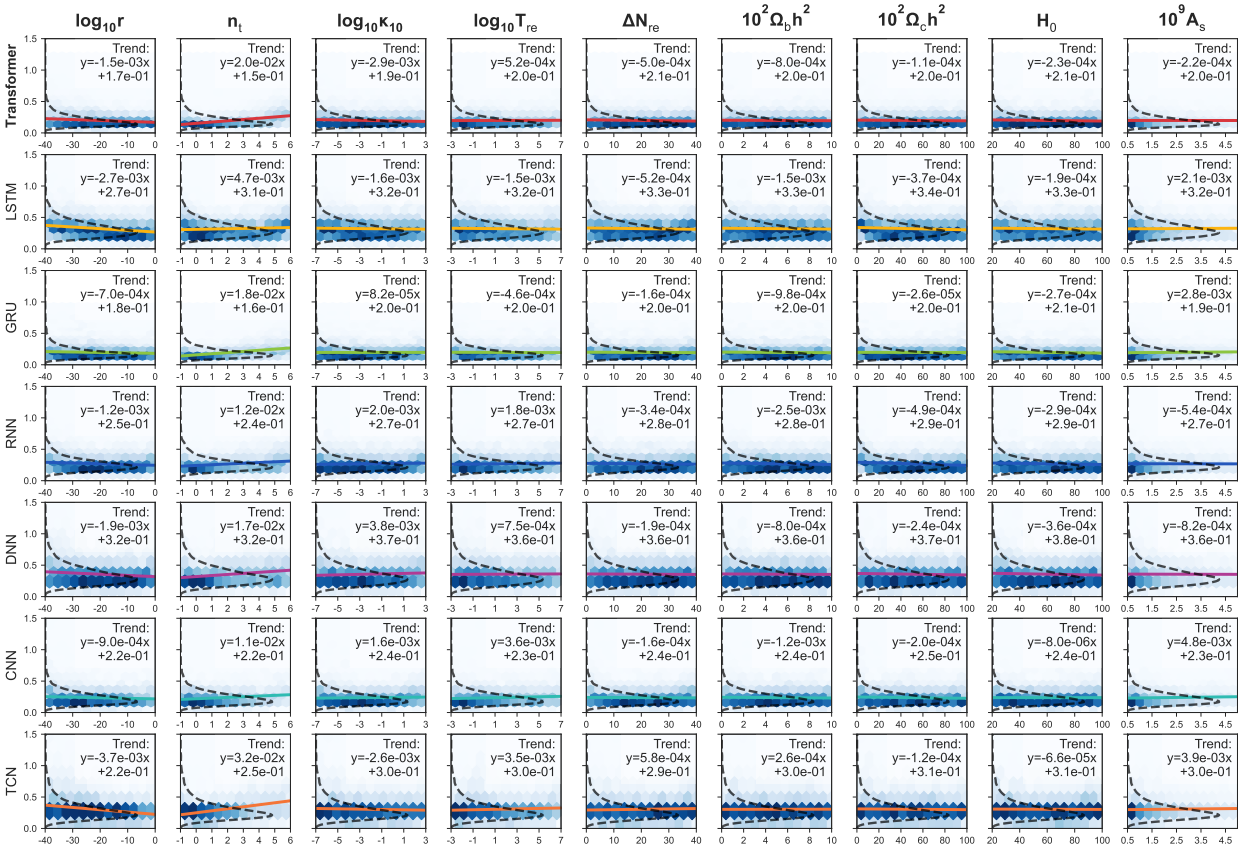
② Absolute boundary logarithmic error

$$\mathrm{ABLE} \equiv \left| \ln \frac{f_{\min}^{\mathrm{pred}}}{f_{\min}^{\mathrm{true}}} \right| + \left| \ln \frac{f_{\max}^{\mathrm{pred}}}{f_{\max}^{\mathrm{true}}} \right|$$

