



21cmVAE: A Very Accurate Emulator of the 21 cm Global Signal

Christian H. Bye^{1,2} , Stephen K. N. Portillo³ , and Anastasia Fialkov^{4,5} ¹Department of Astronomy, University of California, Berkeley, CA 94720, USA; chbye@berkeley.edu²Department of Physics, McGill University, Montréal, QC H3A 2T8, Canada³DIRAC Institute, Department of Astronomy, University of Washington, 3910 15th Avenue NE, Seattle, WA 98195, USA⁴Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge, CB3 0HA, UK⁵Kavli Institute for Cosmology, Madingley Road, Cambridge, CB3 0HA, UK

Received 2021 July 12; revised 2022 March 14; accepted 2022 April 3; published 2022 May 5

Abstract

Considerable observational efforts are being dedicated to measuring the sky-averaged (global) 21 cm signal of neutral hydrogen from Cosmic Dawn and the Epoch of Reionization. Deriving observational constraints on the astrophysics of this era requires modeling tools that can quickly and accurately generate theoretical signals across the wide astrophysical parameter space. For this purpose artificial neural networks were used to create the only two existing global signal emulators, 21CMGEM and GLOBALEMU. In this paper we introduce 21CMVAE, a neural-network-based global signal emulator, trained on the same data set of $\sim 30,000$ global signals as the other two emulators, but with a more direct prediction algorithm that prioritizes accuracy and simplicity. Using neural networks, we compute derivatives of the signals with respect to the astrophysical parameters and establish the most important astrophysical processes that drive the global 21 cm signal at different epochs. 21CMVAE has a relative rms error of only 0.34%—equivalently 0.54 mK—on average, which is a significant improvement compared to the existing emulators, and a run time of 0.04 s per parameter set. The emulator, the code, and the processed data sets are publicly available at <https://github.com/christianhbye/21cmVAE> and through <https://zenodo.org/record/5904939>.

Unified Astronomy Thesaurus concepts: [Early universe \(435\)](#); [Cosmology \(343\)](#); [Astronomy software \(1855\)](#)

1. Introduction

The 21 cm line of neutral hydrogen is one of the most promising probes of Cosmic Dawn and the Epoch of Reionization (EoR). Emitted by neutral hydrogen at redshifts $z > 6$, the signal is redshifted to frequencies below 200 MHz and can therefore be observed by radio telescopes (Tozzi et al. 2000). The global 21 cm signal, obtained by averaging the spectrum across all sky, traces the cosmology and astrophysics of the high-redshift universe. The intensity of the signal is observed in contrast to the background radiation, which is normally assumed to be the cosmic microwave background (CMB), and is quantified by the differential brightness temperature. This temperature depends on the excitation temperature of the 21 cm transition (the spin temperature), the temperature of the background radiation, and the abundance of neutral hydrogen. During Cosmic Dawn and the EoR, processes including the Wouthuysen–Field effect (Wouthuysen 1952; Field 1959), X-ray heating (Madau et al. 1997), and reionization of neutral hydrogen affect the differential brightness temperature and leave characteristic features in the global signal (Shaver et al. 1999). The Wouthuysen–Field effect describes how the spin states of neutral hydrogen are mixed through absorption and reemission of Ly α photons, thus coupling the spin temperature to the kinetic temperature of the hydrogen gas. This effect was dominant in the beginning of the Cosmic Dawn, when the gas temperature was cooler than the background radiation temperature, hence making the differential brightness temperature

negative. Later, the first X-ray sources heated the intergalactic medium (IGM), making the gas go from absorption to emission against the background. Finally, the abundance of neutral hydrogen decreased owing to reionization of the gas, and the signal vanished. We refer to Furlanetto et al. (2006), Pritchard & Loeb (2012), and Barkana (2016) for in-depth reviews of these physical processes.

There are several ongoing efforts to detect the global signal with one reported detection (Bowman et al. 2018), made by the Experiment to Detect the Global EoR Signature (EDGES; Bowman & Rogers 2010; Monsalve et al. 2017, 2018, 2019). Other experiments include Probing Radio Intensity at high- Z from Marion (PRIZM; Philip et al. 2019), Mapper of the IGM Spin Temperature (MIST),⁶ Shaped Antenna measurement of the background RADIO Spectrum (SARAS; Singh et al. 2022), Radio Experiment for the Analysis of Cosmic Hydrogen (REACH; de Lera Acedo 2019; Cumner et al. 2022), and Dark Ages Polarimeter Pathfinder (DAPPER; Burns et al. 2019). The EDGES detection showed a deep and narrow absorption profile, which cannot be explained by standard cosmological models. In particular, the lower 99% confidence bound on the best-fit amplitude was 0.3 K, which is approximately 50% greater than the largest predicted amplitude (Bowman et al. 2018).

With the surprising first results and awaited new measurements, it is important to model the range of possible 21 cm signals from the early universe and explore the associated astrophysical parameter space. To this end, a flexible method is required to realize the global signal efficiently, allowing parameters to be constrained from measurements with the use of sampling techniques such as Markov Chain Monte Carlo



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

⁶ <http://www.physics.mcgill.ca/mist/>

(MCMC; Liu & Shaw 2020). Simulations of the 21 cm signal (e.g., Visbal et al. 2012; Fialkov et al. 2014) take a few hours to run and, thus, cannot be directly employed in a parameter estimation pipeline. Instead, emulators trained on the results of these simulations are becoming a popular tool for fast model generation. Currently available software include EMUPY (Kern et al. 2017), which emulates the 21 cm power spectrum; 21CMGAN (List & Lewis 2020), which generates tomographic samples of the 21 cm brightness temperature; and the two global signal emulators 21CMGEM (Cohen et al. 2020) and GLOBALEMU (Bevins et al. 2021). Given seven astrophysical parameters, 21CMGEM calculates five auxiliary parameters and uses a series of neural networks, a bagged tree classifier, and principal component analysis to emulate the global signal over redshifts $z = 5\text{--}50$. It is fast and accurate: the emulator has a running time of 0.16 s per parameter set and an average relative error of 1.59%. The emulator has already been used to constrain the astrophysical parameters using the data of the EDGES High Band experiment (Monsalve et al. 2019). GLOBALEMU emulates the global signal from the same parameters as 21CMGEM, using just a single neural network. The simpler prediction algorithm emulates the global signals significantly faster, with a run time of only 1.3 ms, and with a smaller average relative error of 1.12%.

In this paper we present 21CMVAE, a new emulator of the global signal from the same seven astrophysical parameters used in 21CMGEM and GLOBALEMU: the star formation efficiency (f_*), the minimum circular velocity of star-forming halos (V_c), the X-ray radiation efficiency (f_X), the optical depth (τ) of the CMB radiation, the power-law slope (α) and low-energy cutoff (ν_{\min}) of the X-ray spectral energy distribution (SED), and the mean free path of ionizing photons (R_{mfp}). The emulator uses the same data set as 21CMGEM and GLOBALEMU, with global signals simulated from the parameters using the method described in, e.g., Cohen et al. (2020). We refer to Cohen et al. (2020) for more details on the modeling.

