Prepared for submission to JCAP

deep21: a Deep Learning Method for 21cm Foreground Removal

T. Lucas Makinen, [a,b,1] **Lachlan Lancaster,** [a] **Francisco Villaescusa-Navarro,** [a] **Peter Melchior,** [a,c] **Shirley Ho,** [d] **Laurence Perreault-Levasseur,** [d,e,f] **and David N. Spergel** [d,a]

[a] Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton, NJ, 08544, USA
[b] Institut d'Astrophysique de Paris, Sorbonne Université, 98 bis Boulevard Arago, 75014 Paris, France
[c] Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08544, USA
[d] Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY, 10010, USA
[e] Department of Physics, Univesité de Montréal, CP 6128 Succ. Centre-ville, Montréal, H3C 3J7, Canada
[f] Mila - Quebec Artificial Intelligence Institute, Montréal, Canada

E-mail: tmakinen@princeton.edu

Abstract. We seek to remove foreground contaminants from 21cm intensity mapping observations. We demonstrate that a deep convolutional neural network (CNN) with a UNet architecture and three-dimensional convolutions, trained on simulated observations, can effectively separate frequency and spatial patterns of the cosmic neutral hydrogen (HI) signal from foregrounds in the presence of noise. Cleaned maps recover cosmological clustering statistics within 10% at all relevant angular scales and frequencies. This amounts to a reduction in prediction variance of over an order of magnitude on small angular scales ($\ell > 300$), and improved accuracy for small radial scales ($k_{\parallel} > 0.17$ h Mpc^{-1}) compared to standard Principal Component Analysis (PCA) methods. We estimate posterior confidence intervals for the network's prediction by training an ensemble of UNets. Our approach demonstrates the feasibility of analyzing 21cm intensity maps, as opposed to derived summary statistics, for upcoming radio experiments, as long as the simulated foreground model is sufficiently realistic. We provide the code used for this analysis on GitHub, as well as a browser-based tutorial for the experiment and UNet model via the accompanying Colab notebook.

Keywords: cosmology: radio – reionisation, large-scale structure – foregrounds – deep learning, signal processing

[1] Corresponding author.

Contents

1 Introduction

Observations of cosmic neutral hydrogen emission hold the promise of making precision measurements of the universe's evolution at intermediate to late redshifts ($0.5 > z > 30$) [34, 38, 63, 68], providing an observable to trace both the growth of massive structure from the time of the Cosmic Microwave Background (CMB), as well as constrain the physics of Reionization ($10 > z > 6$) [20, 48, 53].

Measurement of the 21cm line relies on intensity mapping, in which large fractions of the sky are observed to capture wide-field statistics, instead of resolving individual sources [for

an overview see e.g. 34, 48]. This technique has already been used to place constraints on the formation of the first stars and galaxies at $z \sim 9$ [45], but the precise nature of the Epoch of Reionization (EoR) remains unknown [20]. 21cm intensity maps also promise to be a tracer of three-dimensional, large-scale structure growth at later redshifts $z \lesssim 4$, linking late-stage structure to underlying gravitational theory and the primordial density [9, 26]. Upcoming experiments, such as the Square Kilometer Array (SKA) promise to trace EoR physics to large-scale structure formation using this single observable [4].

The greatest challenge for these measurements is mitigating systematics and removing enormous foreground contamination from both galactic and extragalactic radio sources, such as synchrotron and free-free radio emission [48, 53]. These contaminants tend to be three to four orders of magnitude brighter than the interesting cosmological signal [27, 48]. Furthermore, foregrounds lack detailed analytic descriptions, making 21cm likelihoods hard to specify [3].

The foreground signals have different statistical properties than the cosmological signal, a phenomenon thoroughly covered in the literature [6, 7, 22, 29, 43, 47, 49, 51, 56, 75], with several proposed methods for signal separation [35, 37, 42, 60, 61, 79, 81]. Most foreground contaminants are forecast to be spectrally smooth in frequency, motivating the application of blind signal separation techniques, such as Principal Component Analysis (PCA) [e.g. 2, 16], which require no prior knowledge of the expected signals. However, blind separation techniques are not linked to physical processes and therefore make no use of our physical understanding of these foregrounds. This means that there exists information in the observed signal that is not fully exploited for separation. Blind subtraction can also lead to biased summary statistic recovery. Furthermore, the mean of the HI intensity spectrum is irretrievably removed in blind subtraction in single-dish experiments [14], placing the focus of analyses on interpreting compressed, sometimes biased [see e.g. 2, 64, 79], summary statistics derived from these maps, such as power spectra. With clean 21cm intensity maps, more fundamental large-scale structure analyses and parameter extraction would be possible using the maps themselves [21, 40, 76, for a review].

We address these problems by constructing a convolutional neural network to recover cosmological 21cm maps from PCA-reduced inputs. Deep learning architectures are ideally suited to similar high-dimensional problems such as image segmentation, classification, and computer vision tasks (see e.g. [24] for a review, and e.g. [28], [30], [73] for applications), in which patterns and higher-order correlations must be captured over a large number of input data. To incorporate an estimate of uncertainties of the separated maps we train an ensemble of networks on a suite of simulated radio skies. We then test these architectures on simulations with altered foreground parameters to assess how well the approach generalizes beyond the fiducial choice of model.

Recent studies have incorporated deep learning techniques to analyze EoR cosmology [11, 21, 31, 39, 41, 72], but have largely focused on retrieving compressed cosmological statistics or higher-order correlations from clean intensity maps. Yao [80] presents a foreground removal method for CMB maps based on Generative Adversarial Networks (GANs). Li et al. [33] presented an autoencoder-based method for 21cm foreground removal in the context of interferometric intensity mapping, demonstrating a marked improvement in foreground removal for EoR frequencies. Villanueva-Domingo and Villaescusa-Navarro [72] used a UNet architecture to remove far-field astrophysics from 21cm maps.

However, our study is (to our knowledge) the first to leverage a fully 3D UNet architecture to separate clean 21cm maps from radio foregrounds directly from simulated single-dish

observations. These clean maps can then be leveraged in existing 21cm analyses.

The layout of this paper is as follows: in Section 2, we present the physical formalism behind HI intensity mapping and astrophysical foregrounds. In Section 3, we present the foreground subtraction techniques we employ to train our network. We detail the architecture choice and the UNet training procedure in Section 3.3. Results for both blind subtraction and our network are presented in Section 4, followed by tests on foregrounds with altered simulation parameters. We discuss the successes of our network, as well as failure modes in Section 5. The cosmology we assume in our study is the standard flat ΛCDM in agreement with the results from the Planck Collaboration et al. [52], with fiducial parameters $\{\Omega_m, \Omega_b, h, n_s, \sigma_8\} = \{0.315, 0.049,\ 0.67, 0.96, 0.83\}$.