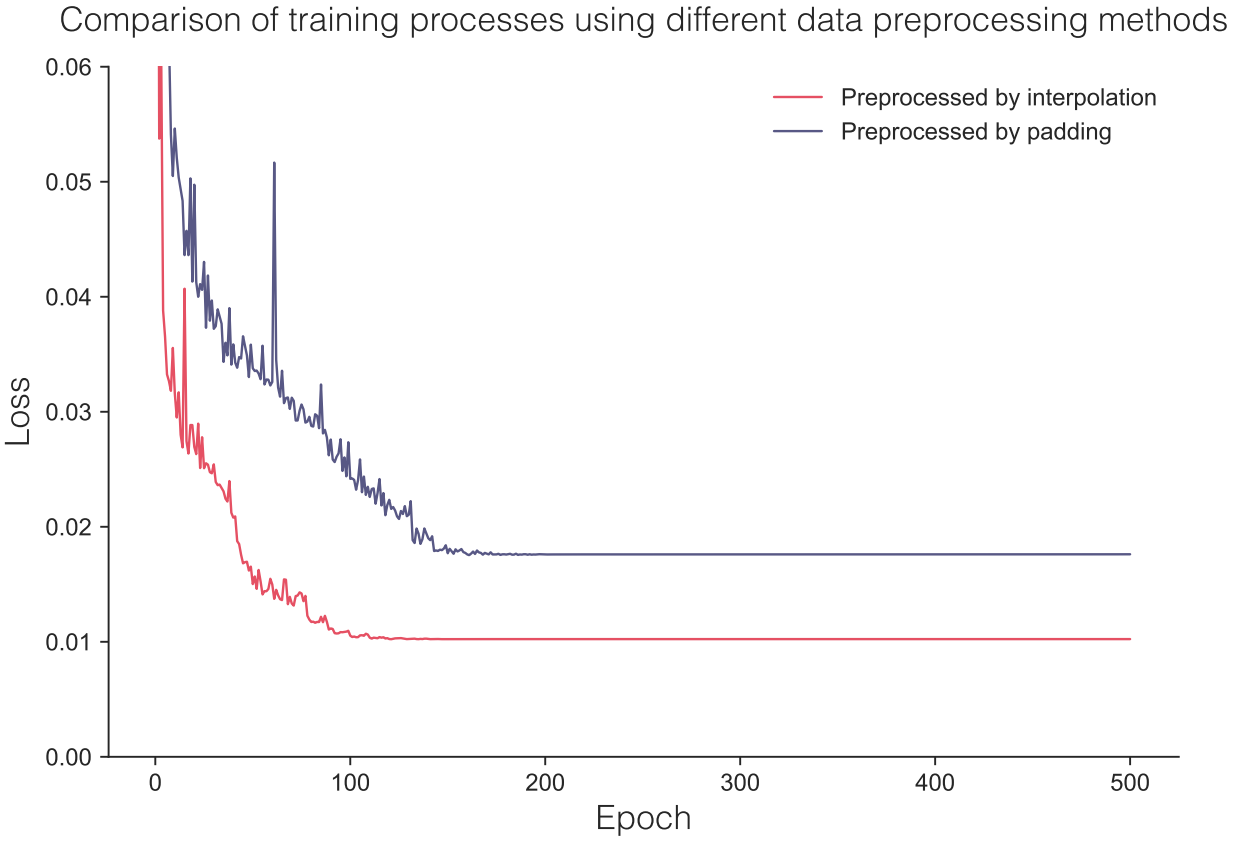
③ Symmetric mean absolute percentage error

$$\mathrm{SMAPE} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\Omega_{\mathrm{GW}}^{\mathrm{true}}(f_i) - \Omega_{\mathrm{GW}}^{\mathrm{pred}}(f_i)|}{|\Omega_{\mathrm{GW}}^{\mathrm{true}}(f_i)| + |\Omega_{\mathrm{GW}}^{\mathrm{pred}}(f_i)|}$$