The objective of this work is to use artificial neural networks to learn relationships between parameters and signals without enforcing any physical models. Compared to 21CMGEM, we aim to use a simpler prediction algorithm that does not calculate auxiliary parameters or depend on the input parameters. However, unlike GLOBALEMU, we do not aim to make the fastest or simplest model that meets a target performance—instead, we explore a wider range of models and prioritize very accurate predictions. As a result, our emulator uses only one neural network that predicts global signals given the seven astrophysical parameters. This allows 21CMVAE to be flexible and predict a wide range of signals. The emulator is written in Python, using the machine-learning libraries TensorFlow (Abadi et al. 2015) and Keras (Chollet et al. 2015), making it easy to modify or retrain. 21CMVAE is available on GitHub,⁷ where the data set and documentation also can be found.

This paper has six sections. After the introduction, Section 2 discusses how we designed and optimized our models, and the performance and speed of the emulator are presented in Section 3. In Section 4 we evaluate the impact of the astrophysical parameters on the global signal, and in Section 5 we use the emulator to interpret an intermediate (latent) representation of the signals. We conclude in Section 6.

2. Methods

This section assumes familiarity with basic principles and terminology of neural networks; we refer to Goodfellow et al. (2016) for a review of this topic.

2.1. Architecture

The emulator has a seven-dimensional input layer (for seven astrophysical parameters) and a 451-dimensional output layer, since it outputs realizations of global signals at 451 redshifts. It also has four hidden layers—of 288, 352, 288, and 224 dimensions, respectively—which all are fully connected and use the activation function ReLU (Fukushima 1969). We also trained an emulator that utilized an autoencoder to create a low-dimensional representation of the global signals; this emulator appeared to be both less accurate and slower, but it is described in Appendix A. We earlier tried to use a variational autoencoder (Kingma & Welling 2013), which uses a regularized latent space that we believed would preserve data structures better than the latent space of a vanilla autoencoder and thus work better in the emulator. However, we found that this regularization came at the cost of increased reconstruction error of the autoencoder and that the emulator error was larger with a variational autoencoder than a standard autoencoder.

We use the rms error as a fraction of signal amplitude to evaluate the performance of the emulator. This is defined for each signal by

$$\text{Error} = \frac{\sqrt{\langle (T(\nu) - \hat{T}(\nu))^2 \rangle}}{\max(|T(\nu)|)}. \quad (1)$$

Here $\hat{T}(\nu)$ denotes the frequency-dependent signal predicted by the emulator, whereas $T(\nu)$ represents the simulated signal from the data set. The figure of merit is the same as used by 21CMGEM, making the results directly comparable.

2.2. Data Set

We use the publicly available data set created for 21CMGEM of 29,641 signals, including a training set with 27,455 signals and a test set with 2186 signals. We restrict the data sets to only include signals in the parameter ranges given by Cohen et al. (2020) (the ranges for the test set in parentheses): $f_* = 0.0001\text{--}0.50$ (0.0003–0.50), $V_c = 4.2\text{--}100$ (4.2–76.5) km s^{-1} , $f_X = 0\text{--}1000$ (0–10), $\tau = 0.04\text{--}0.2$ (0.055–0.1), $\alpha = 1.0\text{--}1.5$ (1.0–1.5), $\nu_{\min} = 0.1\text{--}3.0$ (0.1–3.0) keV, $R_{\text{mfp}} = 10\text{--}50$ (10–50) Mpc. Figure 1 shows a representative sample of the global signal models in the training set.

Approximately 10% of the training set is used as a validation set. This results in a total of 28,996 global signals split into a training set with 24,562 signals, a validation set with 2730 signals, and a test set with 1704 signals. The three data sets were divided randomly, ensuring distributions that are statistically similar, with the only difference being the restrictions on the parameter ranges. The larger parameter ranges on the training and validation set make the emulator capable of exploring a variety of astrophysical scenarios, but we keep narrower ranges on the parameters in the test set to make the results comparable to 21CMGEM and GLOBALEMU.

Before training the emulator, the training signals are preprocessed. This is done by computing the mean temperature at each frequency across all signals and subtracting this from the respective signals. Afterward, every signal is divided by the

⁷ <https://github.com/christianhbye/21cmVAE>

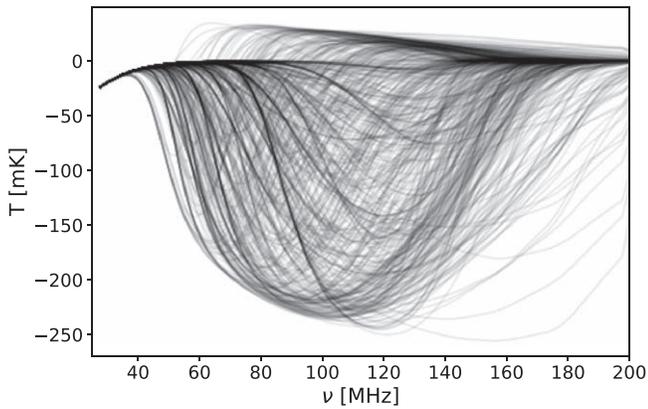


Figure 1. 500 global signal models randomly drawn from the training set.

standard deviation across all signals and frequencies. The preprocessed signals thus have zero mean in every frequency bin and are scaled to units of standard deviation. Standard practice is to divide by the standard deviation in each frequency bin; we deviate from this because the low standard deviation between signals at small frequencies would make the preprocessed signals blow up.

We train the emulator on the signals in the training set, while monitoring the performance on the training set and validation set. Only the errors on the training set are being back-propagated through the neural networks; thus, the errors on the validation set measure whether the model is able to emulate signals that are statistically similar to the training signals but not propagated through the network. The validation error is therefore an important metric for overfitting and the ability of the emulator to generalize to unseen signals. We use it to regulate the learning rate of the model, to trigger early stopping, and during hyperparameter tuning; see Sections 2.3 and 2.4 for more details.

The test set is not seen by the emulator until after training and hyperparameter tuning. The performance on the test set is therefore a measure of how well the emulator can predict global signals from parameters it was not optimized for. Hence, the test error is the best estimate of the performance of the emulator in a real setting and is therefore the figure of merit we use.

2.3. Training

We trained the emulator for 350 epochs, using minibatches with 256 signals in each, and a loss function defined as the square of the figure of merit (Equation (1)). We used the *Adam* (Kingma & Ba 2014) optimizer and an initial learning rate of 0.01. Together, these define the gradient descent algorithm used to minimize the loss function.

During training, we also used a learning rate schedule, which reduces the learning rate if the validation loss does not decrease over 5 epochs, and early stopping, which stops the training if the validation loss does not decrease over 15 epochs. We also saved the model weights and biases after each epoch and loaded the weights and biases that correspond to the epoch with the smallest validation loss after training. This is normally toward the end of the training, but it will be at an earlier epoch in the case of overfitting.