2 Methods and Formalism

In this section we describe the theoretical formalism behind the three main components of the observed HI 21cm sky: the cosmological signal itself, the various galactic and extra-galactic foregrounds, and observational noise. For a deeper discussion of various aspects of these components, we refer the reader to one of the several review papers on the subject such as [20, 48, 53] and [34]. We then describe how these various components are created in the simulated skies that we use to train our machine learning framework.

2.1 Cosmological HI Signal

The component of the 21cm sky that we care about most is the redshifted HI signal itself. This signal is often described in terms of a brightness temperature, T_b, which relates the observed intensity of the signal at a given sky position and frequency to a temperature [17, 18]. In the Rayleigh-Jeans limit ($\hbar\nu_{21} \ll k_B T_b$), this brightness temperature can be related to the underlying cosmology at a given line of sight $\hat{\mathbf{n}}$, and frequency, ν as [38]

$$T_b(\hat{\mathbf{n}}, \nu) = \frac{3\hbar c^3 A_{21}}{16 k_B \nu_{21}^2} \frac{(1+z)^2}{H(z)} n_{\rm HI}(z, \hat{\mathbf{n}})\,, \tag{2.1}$$

where $n_{\rm HI} \propto (1+\delta_{\rm HI})$ is the comoving number density of HI, $H(z)$ is the Hubble parameter as a function of redshift, z, k_B is Boltzmann's constant, ν_{21} is the frequency associated with the HI fine-structure line, $\hbar$ is the reduced Planck's constant, and $A_{21} = 2.876 \times 10^{-15}$ Hz is the 21cm line Einstein emission coefficient. Using the standard values for the various constants along with a standard flat ΛCDM cosmology for the evolution of $H(z)$, this expression can be written in terms of the HI overdensity, $\delta_{\rm HI} = \rho_{\rm HI}/\bar{\rho}_{\rm HI} - 1$, redshift, and cosmological parameters [20, 38] as

$$\begin{aligned} T_b(\hat{\mathbf{n}}, z) &= 0.19055 \times \\ &\frac{\Omega_b h (1+z)^2 x_{\rm HI}(z)}{\sqrt{\Omega_m (1+z)^3 + \Omega_\Lambda}} (1 + \delta_{\rm HI}\,(\hat{\mathbf{n}}, z))\ \text{mK} \end{aligned} \tag{2.2}$$

where $h \equiv H_0/(100\,\text{km/s/Mpc})$ is the dimensionless Hubble constant, $x_{\rm HI}$ is the fraction of baryonic mass comprised of HI, and Ω_b and Ω_m are the baryon and total matter fractions, respectively. The observed 21cm signal can thus be related to the underlying cosmological model and relevant parameters for inference studies [1, 8, 58, 70, 71]. The 21cm signal can thus be used as a tracer of the large-scale structure of the Universe.

Foreground Component	A [mK^2]	β	α	ξ
Galactic Synchrotron	1100	3.3	2.80	4.0
Point Sources	57	1.1	2.07	1.0
Galactic free-free	0.088	3.0	2.15	35
Extragalactic free-free	0.014	1.0	2.10	35

Table 1: Fiducial foreground $C_\ell(\nu_1, \nu_2)$ model parameters used in this study, adapted from [56] for the pivot values $\ell_{\rm ref} = 1000$ and reference frequency $\nu_{\rm ref} = 130$ MHz.

The `CRIME` simulation code, described in [3], generates a dark matter field and then utilizes a log-normal model to generate the cosmological HI intensity maps. Concisely, Gaussian density and velocity perturbations are generated on a Cartesian grid with no redshift effects. The "observer" is placed at the center of the grid, and the signal is projected onto the observer's light cone. The Gaussian density field then undergoes localized log-normal transformations to generate the non-uniform HI density field. The results are projected onto spherical sky maps at different frequencies corresponding to the redshift of the structure from the observer, using the `HEALPix` pixelization scheme [25]. The brightness temperature T_b is related to the underlying HI number density $n_{\rm HI}$, as shown in Equation 2.1. The simulations are generated in a box of size 8850 h^{-1} Mpc per side with 3072^3 box cells. The interpolated sky maps were generated using a `HEALPix` resolution of $N_{\rm side} = 256$, which corresponds to a per-pixel frequency-independent resolution of $\theta_{\rm pix} \approx 14'$.

2.2 Foregrounds

21cm foregrounds currently lack a detailed analytic description, making them difficult to separate from the cosmological HI signal. However, descriptive numerical simulations exist [e.g. 74], largely extrapolated from observed radio maps, such as the Haslam and Planck maps [27, 52]. The foregrounds that we hope to remove from the observed sky can be separated into galactic and extra-galactic components. Extragalactic foreground sources are expected to be distributed according to a clear power spectrum [13, 43, 56], while galactic foregrounds, such as synchrotron emission, are expected to be localized, particularly in the galactic plane [56, 74].

For galactic sources, the `CRIME` simulations extrapolate foregrounds from the 408 MHz map of Haslam et al. [27] to the relevant frequencies, as described in Santos et al. [56]. For weaker foregrounds such as point sources and free-free emission, as well as for synchrotron effects on small scales, we adopt Gaussian realizations of the generic power-spectrum based model

$$C_\ell(\nu_1, \nu_2) = A \left(\frac{\ell_{\rm ref}}{\ell}\right)^\beta \left(\frac{\nu_{\rm ref}^2}{\nu_1 \nu_2}\right)^\alpha \exp\left(\frac{-\log^2(\nu_1/\nu_2)}{2\xi^2}\right) \tag{2.3}$$

with values given in Table 1 (see [56] for details). While we train our network on these fiducial values, we explore model generalization to new foreground parameters in Section 4.6.

2.2.1 Polarized Foregrounds.

Foreground polarization arises when synchrotron emitting electrons traverse the Milky Way's magnetic fields, changing their polarization angles due to Faraday rotation [see e.g. 55]. De-

spite some empirical observations [e.g. 15, 57, 78], this effect on radio foregrounds is poorly understood, except perhaps at very low radio frequencies, as purported by the Experiment to Detect the Global Epoch of Reionization Signature (EDGES) [46]. Several models have been proposed to describe this phenomenon. The `CRIME` simulation package defines the Faraday depth at a distance s along a line of sight (LOS), $\hat{\mathbf{n}}$, as:

$$\psi(s, \hat{\mathbf{n}}) = \frac{e^3}{2\pi(m_e c^2)} \int_0^s ds' n_e(s', \hat{\mathbf{n}}) B_\parallel(s', \hat{\mathbf{n}}) \tag{2.4}$$

where m_e is the electron mass, and $n_e(s, \hat{\mathbf{n}})$ $B_\parallel$ are the number density of electrons and galactic magnetic field contribution for the given LOS. As shown in [3], the correlation over frequency for the polarization leakage field μ can then be written as:

$$\langle \mu_{lm}(\psi) \mu^*_{l'm'} \rangle \propto \delta_{ll'} \delta_{mm'} \left(\frac{\ell_{\rm ref}}{\ell} \right) e^{-\frac{1}{2}\left[\frac{\psi - \psi'}{\xi_{\rm polar}}\right]^2} \tag{2.5}$$

where the correlation length, $\xi_{\rm polar}$, and amplitude are free parameters. The rest of the numerical values are given in [3]. The polarized emission correlation length is related to the Fourier coordinate, k, via:

$$\xi_{\rm polar} = \frac{\pi H(z)}{c(1+z)k}. \tag{2.6}$$

For example, [62] use a correlation length scale of $k = 0.01 - 0.02$ Mpc^{-1}, yielding a $\xi_{\rm polar}$ of $0.1 - 0.05$ rad m^{-2}. We vary this parameter to probe failure modes in our foreground subtraction method in Section 4.6.3.