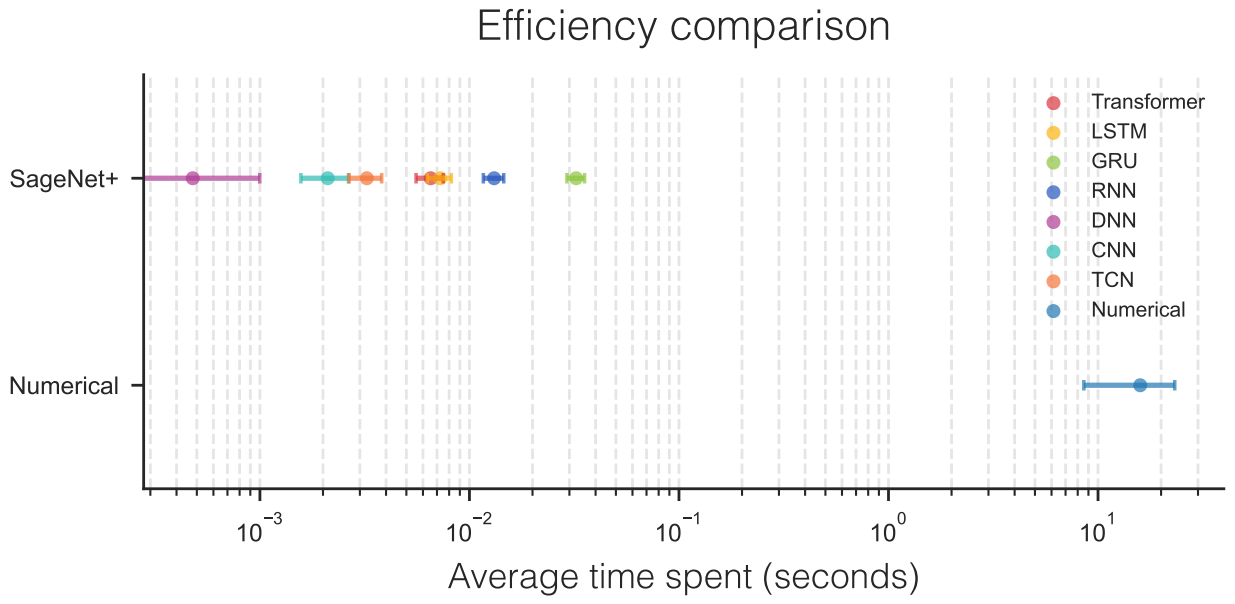
Extended Data Figure 1: **Interpreting the proposed new metrics.** When measuring the differences between predicted and true points on a curve, a common method is to pair points with the same sequence position and calculate their relative positional differences. This approach intuitively reflects the model's prediction accuracy at specific sequence positions and offers good accuracy when measuring errors in ordered data. However, this method assumes a strict correspondence in the sequence order of predicted and true points. If the frequency points f_i in the predicted sequence are not strictly aligned with the original sequence, the error can be significantly amplified. Additionally, considering that the original curve employs an adaptive sampling strategy with varying point densities across different frequency ranges, slight offsets in predicted point positions in sparsely sampled regions can lead to large local errors, even if the overall curve shape closely matches the true value. In practice, since the final predicted curve can be densely interpolated, we are more concerned with the consistency of the overall curve shape rather than differences at specific points. To address this, we adopted the relative area error to measure the prediction error of the curve, directly evaluating the overall physical characteristics of the curve.