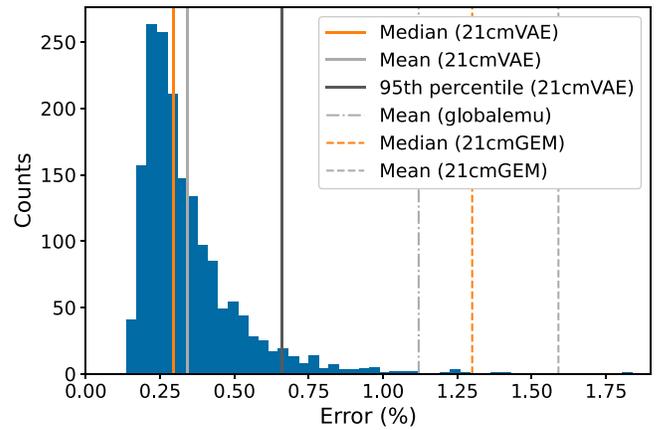


Figure 2. The histogram shows the distribution of errors for the test set of 1704 astrophysical signals with mean, median, and 95th percentile marked (solid lines). It also shows the results for GLOBALEMU (dashed-dotted line) and 21CMGEM (dashed lines). Note that GLOBALEMU does not report the median error and that the 95th percentile errors of GLOBALEMU and 21CMGEM are 2.41% and 3.49%, respectively, hence out of the range used for the x-axis here.

2.4. Hyperparameter Tuning

In our neural networks, there are several parameters that cannot be optimized during training. These are called hyperparameters; examples include the number of layers in the network, dimensionality of each layer, and the loss functions. We performed hyperparameter tuning to optimize the number of layers and dimensionality of each layer. Although there exist dedicated Python packages to do this, we wrote our own script based on a random grid search to get full control over the tuning. We specified an appropriate hyperparameter space, used the script to randomly generate parameters in the space, and built the emulator with these parameters. The model was trained, and the validation loss of the emulator was saved. We trained 500 emulators with different hyperparameters. The hyperparameters associated with the emulator with the lowest validation loss were used to build the final product; these define the architecture, which is described in Section 2.1.

3. Performance

3.1. Test Errors

The training algorithm is stochastic; hence, training the same model twice on the same data set may yield different results. We therefore trained the emulator 20 times after optimizing the architecture with hyperparameter tuning. For each of the 20 trials, we computed the error for each signal in the test set. We get a mean error across signals for each trial; the distribution of these mean errors is showed in Appendix A, in Figure 6. 21CMVAE has a mean error across the 20 trials of $0.354\% \pm 0.001\%$ and a median error of $0.305\% \pm 0.001\%$. The results of the best trial of 21CMVAE are displayed in the histogram in Figure 2, which also shows how the performance compares to 21CMGEM and GLOBALEMU.

We list the relative (as defined in Equation (1)) and absolute (the rms in mK) errors of the best trial across the entire simulated frequency range (approximately 28–237 MHz) and for selected frequency bands in Table 1: 50–100 MHz (corresponding to EDGES Low Band), 60–120 MHz (corresponding to EDGES Mid Band), 90–200 MHz (corresponding to EDGES High Band), and 50–200 MHz. The mean error across

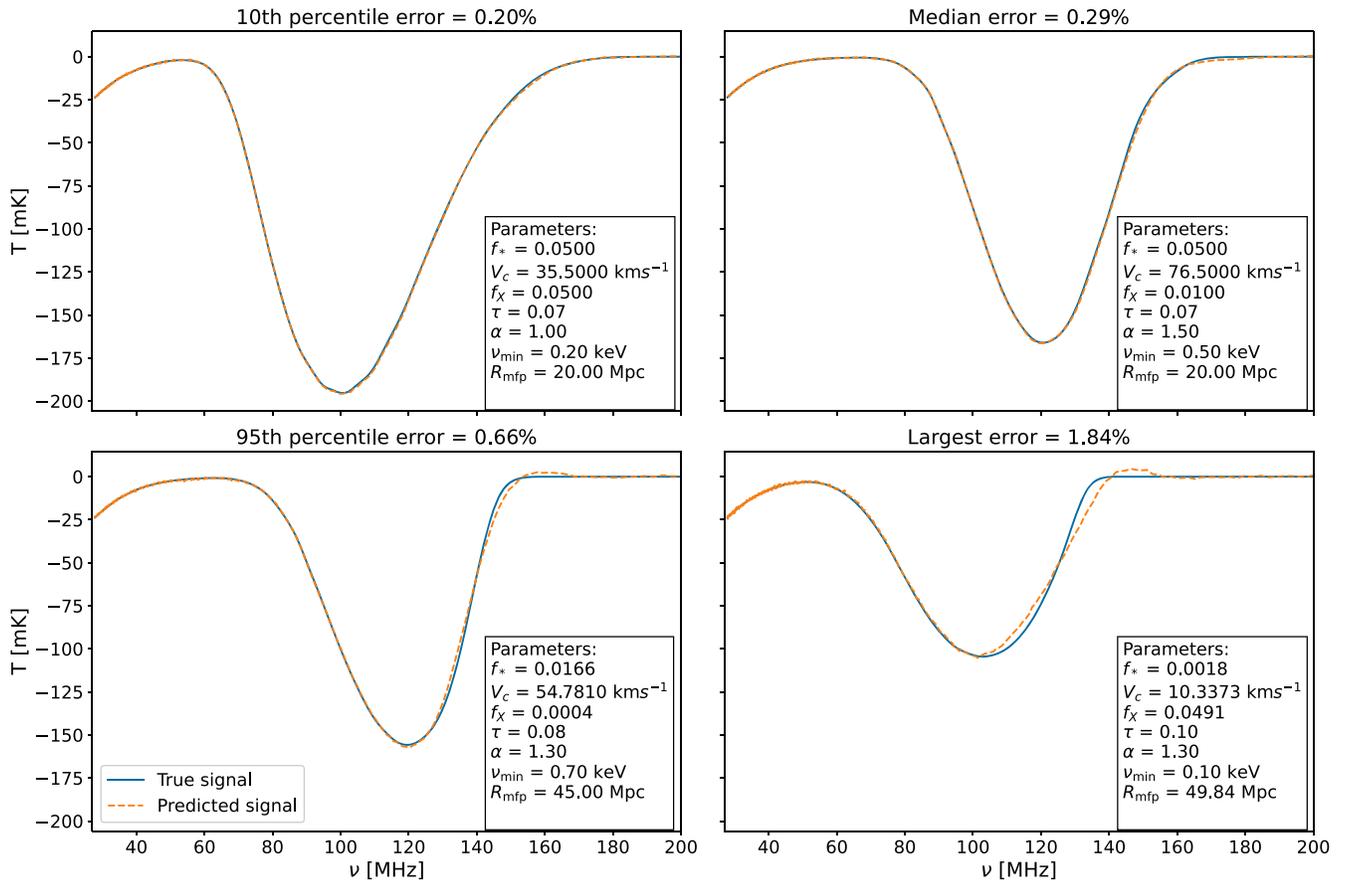


Figure 3. Comparison of true signal (blue solid line) in the test set and emulated signals (orange dashed line) for the case with error closest to the 10th percentile, median, 95th percentile, and the largest error. The errors and the values of the astrophysical parameters of each corresponding model are shown in each panel of the figure.