2.3 Observational Noise

The third component of the observed 21cm signal is (largely thermal) observational noise. Radio observational noise can be simply modeled as zero-centered Gaussian noise for single-dish experiments [8, 34].

We modify the white noise model with a stochastic component in order to train and test cleaning methods on a wide range of possible observational thermal noise, capturing a range of possible current and future intensity mapping configurations. For each full-sky simulation, we adopt a frequency-dependent hierarchical noise model, which has the added advantage of better training our networks (see Section 3.3):

$$\alpha_{\rm noise} \sim \log \mathcal{U}(0.05, 0.5) \tag{2.7a}$$

$$\sigma_{\rm noise} = \alpha_{\rm noise} \langle T_b(\nu) \rangle \tag{2.7b}$$

$$\epsilon_{b,i} \sim \mathcal{N}(0, \sigma_{\rm noise}) \tag{2.7c}$$

$$\hat{T}_{b,i} = T_{b,i} + \epsilon_{b,i} \tag{2.7d}$$

where we relate the variance of the noise to the average fiducial cosmological temperature at a given frequency, $\langle T_b(\nu) \rangle$. The observed signal at pixel i, written as $\hat{T}_{b,i}$ is then given by the true signal, $T_{b,i}$, with the addition of the Gaussian noise $\epsilon_{b,i}$. For comparison, the noise model employed by [2] and [70] correspond to an amplitude of $0.025 < \alpha_{\rm noise} < 0.12$. By sampling a large and competitive range of amplitudes for realizations of the per-pixel Gaussian noise, we allow the network to learn despite a variable range of noise.

It should be noted that, strictly speaking, our noise model allows for the observed signal $\hat{T}_{b,i}$ to be negative, which is unphysical, but not an uncommon observable in radio astronomy as a result of readout noise [see e.g. 77]. However, given the range of $\alpha_{\rm noise}$ values taken above, this is a very rare occurrence.

3 Foreground Removal Methods

Since foreground contaminants are several orders of magnitude brighter than cosmological signal, the biggest challenge for upcoming observational data analysis will be to separate the two signals [16, 43]. In this section we review blind data preprocessing for foreground subtraction, and introduce our improved method.

3.1 Blind Foreground Subtraction

Fortunately, foregrounds are expected to be spectrally smooth in frequency [36, 52, 66], while the cosmological signal is expected to vary with frequency according to Equation 2.1 [34, 43]. Current separation techniques therefore rely on the statistical distinctions between 21cm spectral components.

As we outlined in the last section, the observed 21cm signal can be largely modeled as the sum of three components: cosmological, noise, and foreground modes. Formally, we write:

$$T_{\rm obs}(\nu, \hat{\mathbf{n}}) = T_{\rm fg}(\nu, \hat{\mathbf{n}}) + T_{\rm cosmo}(\nu, \hat{\mathbf{n}}) + T_{\rm noise}(\nu) \tag{3.1}$$

where each component is described with respect to a given frequency, ν, and line of sight direction, $\hat{\mathbf{n}}$. In a system of discrete frequencies, we can write a linear system for each line of sight over frequency:

$$\mathbf{x} = \hat{\boldsymbol{A}} \cdot \mathbf{s} + \boldsymbol{\mathcal{C}_0} \tag{3.2}$$

where each $x_i = T_{\rm obs}(\nu_i, \hat{\mathbf{n}})$, and $A_{ik} = f_k(\nu_i)$ and $s_k = S_k(\hat{\mathbf{n}})$ are linearly separable basis functions and foreground sky components, respectively. The cosmological signal and thermal noise can then be packaged as $\boldsymbol{\mathcal{C}_0} = T_{\rm cosmo}(\nu, \hat{\mathbf{n}}) + T_{\rm noise}(\nu)$. We see that in these terms, foreground subtraction becomes a residual learning problem, such that we aim to reconstruct $\boldsymbol{\mathcal{C}_0} = \mathbf{x} - \hat{\boldsymbol{A}} \cdot \mathbf{s}$ as accurately as possible.

3.2 PCA Residual Analysis

Principal Component Analysis (PCA) makes use of the statistical properties of foreground signals by simultaneously fitting foreground sky maps, s_k, and smooth foreground functions, A_{ik} [2, 16]. Intuitively, PCA can be thought of as fitting a multidimensional ellipsoid to a feature space, with orthogonal axes (eigenvectors) pointing in the directions of the largest variance. PCA is an orthogonal transformation which maps the data from a set of basis vectors in which the data is correlated, to a basis in which the data is linearly uncorrelated. Since foregrounds are expected to be smooth and highly correlated in frequency [16, 43, 53], removing the components of largest eigenvalue (see Figure 10 in this work and Figure 1 in Alonso et al. [2]) is expected to preserve the cosmological signal on large angular scales relevant to cosmology. The method has been employed for foreground cleaning in both simulated and real 21cm data [10, 42, 65].