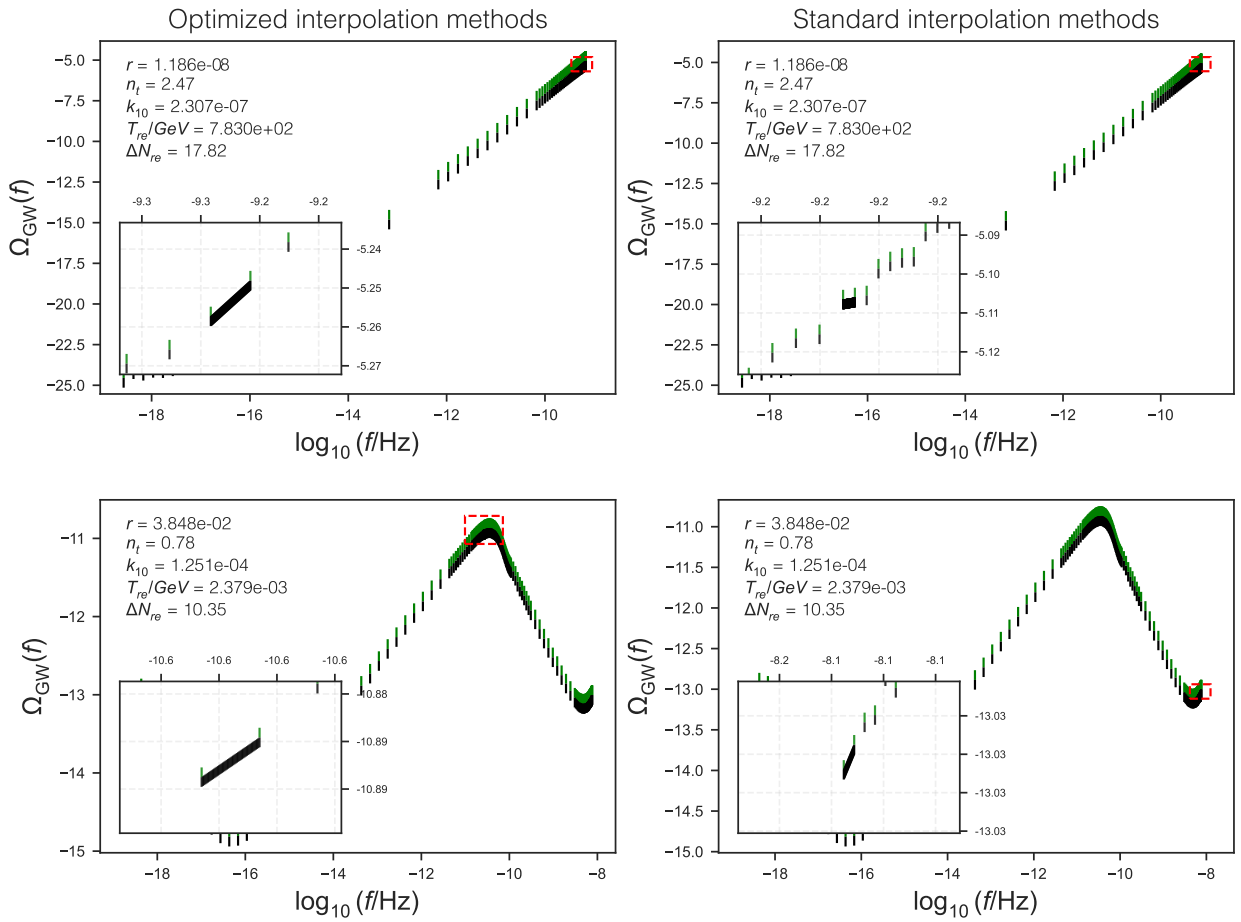
Extended Data Figure 2: **Performance analysis on parameters using the MLAE metric.** Different with Fig. 5, the analysis using the MLAE metric shows a consistently positive slope for $n_{\rm t}$ across all models. This positive trend aligns with the physical interpretation that a steeper inflationary tensor spectrum (higher $n_{\rm t}$) reduces the maximum frequency ($f_{\rm max}$) due to the ultraviolet cutoff, leading to a smaller denominator in the MLAE, $\ln(f_{\rm max}/f_{\rm min})$, and thus an increased prediction error.



Extended Data Figure 3: **Training efficiency with different data preprocessing methods.** The figure shows the loss curves during training Transformer model to handle $(f, \log_{10}\Omega_{\rm GW})$ sequences with variable length. To allow the model to process them, we used interpolation and padding methods, which is conventional in NLP tasks. When using the interpolation method, the model's convergence speed and accuracy are higher than those implemented by padding the dataset during training.



Extended Data Figure 4: **Average time cost of SageNet+ and numerical method.** When tested on a CPU with a Transformer as the feature extractor, the average prediction time is 6ms with a standard deviation of 1ms, while the average prediction time for numerical computation is 15.9s with a standard deviation of 7.3s. Compared to numerical computation, SageNet+ achieves significant efficiency improvements while maintaining accuracy, making large-scale MCMC inference feasible.



Extended Data Figure 5: **Comparision between different interpolation methods.** To prevent the interpolation method from overly focusing on the high-frequency stage of the curve and missing the physical information carried by dense variations in the mid-to-low frequency range, we selected the first 60% of the curve as the interpolable range. Under this choice, if the curve exhibits significant changes only in the high-frequency stage, the algorithm will still perform interpolation in that part; however, if the curve shows significant changes in the more critical mid-to-low frequency range, the interpolation window will immediately shift to the mid-to-low frequencies, thereby allowing the interpolation algorithm to focus maximally on the frequency band with the greatest physical significance.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation.pdf