Table 1

The Mean and Median of the Error of 21CMVAE across the Full Frequency Range the Global Signals Are Sampled at, and across Examples of Frequency Bands Used by the EDGES Experiment

| ν (MHz) | z | Mean Error (%) | Median Error (%) | Mean Error (mK) | Median Error (mK) |
|----------------|-----------|----------------------|------------------------|-----------------------|-------------------------|
| 28–237 | 5.0–50.0 | 0.34 | 0.29 | 0.54 | 0.50 |
| 50–100 | 13.3–27.4 | 0.36 | 0.29 | 0.50 | 0.45 |
| 60–120 | 10.9–22.6 | 0.38 | 0.30 | 0.57 | 0.52 |
| 90–200 | 6.2–14.7 | 0.82 | 0.56 | 0.90 | 0.79 |
| 50–200 | 6.2–27.4 | 0.45 | 0.39 | 0.71 | 0.65 |

the entire frequency range is 0.54 mK, or 0.34% of the signal amplitude. For comparison, the rms between the global signal predicted from all parameters in the middle of the test range and the global signals of a model with all the parameters simultaneously shifted by $\pm 1\%$ is 0.70 mK. Hence, the emulator is sensitive to $\mathcal{O}(1\%)$ changes in the astrophysical input parameters.

For a sample view of the performance of 21CMVAE on the test set, we show in Figure 3 the predicted signals with error closest to the 10th percentile, the median, and the 95th percentile, as well as the signal with the largest error. This can be directly compared to the analogous Figure 9 of Bevins et al. (2021)

and Figure 13 of Cohen et al. (2020), which show the performances of GLOBALEMU and 21CMGEM, respectively.

3.2. Speed

We measure the time from inputting astrophysical parameters to when an output is produced on a computer with a GeForce RTX 2080 Ti Rev. A. GPU. By randomly drawing parameters from the parameter ranges of the training set 1000 times, we find that it takes on average 0.0414 ± 0.0007 s to predict one global signal. We also measured the time it takes to predict 1000 signals simultaneously—that is, inputting 1000 parameter combinations to the emulator at once—obtaining an average time of 0.0418 ± 0.0002 s.

4. Impact of Astrophysical Parameters

We use 21CMVAE to investigate the impact of each astrophysical parameter on the global signal, both qualitatively and quantitatively. First, we show visually how each parameter affects the emulated global signal. We use the signal emulated from the mean parameters in the test set as our nominal model and vary each parameter uniformly between its minimal and maximal value in the test set. The mean parameters are given by $\bar{X} \equiv \{\log f_* = -0.800, \log(V_c/1 \text{ km s}^{-1}) = 1.4803, \log f_X = -0.1190, \tau = 0.07, \alpha = 1.25, \nu_{\min} = 0.85 \text{ keV}, R_{\text{mfp}} = 29.60 \text{ Mpc}\}$.

Figure 4 shows the effect of each parameter on the global signal. In general, larger values of f_* and smaller values of V_c

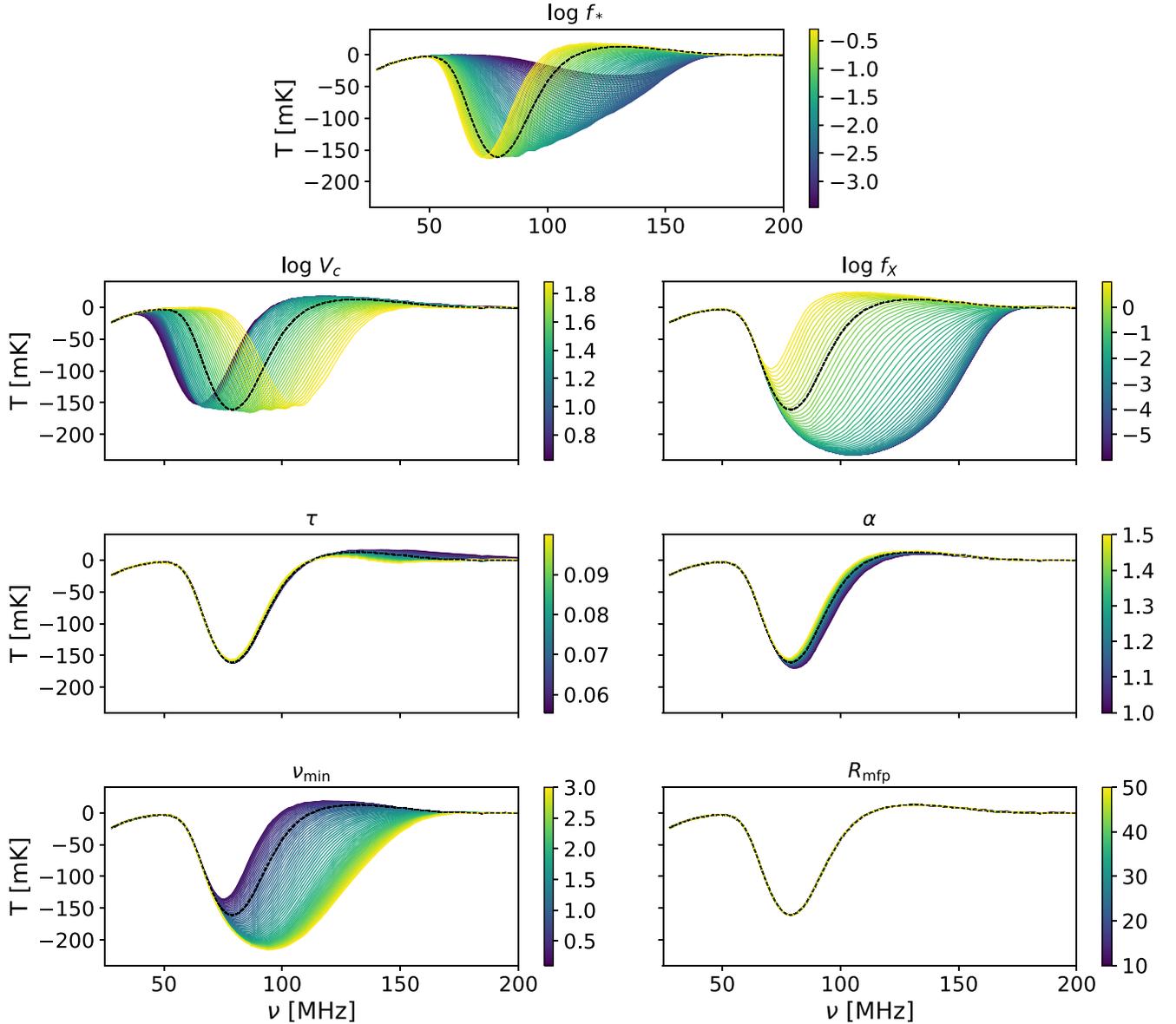


Figure 4. The effect of each astrophysical parameter on the emulated global signal. The black dashed line in each panel is the nominal model. Each panel shows the global signal obtained by varying one parameter, while keeping all the others fixed at their nominal values. The color of each signal represents the value of the varied parameter—on a logarithmic (base-10) scale in the cases of f_* , V_c , and f_X .

are both associated with the center of the absorption trough being shifted to earlier times (lower frequencies); this is as expected since this combination of parameters corresponds to earlier star formation. Furthermore, we note that the amplitude of the global signal is regulated by X-ray heating: it increases with decreasing f_X and increasing ν_{\min} , both of which correspond to less X-ray heating. Less heating also led to later absorption, as can be seen in the figure. Whereas the four parameters discussed have combined a significant impact on the center, amplitude, and width of the absorption trough, the global signal is much less sensitive to the last three parameters. As anticipated, τ is only important at late times and regulates the amplitude during reionization. The variations in amplitude are below the 10 mK level, but the 21 cm line can be seen in emission at low redshifts for sufficiently small values of τ . Being associated with X-ray heating, the global signal varies with α along the same trends as it varies with f_X and ν_{\min} , but the variations are smaller and α is more constrained than the

other parameters. Finally, R_{mfp} has no apparent effect on the global signal at the scales considered here.