For our analysis, we repeat Alonso et al. [2]'s PCA removal procedure here. First we bin our observed maps (foreground and cosmological signal) into $N_\nu = 64$ frequency bands. We define a correlation matrix, $\mathbf{C}$ in frequency for all $N_{\rm pix}$ pixels in our simulation:

$$C_{ij} = \frac{1}{N_{\rm pix}} \sum_{n=1}^{N_{\rm pix}} \frac{T(\nu_i, \hat{\mathbf{n}}_n)\, T(\nu_j, \hat{\mathbf{n}}_n)}{\sigma_i \sigma_j}, \tag{3.3}$$

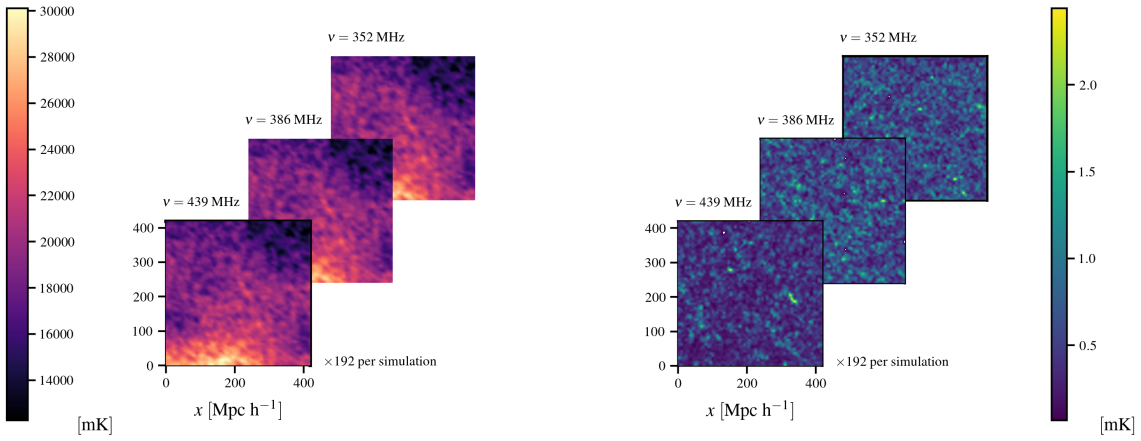


Figure 1: 2D slices from input foreground (left) and output cosmological (right) voxels for the `deep21` network. Each full-sky simulation is comprised of 192 `HEALPix` pixels at 690 different frequencies. We first diagonalize each sky in frequency and remove the first 3 principal components. We then take 64 frequencies from the first bin in Table 2 to generate 3D voxels of dimension 64^3 for `deep21` to process in batches of 16. Each epoch we process 80×192 training and 10×192 validation voxels.

where $T(\nu_i, \hat{\mathbf{n}}_i)$ is the observed 21cm map signal and σ_i are root-mean-square fluctuations of $\boldsymbol{\mathcal{C}_0}$ in mK, in the ith frequency band. Each σ_i is estimated iteratively from the data [2, 67]. The covariance $\mathbf{C}$ can then be diagonalized via eigenvalue decomposition:

$$\mathbf{\Lambda} = \mathbf{UCU}^T = \mathbf{diag}(\lambda_1, \ldots, \lambda_{N_\nu}), \tag{3.4}$$

where $\mathbf{\Lambda}$ is the diagonal eigenvalue matrix for $\mathbf{C}$, and $\mathbf{U}$ is an orthogonal matrix comprised of the corresponding eigenvectors. $\mathbf{\Lambda}$ is ordered by decreasing eigenvalues (principal components). For every pixel n we compute the PCA spectrum in frequency, and then project onto $\mathbf{\Lambda}$. We then remove the first $N_{\rm comp}$ components, and generate a filtered spectrum from the remainder, $\boldsymbol{\mathcal{C}_0}$, which is then assigned to the pixel as the output.

For our analyses, we preprocess observed radio sky maps and remove both the first three and first six principal components from our foreground maps. Henceforth, the notation PCA$-N_{\rm comp}$ corresponds to signal for which the first $N_{\rm comp}$ principal components have been removed.

Despite its ability to recover cosmological statistics in some frequency ranges, as demonstrated in [2, 16], it is important to realize that blind subtraction techniques do not produce clean HI intensity maps, since by definition PCA reduction removes the mean of the observed signal from single-dish observations. PCA is best suited to removal of foregrounds that behave smoothly in frequency. Unfortunately, the cosmological signal exhibits similar, smoothly varying structure on large scales, meaning PCA will remove cosmological clustering information needed for different studies, e.g. primordial non-Gaussianities [59]. Furthermore, polarization leakage from galactic synchrotron emission can create choppy foreground signal that is difficult for blind methods to distinguish from the cosmological signal (see [69] and Section 4.6.3).

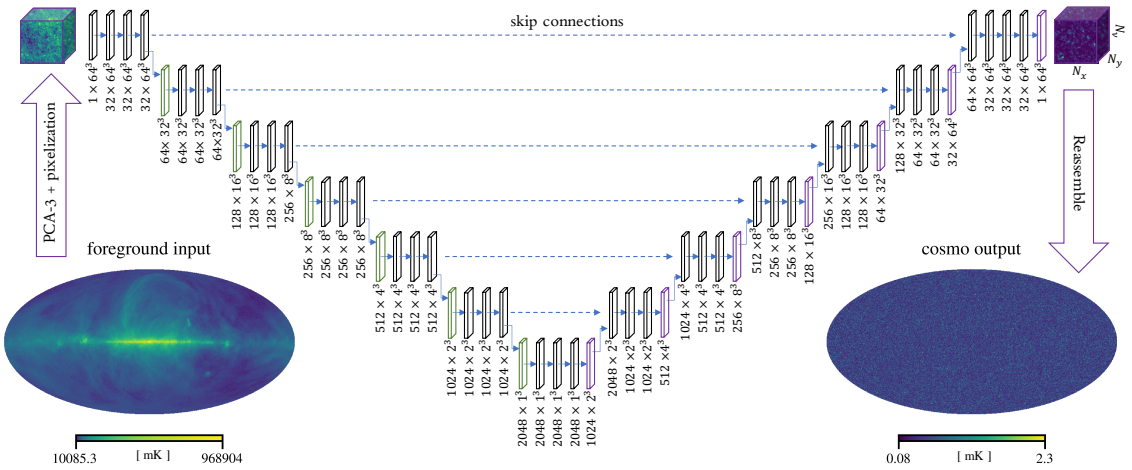


Figure 2: UNet Architecture and training scheme. We first remove the first three principal components from the simulated observed maps. We then split each map via the `HEALPix` pixelization scheme for the network to process. On the encoder side, input data undergo `w`=3 convolutions (black prism) at each level and are subsequently downsampled (green prism) `h`=6 times, halving the spatial dimensionality while simultaneously doubling the number of filters at each level. The data are then decoded via symmetric transposed convolutions (purple prism). Skip connections concatenate features at each depth, allowing the network to learn a specific correlation scale at a time.

3.3 `deep21` Neural Network

In this section we present our novel method for cleaning foregrounds from 21cm maps. We adopt a convolutional neural network (CNN) architecture based on the UNet model of [54], which maps images to images via a symmetric encoder-decoder convolution scheme.

3.3.1 Input Preprocessing

Since PCA has been shown to effectively remove the majority of foregrounds [2, 16], we focus our analysis on recovering the physical, cosmological signal from the PCA residuals. We feed in PCA-3 residual maps, processing input maps using `scikit-learn` [50], which implements [67]'s probabilistic PCA algorithm. Removing the first three components centers input signal on zero and drastically reduces input amplitudes, but does not remove too much small-scale cosmological clustering, as shown in [2]. PCA preprocessing thus has the added benefit of scaling inputs appropriately for neural networks, which perform best for inputs in the range $[-1, 1]$ [24]. Unlike previous analyses (see e.g. [2, 14, 16, 70]), we do not perform instrument-dependent Gaussian beam smoothing as a preprocessing step, since we want to test the ability of our model to recover signal in the limit of pixel size resolution.