Since 21CMVAE maps astrophysical parameters to realizations of the global signal, we can compute the derivative of this map with respect to the input parameters. This allows us to quantify the impact I of each astrophysical parameter on the global signal, which we define as the rms of the derivative of the mean global signal—that is, the global signal corresponding to the mean parameters of the test set (\bar{X})—with respect to its inputs. We define the impact of each parameter \bar{X}_i by

$$I(\bar{X}_i) \equiv \sqrt{\left\langle \left(\frac{d\bar{T}(\nu)}{dX_i} \right)^2 \right\rangle_{\nu}} \Delta X_i. \quad (2)$$

We also compute the impact of the combinations $\log(f_* \cdot f_X)$, which is proportional (up to the minor effect of X-ray SED) to the total amount of X-rays that goes into heating

Table 2
The Impact in mK of Each Parameter on the Mean Global Signal across the Frequency Bands Considered in Table 1

| ν (MHz) | z | $I(\log f_*)$ | $I(\log V_c)$ | $I(\log f_X)$ | $I(\tau)$ | $I(\alpha)$ | $I(\nu_{\min})$ | $I(R_{\text{mfp}})$ | $I(\log(f_* \cdot f_X))$ | $I(\log(f_*/V_c^3))$ |
|-------------|-----------|---------------|---------------|---------------|-----------|-------------|-----------------|---------------------|--------------------------|----------------------|
| 28–237 | 5.0–50.0 | 64.56 | 91.64 | 121.46 | 6.42 | 13.88 | 114.28 | 0.49 | 269.28 | 218.68 |
| 50–100 | 13.3–27.4 | 104.76 | 156.36 | 189.19 | 6.71 | 21.57 | 177.64 | 0.79 | 420.62 | 365.40 |
| 60–120 | 10.9–22.6 | 124.60 | 175.36 | 236.13 | 8.27 | 26.99 | 222.06 | 0.85 | 521.96 | 420.31 |
| 90–200 | 6.2–14.7 | 98.47 | 107.93 | 224.83 | 13.09 | 25.77 | 211.76 | 0.47 | 486.01 | 293.91 |
| 50–200 | 6.2–27.4 | 93.88 | 133.26 | 176.63 | 9.27 | 20.19 | 166.18 | 0.69 | 391.53 | 318.00 |

of the IGM, and $\log(f_*/V_c^3)$, which measures the efficiency of the Wouthuysen–Field effect and is proportional to the total mass of gas in dark matter halos converted into stars (see, e.g., Barkana 2016). The first factor in the above equation, $\sqrt{\left\langle \left(\frac{d\bar{T}(\nu)}{dX_i} \right)^2 \right\rangle_\nu}$, is proportional to the Fisher information for an instrument that has constant error bars for each pixel in the given frequency band. The other factor, ΔX_i , is the uncertainty on the parameter, which we take to be the width of the parameter range in the test set. Table 2 shows the impact of each parameter in the frequency bands that we reported the results for in Table 1. While this analysis considers derivatives at the mean parameters in our test set, it could be repeated at any point of interest in the parameter space.

The results in Table 2 indicate that the global signal at redshifts $z=5-50$ is most sensitive to changes in the combinations $f_* \cdot f_X$ and f_*/V_c^3 . The four parameters with the greatest impact in this redshift range are f_X , ν_{\min} , V_c , and f_* , and the three parameters with the smallest impact are R_{mfp} , τ , and α . The four parameters with the largest quantitative impact are the same as the parameters that visually appeared to have the most impact on the global signals. Thus, the quantitative impact analysis confirms the intuition provided in Figure 4: the global signal is most sensitive to the strength and time of onset of star formation and heating in the redshift range $z=5-50$.

Table 2 also shows how the impact of the parameters changes with the observed frequency band, and hence how it changes with time. All the parameters except for τ have their largest impact in the redshift range $z=10.9-22.6$, which is expected since the global signal used in this analysis both has its absorption through and goes into emission in this redshift range. The most important parameters in this range are $\log(f_* \cdot f_X)$ and $\log(f_*/V_c^3)$ with impact values of 521.96 and 420.31 mK, respectively. These also have the greatest impact in the range $z=6.2-14.7$, but their impact decreases to about 90% and 70%, respectively, of their values in the range $z=10.9-22.6$. The only parameter that increases in impact is τ , which goes from 8.27 to 13.09 mK. This shows that τ becomes comparatively more important at lower redshifts. On the other hand, the other reionization parameter, R_{mfp} , has consistently the smallest impact and does not appear to become more important at later times.

5. Interpreting the Latent Space

Having seen how the global signals vary with the astrophysical parameters, we analyze how the last hidden layer changes with the parameters. We denote this representation the latent representation or the latent space. It is 224-dimensional, since the last hidden layer has 224 neurons (in comparison, the

frequency representation is 451-dimensional since the global signals are sampled at 451 frequencies). The latent representation is related to the output global signals by only one linear transformation and the additive bias of the neural network. Investigating the structure of this representation can therefore illustrate which of the seven astrophysical parameters have the strongest effect on the global signals in the frequency range $\nu=28-237$ MHz.

As an illustration, we plot (Figure 5) the value of each of the 224 neurons in the last hidden layer for each of the 1704 signals from the test set projected to a plane using the t-SNE technique (t-Stochastic Neighbor Embedding; van der Maaten & Hinton 2008). This technique preserves relative distances, meaning that points close in the latent space are also close in the 2D projection. We require that the 200 closest points, or about 10%, are counted as neighbors (i.e., setting the perplexity parameter for t-SNE to 200). For each of the seven astrophysical parameters, we color a copy of this plot according to the value of the parameters and show the results in Figure 5. To gain further insight, we also show the combinations $f_* \cdot f_X$ and f_*/V_c^3 discussed in Section 4.

Since Figure 5 is an attempt at visualizing a 224-dimensional space in 2D, the shapes and structures are not very meaningful. However, as t-SNE preserves relative distances, the relative placement of the points contains information. By visually exploring Figure 5, we find, as anticipated, that the structure of the latent space is most sensitive to changes in the astrophysical parameters regulating star formation (namely, f_* and V_c shown in panels (a) and (b), respectively) and X-ray heating (via the dependence on f_X and, to a lesser degree, on ν_{\min} ; panels (c) and (f), respectively), as these parameters have the sharpest effect on the global signal within the explored frequency range. We can also see how the latent space is structured: the upper right part of the space is a region with large f_* (panel (a)) and small V_c (panel (b)), indicating that this part of the latent space is associated with early star formation. This behavior is even more clearly seen in panel (i), which is color-coded with respect to f_*/V_c^3 . Similarly, the latent space separates f_X (panel (c)), with large values to the upper left and smaller values to the lower right, and ν_{\min} (panel (f)), with smaller values to the upper left and larger values to the lower right, effectively corresponding to stronger and earlier heating, respectively (Fialkov et al. 2014), and thus agreeing with the distribution of f_X . In other words, the amount of heating generally increases upward and to the left in the t-SNE representation of the latent space, which is highlighted in panel (h) corresponding to $f_* \cdot f_X$.

We also find that the effects of the parameters regulating reionization (in our case it is the CMB optical depth and the mean free path of the ionizing photons; panels (d) and (g), respectively), as well as the slope of X-ray SED (panel (e)), are

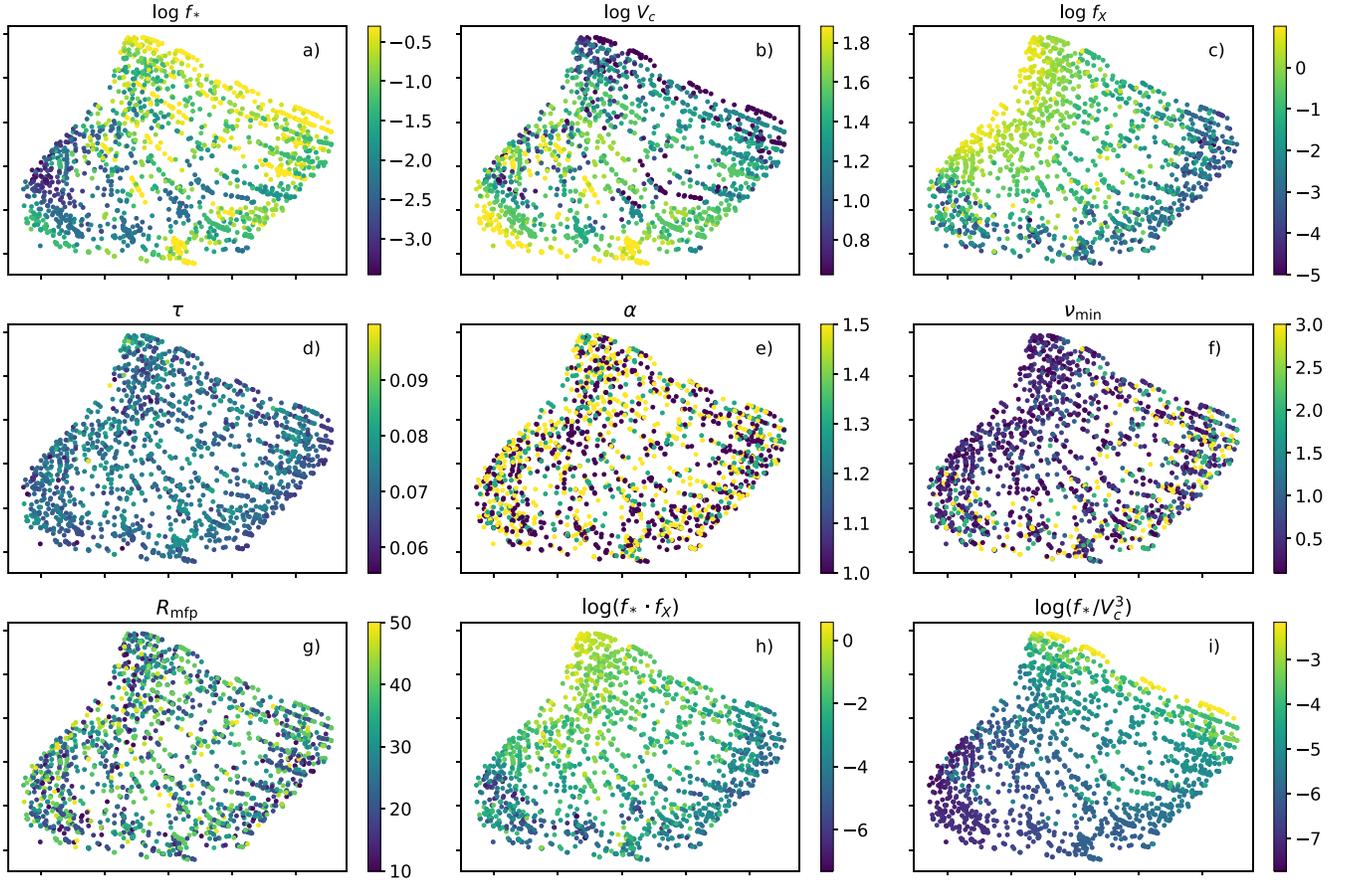


Figure 5. A t-SNE visualization of the latent representation colored by the seven astrophysical input parameters (panels (a)–(g)) and the combinations $f_* \cdot f_X$ (panel (h)) and f_*/V_c^3 (panel (i)). The points shown are the values of the neurons in the last hidden layer of the emulator associated with each global signal. The first three parameters and the two combinations are given on log scales (base-10); otherwise, the units are the same as used elsewhere in the text. In general, the parameters that are most important for the global signal in the frequency band $\nu = 28\text{--}237$ MHz are most separated in the t-SNE representation of the latent space.

not reflected in the t-SNE plots and there is no apparent structure across the corresponding panels. In order to minimize the loss of the emulator, it was not necessary to organize the hidden layer in a way that sorts these parameters in the t-SNE projection, indicating that they have smaller effects on the global signals when explored across the broad frequency range $\nu \approx 28\text{--}237$ MHz.

In summary, we find—as expected—that processes of X-ray heating and Wouthuysen–Field coupling have the strongest impact on the global signal in the explored frequency range, with the latent space separating the amount of star formation along one diagonal in the t-SNE representation and the X-ray heating along the other diagonal. In contrast, the parameters regulating reionization and the slope of X-ray SED have a subdominant effect on the apparent structure of the latent space within the broad frequency range considered.

This is largely consistent with our findings in Section 4, where we saw visually that the center, amplitude, and width of the global signals are most sensitive to the amount of X-ray heating and star formation and quantitatively that the parameters regulating these effects have the greatest impacts. Larger amplitudes are associated with less X-ray heating and, to a lesser degree, more star formation. Since these are largely separated along different diagonals, the t-SNE representation does not separate amplitudes to the same extent as it separates the amount of X-rays and star formation. However, we may loosely identify the lower right region of the latent space with

global signals of large amplitude, since it is a region with low f_X and, in general, above-average values of f_* and ν_{\min} . The opposite side of the plot is a region of smaller amplitudes, with large f_X , more small values of f_* , and small ν_{\min} . On the other hand, the center and width of the global signals appear to be separated with increasing width and the center shifted to high frequencies in the bottom half of the latent space (following generally the separation of $f_* \cdot f_X$).