3.3.2 Dataset Assembly

To test our foreground separation methods, we generated a suite of 100 full-sky cosmological and foreground `CRIME` simulations over 690 frequencies, $350 < \nu < 1044$ MHz, each separated by $\Delta\nu \sim 1$MHz. We then add a copy of the foreground and cosmological maps together, and split the dataset into 80 training simulations, 10 validation, and 10 (hidden) test simulations.

The UNet architecture is ideally suited to image-like data. For this reason, we split each simulation into 192 equal-area windows via the `HEALPix` pixelization scheme [25]. We then stack maps in frequency, drawing 64 frequencies evenly within the designated redshift bin, yielding 192 cubic voxels of dimension $(N_{\theta_x}, N_{\theta_y}, N_\nu) = (64, 64, 64)$ pixels, shown schematically in Figure 1. According to the train-validation-test split, the network sees $80 \times 192 = 15,360$ training voxels each epoch, followed by $10 \times 192 = 1,920$ validation voxels. We set aside 10 hidden test simulations with which to assess our cleaning methods.

The UNet architecture maps input voxels to output voxels via a contracting path, in which input dimensionality is halved and feature channels doubled iteratively via stride-2 downsampling convolutional layers, depicted schematically by green prisms in Figure 2. At each depth of the network, we perform `w` convolutions represented by black prisms. Data are then upsampled via transposed convolutions (purple prism) through a symmetric upsampling path. Skip connections concatenate features from one side of the network to the other, allowing each depth of the network to focus on learning a specific scale of correlations at a time. These correlations are then summed as the network upsamples on the output side. We employ batch normalization between convolutional layers, except for the last convolutional block on the output side. Once voxels have been processed by the network, we reconstruct the full-sky cosmological maps to compute power spectra and clustering statistics.

3.3.3 Loss Function

To train the networks, we would like to minimize a pixel-wise loss function of the form $\mathcal{L}(p,t) = \sum_i L(|p_i - t_i|)$ between prediction, p, and simulation target, t of each i^{th} voxel, and $L(x)$ is the pixel-wise loss function. We find empirically that the standard Mean Square Error (MSE) loss proved volatile early in training. For this reason we selected the Log-Cosh loss function

$$\mathcal{L}(p,t) = \sum_i \log\cosh(p_i - t_i)\,. \tag{3.5}$$

This function behaves much like the L1 norm for poor predictions (large values of $|p_i - t_i|$), making it robust to outliers, and approaches $(p_i - t_i)^2/2$ for small residuals. For network performance validation and test statistics we look at the Log-Cosh loss, as well as the standard MSE metric.

3.4 Training Procedure

In training our selected architecture we make use of a combination of the `AdamW` optimizer and a step-wise learning rate reduction conditioned on validation data. This choice of learning routine provides weight decay regularization, as well as a fine-tuning of network optima.

Every training epoch the network sees 80 simulations of foregrounds added to cosmological signal, subject to a new observational noise realization for a sampled $\alpha_{\rm noise}$. The observed maps are first preprocessed by the PCA-3 subtraction and then pixelized into `HEALPix` voxels described above. The PCA-3 residuals are then processed by the UNet network.

We split our frequency range to test our foreground cleaning in the context of analyses such as [70]. Redshift shells and corresponding co-moving distances are shown in Table 2. For our analysis we focus on evaluating network performances in the co-moving shell of lowest frequency, since these high-redshift regions are interesting for both Baryonic Acoustic Oscillation (BAO) measurement, as well as post-EoR structure analysis [53]. Furthermore, these regions have consistently proven difficult for blind foreground techniques to clean, especially in angular power spectrum recovery [3, 70].

3.5 Hyperparameter Tuning

In choosing our network architecture, we first compared architectures with 2D and 3D convolutional kernels with different network depths. We anticipate that training 3D convolutional kernels perform better than 2D convolutions, since inputs in this scenario are treated as full 3D volumes, capturing frequency patterning in ν, as well as angular patterns in θ_x and θ_y. However, we explore both 2D and 3D kernels as hyperparameters of our architecture.

Other important hyperparameters we considered were UNet depth, `h`, or number of down-convolutions (denoted by green prisms in Figure 2), and the number of convolutions at a given dimension, `w` (convolutional block width). To test hyperparameters, we developed a dynamic UNet model compatible with the `HyperOpt` Python library [5]. Our architecture draws hyperparameters from proposal distributions and trains the resulting architecture on a smaller set of training data. We selected the hyperparameter combination that yielded the lowest validation loss after testing 550 trial architectures. The priors from which we drew our hyperparameters are listed in Appendix A. The results yielded the interesting result that deeper, wider UNets outperformed shallower networks (see [44] for a theoretical investigation). We found that in particular, architectures with network width $2 < \mathtt{w} < 4$ and height satisfying

$$\mathtt{h} = \log_{\mathtt{stride}} \mathtt{n}_{\mathrm{filters}}; \quad \mathtt{stride} = 2 \tag{3.6}$$

consistently yielded the lowest losses. Too many convolutions per block frequently obscured the sharp T_b distribution (see Figure 6), and Equation 3.6 guarantees an architecture that compresses inputs down to dimension 1^3 for stride-2 down-convolutions, meaning the network learned correlations on a pixel-sized scales. Our optimized architecture, henceforth `deep21`, is displayed graphically in Figure 2.

Neural network outputs are generally not probabilistically interpretable [12], and no tractable image-producing Bayesian neural networks currently exist, so quantifying the variability of foreground cleaning on a given dataset is not possible with a single UNet.

Motivated by the methodology of deep ensembles [19, 32] we train an ensemble of $M = 9$ networks with independently Glorot-Uniform-initialized weights [23] for 300 epochs in parallel. The `HEALPix` data inputs are also subject to the stochastic noise model as before, as well as random sky-sized rotations on the sphere, such that the networks learn to denoise pixels independently of orientation. To gauge the uncertainty of the cleaning method, each ensemble member then performs foreground separation on ten test simulations, subject to competitive observational noise with a fixed $\alpha_{\mathrm{noise}} = 0.25$, falling in the upper range of our amplitude prior and roughly twice the maximum noise amplitude utilized by Alonso et al. [2]. This practice allows us to estimate the epistemic uncertainty on the summary statistics obtained from the predicted full-sky maps.

4 Results and Analysis

4.1 Visual Inspection

Most current and upcoming cosmological experiments are aimed at reconstructing relevant clustering statistics from brightness temperature maps. Therefore it is prudent to ensure that our network outputs clean maps that capture temperature distributions at given frequencies. An initial check for `deep21`'s performance is a qualitative one: Figure 3 shows each cleaning method's performance on a test `HEALPix` pixel slice. From left to right we display simulated target cosmological signal, network PCA-3 inputs, and `deep21` prediction. `deep21` is able to

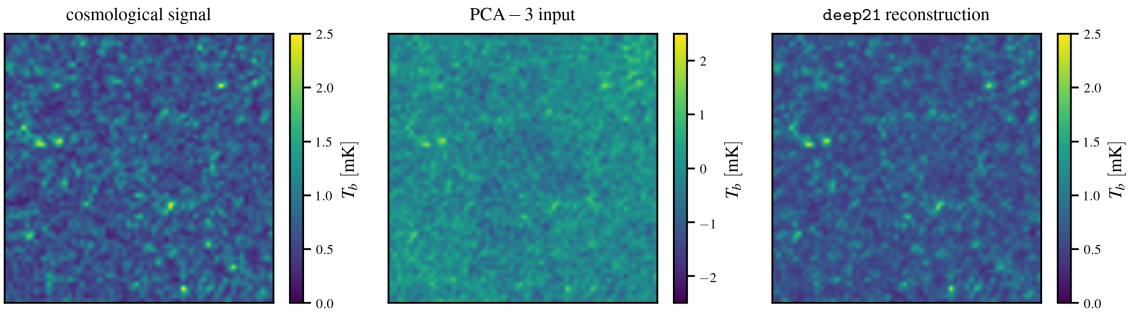


Figure 3: 2D slices at $\nu = 392$ MHz comparing PCA-3 UNet inputs (*middle*) to UNet ensemble prediction (*right*) and target cosmological signal (*left*). `deep21` is able to reconstruct the signal even after PCA subtraction removes the mean of the observations.

recover intricate cosmological signal from corrupted inputs almost indistinguishable from the simulated targets.

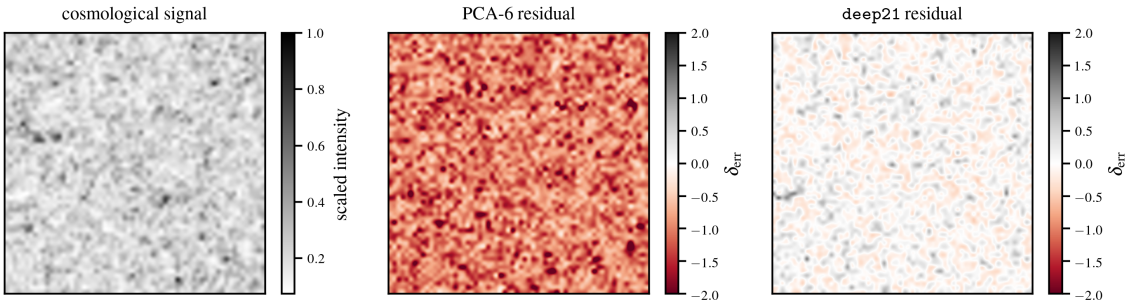


Figure 4: Temperature map residuals for PCA-6 (*middle*) and `deep21` (*right*) map residuals compared to scaled intensity of the simulated cosmological signal (*left*) for the same 2D slices as Figure 3. The UNet ensemble recovers a much more accurate tomography than the PCA method, the latter failing to capture details around high-intensity regions and voids. Removal of the first moment in the PCA method results in a significant deviation ($\sim 200\%$) from the true signal, while `deep21` predictions yield small, positive residuals.

In Figure 4, we compare the per-pixel scaled cosmological signal (greyscale panel) to the relative residual error for cosmological brightness temperature T_b,

$$\delta_{\mathrm{err},i} = \frac{p_i - t_i}{t_i} \tag{4.1}$$

where p_i and t_i are the pixel-wise predicted and target signals, respectively. Here we compare the best-case (PCA-6) blind subtraction to `deep21`. We note that PCA predictions over-subtract the signal, particularly in low-density regions. `deep21`'s residuals are all well within order unity of the target, with some deviations above zero.

4.2 Clustering Statistics

The most important cosmological parameter constraints from HI intensity mapping will most likely come from power spectra of the 21cm brightness temperature, since two-point correlation functions contain the vast majority of information regarding underlying cosmology on large, linear, scales. For this study, we consider angular and radial power spectra separately, capturing clustering patterning on the sky and along each line of sight, respectively.

For a fixed frequency and assuming a full-sky survey, the angular power spectrum of the brightness fluctuations ΔT_b is computed first by calculating the spherical harmonic components:

$$a_{\ell m}(\nu) = \int d\hat{\mathbf{n}}^2 \Delta T_b(\nu, \hat{\mathbf{n}}) Y^*_{\ell m}(\hat{\mathbf{n}}), \tag{4.2}$$

where $Y_{lm}(\hat{\mathbf{n}})$ are the spherical harmonic basis functions. We can then estimate the power spectrum by averaging over the moduli of the harmonics:

$$\widetilde{C}_l = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} |a_{lm}|^2 \tag{4.3}$$

where small ℓ correspond to the largest scales. We calculated the angular power spectra for our maps using the `healpy` Python library [82].

To capture radial clustering in an HI survey independent of redshift effects, one must make two assumptions, namely 1) each line-of-sight (`HEALPix` pixel) window satisfies the flat-sky assumption [3], and 2) that the redshift bin under consideration is narrow enough that no significant cosmological expansion occurs between the edges of the bin. The resulting power spectrum then describes the clustering distribution independently of cosmological expansion effects. Given these two assumptions, we can average over all possible radial lines of sight, $i = 1, \ldots, N_\theta$, to obtain the radial power spectrum:

$$P_\parallel(k_\parallel) = \frac{\Delta\chi}{2\pi N_\theta} \sum_{i=1}^{N_\theta} |\widetilde{\Delta T_b}(\hat{\mathbf{n}}, k_\parallel)|^2 \tag{4.4}$$

where the Fourier coefficients $\widetilde{\Delta T_b}$ are estimated using the Fast Fourier Transform (FFT) over each line of sight, and $\Delta\chi = \chi(z_{\rm max}) - \chi(z_{\rm min})$ is the comoving width of the given redshift bin. Given a constant frequency interval separating the spherical surfaces, $\delta\nu$, the same interval is expressed in the conjugate space as $\delta k_\nu = 2\pi/\Delta\nu$. The radial coordinate, $k_\parallel$, can then be defined as [3]:

$$k_\parallel = \frac{\nu_{21} H(z_{\rm eff})}{(1 + z_{\rm eff})^2} k_\nu, \tag{4.5}$$

where $z_{\rm eff}$ is the effective redshift for the comoving volume under consideration.