The parameters τ , α , and R_{mfp} both have the smallest effect on the global signal and do not separate the latent space in the t-SNE representation. This shows that the emulator, as anticipated, does not prioritize separating parameters that do not significantly change the shape of the global signal.

6. Conclusions

We have presented 21CMVAE, a new emulator of the 21 cm global signal, which predicts signals based on a seven-parameter input. The prediction pipeline is simple, with one neural network, whose architecture is optimized with hyper-parameter tuning. The code used to optimize the network is publicly available, together with the final product. 21CMVAE emulates global signals in approximately 0.04 s on average and with a mean rms error of 0.34% of the signal amplitude, a significant improvement in error from existing emulators. The absolute error is on average smaller than the rms between two signals with parameters changed by 1%.

Compared to GLOBALEMU, which is the fastest and most accurate existing emulator of the global 21 cm signal, as well as the most similar in architecture, 21CMVAE has overall smaller errors in prediction of the same global signals: both the mean and maximum errors of 21CMVAE are more than a factor of 3 smaller than the mean and maximum errors of GLOBALEMU over the redshift range $z=5-50$. This is due to different priorities: GLOBALEMU aims to be as simple and fast as possible given a target accuracy of about 10% of the expected noise of the REACH experiment, whereas 21CMVAE aims to be as accurate as possible without compromises. GLOBALEMU is indeed able to emulate the global signals in 1.3 ms, which is even faster than 21CMVAE.

By exploring different representation of the global signal and its derivatives, we were able to qualitatively and quantitatively establish which of the model parameters create the most significant changes in the global 21 cm signal across a broad frequency range. As anticipated, we find that processes of X-ray heating and Wouthuysen–Field coupling have the strongest impact, while the parameters regulating reionization and the slope of X-ray SED have no apparent effect on the latent representation for the redshift range $z=5-50$. As the analysis in Section 4 shows, the impact of each parameter depends on the frequency band considered, and the latent representation will therefore depend on the frequency sampling of the global signals in the training set. The visual latent representation and the quantitative derivative analysis are a potentially powerful diagnostic that can point out dominant astrophysical processes and help optimize theoretical modeling when targeting specific frequency bands of different experiments.

In summary, 21CMVAE achieves unprecedentedly small errors for a range of 21 cm models across a wide frequency band. Combined with the short running time and the implementation in Python, this makes 21CMVAE ideal for parameter fitting such as MCMC.

S.K.N.P. acknowledges support from the DIRAC Institute in the Department of Astronomy at the University of Washington. The DIRAC Institute is supported through generous gifts from the Charles and Lisa Simonyi Fund for Arts and Sciences and the Washington Research Foundation. A.F. was supported by the Royal Society University Research Fellowship.

Appendix A Autoencoder-based Emulator

In addition to 21CMVAE, we trained an autoencoder to generate a low-dimensional—latent—representation of the global signals and an emulator based on the autoencoder. These two networks were trained the same way as 21CMVAE, and the hyperparameters were optimized with the same hyperparameter tuner. The emulator maps astrophysical parameters to the latent representation, which we set to nine dimensions after hyperparameter tuning. The latent representation is then decoded by the autoencoder. By compressing the data to the latent space, the autoencoder is forced to learn the most robust features of the global signal. We believed that this would aid the emulator in predicting the global signals and that the nine latent parameters would be easier to interpret as combinations of the input astrophysical parameters than the hidden layers of 21CMVAE, which all have more than 200 dimensions. Despite this, 21CMVAE actually performs better than the autoencoder-based emulator and can be interpreted both qualitatively and quantitatively, as done in Sections 4 and 5. We still show the methods and results here for comparison.

The autoencoder attempts to reconstruct global signals but has a nine-dimensional latent layer that forces it to reduce the dimensionality of each global signal from 451 to 9. The encoder has one hidden layer with 352 dimensions, whereas the decoder has two hidden layers with 32 and 352 dimensions, respectively. The emulator takes in the seven astrophysical parameters and outputs the latent representation, using four hidden layers of dimensions 352, 352, 352, and 224.

As with 21CMVAE, we trained the tuned autoencoder-based emulator 20 times and computed the mean test error for each trial. These are shown in Figure 6.

21CMVAE has a mean error across trials of $0.354\% \pm 0.001\%$ and a median error of $0.305\% \pm 0.001\%$, whereas the autoencoder-based emulator has a mean error of $0.44\% \pm 0.01\%$ and a median error of $0.39\% \pm 0.01\%$. We see that the distribution of mean errors of 21CMVAE is strictly at smaller errors than the distribution of errors of the autoencoder-based emulator. However, the errors due to the autoencoder itself are in general smaller than the errors of 21CMVAE, as shown in Figure 6. The autoencoder has a mean error of $0.332\% \pm 0.003\%$ and a median error of $0.292\% \pm 0.002\%$. The errors of the autoencoder-based

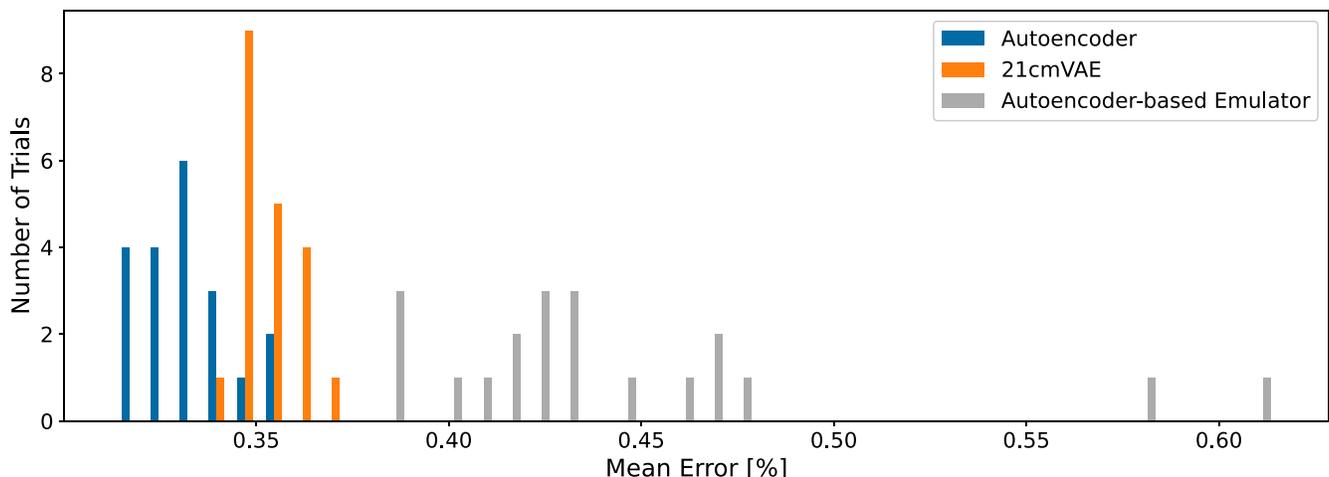


Figure 6. Histogram of the mean errors of the 20 trials for the autoencoder (blue), 21CMVAE (orange), and the autoencoder-based emulator (gray).