To capture uncertainty over the space of the `deep21` ensemble, we compute a weighted average, $\bar{Z}_w$, and standard deviation, $\sigma_w(Z)$, of each network's independent estimate of a given statistic, Z. We employ proper scoring weights by computing the inverse globally-averaged MSE for each network's prediction over test data: $w_{\rm m} = \frac{1}{\langle {\rm MSE} \rangle}$ for $m = 1, \ldots 9$ independent networks. Here we do not explicitly assume a Gaussian form, so σ does not correspond to the 68% inclusion interval. We note that to apply this procedure in a real observational setting, scores for each ensemble member would need to be computed for a set of validation simulations, and then used to compute statistics obtained from the cleaned observed sky.

ν (MHz)	z	$\langle z \rangle$	Vol. $(h^{-1}\text{Gpc})^3$	N_{side}
$[886 - 1044]$	$[0.36 - 0.60]$	0.47	12	256
$[667 - 886]$	$[0.60 - 1.12]$	0.87	50	256
$[476 - 667]$	$[1.12 - 1.88]$	1.50	108	256
$[350 - 491]$	$[1.88 - 3.05]$	2.47	187	256

Table 2: Comparison of the four redshift co-moving shells used in the UNet assessments. Co-moving shells were chosen by splicing the simulation into bins with equal numbers of frequency. Our analysis focuses on the highest-redshift bin because this is where blind foreground techniques such as PCA reduction have been shown to perform the worst in the literature.

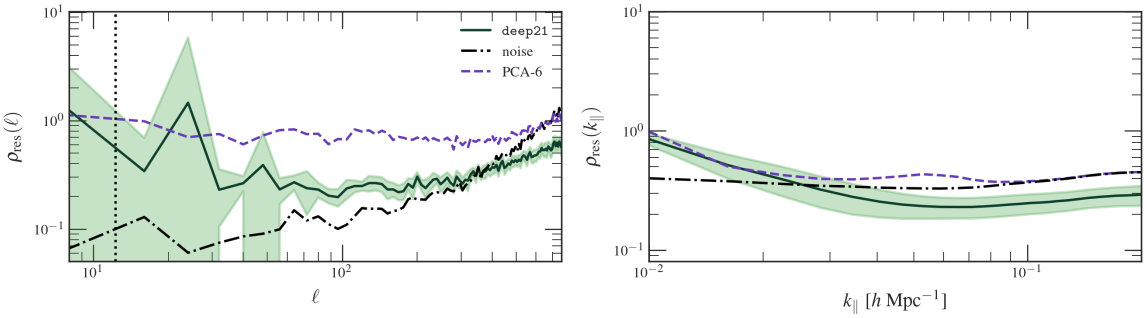


Figure 5: Comparison of angular (*left*) and radial (*right*) residual map power spectra as a fraction of the target cosmological signal predicted by the `deep21` UNet ensemble (green) blind PCA reduction (purple), and noise realization (black) with $\alpha_{\text{noise}} = 0.25$ over a single full-sky test simulation. Confidence intervals corresponding to $\pm 2\sigma_w$ over the space of ensemble parameters is estimated in shaded green. We display `deep21`'s angular resolution via the black dashed vertical line. The angular power spectrum shown is computed for a single frequency, $\nu = 357$ MHz, while the radial power spectrum is computed for the lowest frequency bin, with mean redshift $\langle z \rangle = 2.5$. The network successfully learns to marginalize out additive observational noise in the radial direction at smaller scales ($k_\parallel > 0.015$).

We consider the power spectra calculated for the residual maps for each cleaning method, defined for $P \in \{C_\ell, P_\parallel\}$ as:

$$\rho_{\text{res}} = \frac{P_{\text{res}}}{P_{\text{cosmo}}} = \frac{P(p-t)}{P(t)} \tag{4.6}$$

where t is the target cosmological signal and $p \in \{T_{\text{PCA}}, T_{\texttt{deep21}}\}$ is the given cleaning method's predicted map. We additionally consider ρ_{res} computed for the noise map, $P(T_{\text{noise}})$, generated for the test simulation.

This statistic quantifies each cleaning method's residuals as a fraction of the true cosmological signal in both the angular and radial directions. The noise residual demonstrates to what degree observational error obscures the structure estimate.

We compare `deep21` to the PCA-6 residual noise realization generated with $\alpha_{\text{noise}} = 0.25$ in Figure 5. `deep21` (green) outperforms PCA (purple) in both the angular and radial directions, and successfully fits cosmological signal at small scales despite high observational noise contribution (black dashed line). We interpret this result as a successful marginalization of the observational noise. Through training, `deep21` has learned to distinguish cosmological clustering from noise fluctuations at small scales. By contrast, the PCA-6 residual asymptot-

ically approaches the noise boundary in both plots, indicating a limit to the blind foreground cleaning at small scales. `deep21` also substantially reduces the loss of signal at large radial scales incurred by the PCA method. The larger PCA-6 residual at small $k_{\parallel}$ indicates large-scale information lost to the foreground subtraction as demonstrated in [2].

4.3 Intensity Distributions

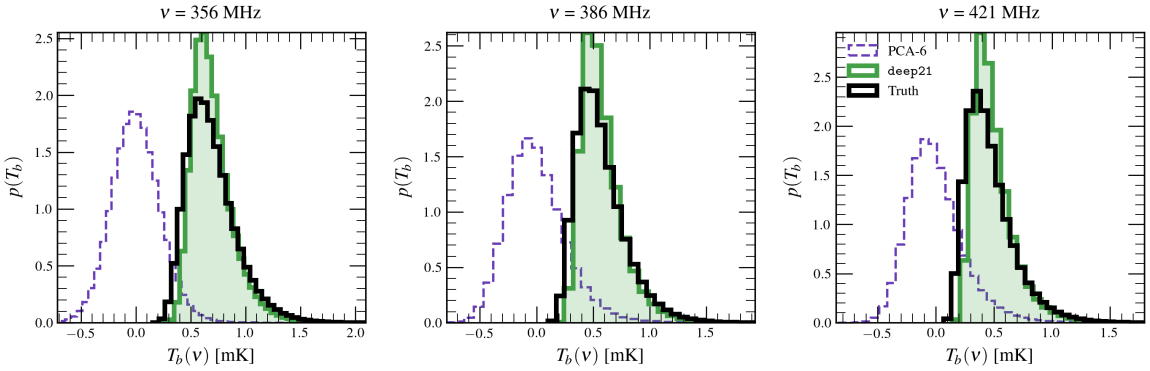


Figure 6: Comparison of the distribution of pixel temperatures from the PCA-6 (purple) and `deep21` (green) cleaned maps to those of the cosmological simulations (black) at several different frequencies. PCA reduction centers temperature distributions on zero and does not capture the asymmetric cosmological signal.