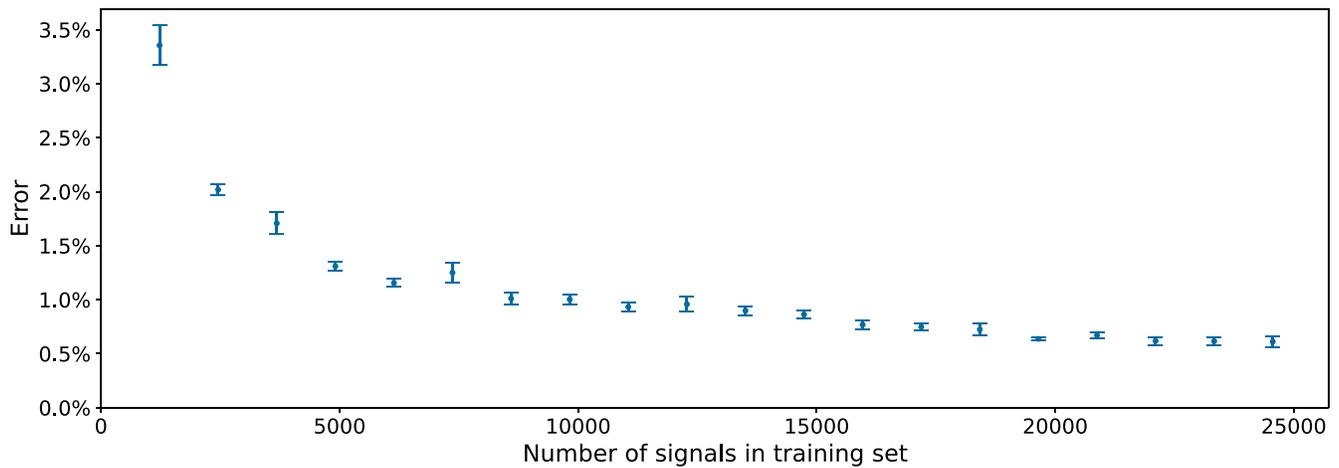


Figure 7. Test error vs. the size of the training set as a percentage of the full training set for the emulator. The points and error bars represent the mean and standard error in the mean, respectively.

emulator would be identical to the autoencoder errors if it could perfectly map the astrophysical parameters to the autoencoder latent parameters; thus, the difference between these distributions is due to errors in the map from astrophysical parameters to latent parameters. An improved autoencoder-based emulator could reduce this difference, but the autoencoder errors are likely the limit of how small the autoencoder-based emulator errors can be. For the best trial, the mean error of the autoencoder-based emulator on the test set is 0.39%, and the median error is 0.35%. This is worse than 21CMVAE (mean: 0.34%; median: 0.29%) but a significant improvement over 21CMGEM (mean: 1.59%; median: 1.30%). The autoencoder that is used by that emulator has a mean error of 0.33% and a median error of 0.29%.

Appendix B Data Set Size

The performance of the emulator is correlated with the size of the training data set. To test this, we randomly sample subsets of the training set and evaluate the performance of the emulator trained on each of the subsets. The subsets range in size from 5% to the full data set in increments of 5%. We sample the subsets at each given size 10 times and average the error across the samples. This ensures that the results depend only on the size of the subsets and not on the composition of the subsets. This relationship is displayed in Figure 7. We see that the emulator error steadily decreases with increasingly more signals until the data set size is $\sim 35\%$ (about 8600 signals), after which it only slowly decreases. Therefore, the marginal effect of increasing the size of the training set used to train 21CMVAE would likely be small.

ORCID iDs

Christian H. Bye  <https://orcid.org/0000-0002-7971-3390>
 Stephen K. N. Portillo  <https://orcid.org/0000-0001-8132-8056>
 Anastasia Fialkov  <https://orcid.org/0000-0002-1369-633X>

References

- Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, <https://www.tensorflow.org/>
- Barkana, R. 2016, *PhR*, **645**, 1
- Bevins, H. T. J., Handley, W. J., Fialkov, A., de Lera Acedo, E., & Javid, K. 2021, *MNRAS*, **508**, 2923
- Bowman, J. D., & Rogers, A. E. E. 2010, *Natur*, **468**, 796
- Bowman, J. D., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J., & Mahesh, N. 2018, *Natur*, **555**, 67
- Burns, J. O., Bale, S., & Bradley, R. F. 2019, AAS Meeting, **234**, 212.02
- Chollet, F., et al. 2015, Keras, <https://keras.io/>
- Cohen, A., Fialkov, A., Barkana, R., & Monsalve, R. A. 2020, *MNRAS*, **495**, 4845
- Cumner, J., De Lera Acedo, E., de Villiers, D. I. L., et al. 2022, *JAI*, **11**, 2250001
- de Lera Acedo, E. 2019, in IEEE Electromagnetics in Advanced Applications (ICEAA) (Piscataway, NJ: IEEE), 0626
- Fialkov, A., Barkana, R., & Visbal, E. 2014, *Natur*, **506**, 197
- Field, G. B. 1959, *ApJ*, **129**, 536
- Fukushima, K. 1969, *IEEE Trans. Syst. Sci. Cybern.*, **5**, 322
- Furlanetto, S. R., Oh, S. P., & Briggs, F. H. 2006, *PhR*, **433**, 181
- Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (Cambridge, MA: MIT Press), <https://www.deeplearningbook.org/>
- Kern, N. S., Liu, A., Parsons, A. R., Mesinger, A., & Greig, B. 2017, *ApJ*, **848**, 23
- Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980
- Kingma, D. P., & Welling, M. 2013, arXiv:1312.6114
- List, F., & Lewis, G. F. 2020, *MNRAS*, **493**, 5913
- Liu, A., & Shaw, J. R. 2020, *PASP*, **132**, 062001
- Madau, P., Meiksin, A., & Rees, M. J. 1997, *ApJ*, **475**, 429
- Monsalve, R. A., Fialkov, A., Bowman, J. D., et al. 2019, *ApJ*, **875**, 67
- Monsalve, R. A., Greig, B., Bowman, J. D., et al. 2018, *ApJ*, **863**, 11
- Monsalve, R. A., Rogers, A. E. E., Bowman, J. D., & Mozdzen, T. J. 2017, *ApJ*, **847**, 64
- Philip, L., Abdurashidova, Z., Chiang, H. C., et al. 2019, *JAI*, **8**, 1950004
- Pritchard, J. R., & Loeb, A. 2012, *RPPh*, **75**, 086901
- Shaver, P. A., Windhorst, R. A., Madau, P., & de Bruyn, A. G. 1999, *A&A*, **345**, 380
- Singh, S., Nambissan, T. J., Subrahmanyam, R., et al. 2022, *NatAs*, in press
- Tozzi, P., Madau, P., Meiksin, A., & Rees, M. J. 2000, *ApJ*, **528**, 597
- van der Maaten, L., & Hinton, G. 2008, *JMLR*, **9**, 2579, <https://www.jmlr.org/papers/v9/vandermaaten08a.html>
- Visbal, E., Barkana, R., Fialkov, A., Tselikhovich, D., & Hirata, C. M. 2012, *Natur*, **487**, 70
- Wouthuysen, S. A. 1952, *AJ*, **57**, 31