We also compare how well each cleaning method captures the distribution of the cosmological signal at each frequency. Figure 6 compares cosmological temperature distributions at several frequencies throughout a single test simulation. We see that the PCA method reproduces a more symmetric temperature distribution that is zero-centered. This result is anticipated by the definition of the method (which removes the mean of the distribution to diagonalize the signal in frequency). It is clear from comparison to the true signal that `deep21` captures the asymmetric distribution much more accurately than the PCA.

4.4 Power Spectrum Recovery

We additionally report clustering statistics based on the analysis by [2]. We introduce the statistic

$$\varepsilon = \frac{P_p - P_{\rm cosmo}}{P_{\rm cosmo}} \tag{4.7}$$

where P_p is the predicted power spectrum from the foreground cleaning method under consideration, and $P_{\rm cosmo}$ is the power spectrum generated by the target cosmological map. We also compute ε for the power spectrum computed from a noisy cosmological map. This statistic quantifies the bias introduced to the power spectrum by each cleaning method (or noise) as a fraction of the target power spectrum. For a perfect foreground removal, ε approaches zero for all angular scales and frequencies. We can equivalently define the transfer function for coordinate $k \in \{\ell, k_{\parallel}\}$

$$\mathrm{T}(k) = \sqrt{\frac{P_p(k)}{P_{\rm cosmo}(k)}} \tag{4.8}$$

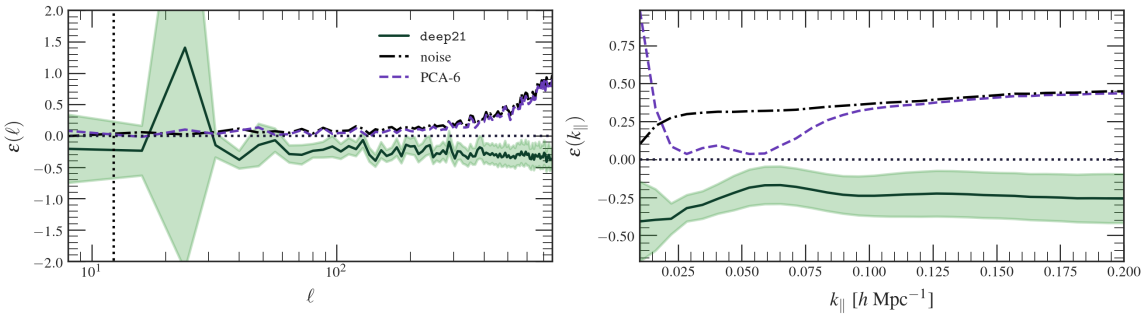


Figure 7: Residuals of power spectrum statistics for both angular (left) and radial (right) power spectra. The angular spectrum is computed at $\nu = 401$ MHz, with the angular resolution of a `deep21` input voxel shown as the black dashed vertical line. `deep21` outperforms the PCA-6 subtraction on most angular scales, with $\pm 2\sigma_w$ confidence intervals largely consistent with zero. The PCA method outperforms the network on intermediate radial scales, but is subject to poor performances on the largest radial scales and converges to the observational noise residual at small radial scales. `deep21`, by contrast, is able to reproduce the radial power spectrum with $\sim 90\%$ accuracy at most radial scales, illustrating its consistency in foreground cleaning in the line-of-sight direction. `deep21`'s outperformance of the noise boundary indicates that the network has also effectively learned to marginalize out added white noise in addition to the remaining foreground components following a PCA-3 cleaning.

which approaches 1 for a perfect foreground cleaning.

We compare the ε statistic for angular and radial power spectra in Figure 7. `deep21` (green) performs consistently at both large and small angular and radial scales, capturing power spectra with $\sim 90\%$ accuracy within the comoving bin. Furthermore, `deep21` outperforms ε computed for the noisy map, indicating that the UNet method's marginalization of the noise also prevents power spectrum bias at small scales. The PCA-6 subtraction (purple) recovers a lower $\varepsilon(k_\parallel)$ than the UNet only on intermediate radial scales ($0.02 < k_\parallel < 0.065$), subject to loss of signal at large scales and convergence with the observational noise at small angular and radial scales. This behavior demonstrates that blind methods are unable to remove the power spectrum bias introduced by noise at small scales.

The consistency of the deep learning approach should be emphasized here: the `deep21` method is largely scale-independent in foreground removal and signal reconstruction. Moreover, these results are not reliant on statistical marginalization over many data realizations like the analyses done in [2] and [70], meaning fewer observations need to be made in a realistic setting in order to achieve a consistent, successful foreground removal with estimated uncertainties. Furthermore, with more computational power, larger voxels can be expected to be processed in the future, resulting in an improvement in large-scale network recovery.

4.5 Noise Performance

`deep21`'s training procedure includes variable levels of observational noise to improve test performance and incorporate uncertainty regarding noise levels in upcoming intensity mapping experiments. To test whether or not the ensemble learned to marginalize out this effect, we tasked the trained network with cleaning the same foreground simulation with a variable noise amplitude, $\alpha_{\rm noise}$. We plot the corresponding MSE metric in Figure 8. The network's

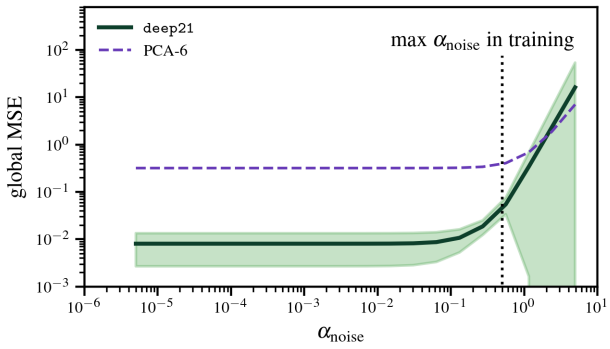


Figure 8: Noise performance testing for `deep21` and PCA-6 cleaning routines. 20 maps were generated with increasing noise amplitude $\alpha_{\rm noise}$. As anticipated, `deep21` performs consistently for $\alpha < \alpha_{\rm max} = 0.5$, since this represents the level of noise the network encountered during training according to Equation 2.7. The network's error quickly increases in variance and magnitude beyond this threshold.

test MSE remains fairly constant until we exceed the noise threshold, $\max\{\alpha_{\rm noise}\} = 0.5$, encountered in training. This shows that `deep21` indeed captures the statistical properties of the observational noise it encountered during training, and subsequently marginalizes over it. In contrast, the blind PCA subtraction, whose MSE is dominated by the removal of the first moment of signal, does not increase in MSE until $\alpha_{\rm noise} = 1$, or order unity with the mean cosmological signal at a given frequency (see Equation 2.7). This result shows that a network ensemble can be trained to be robust to observational noise, so long as a statistical model is specified and varied enough during training.

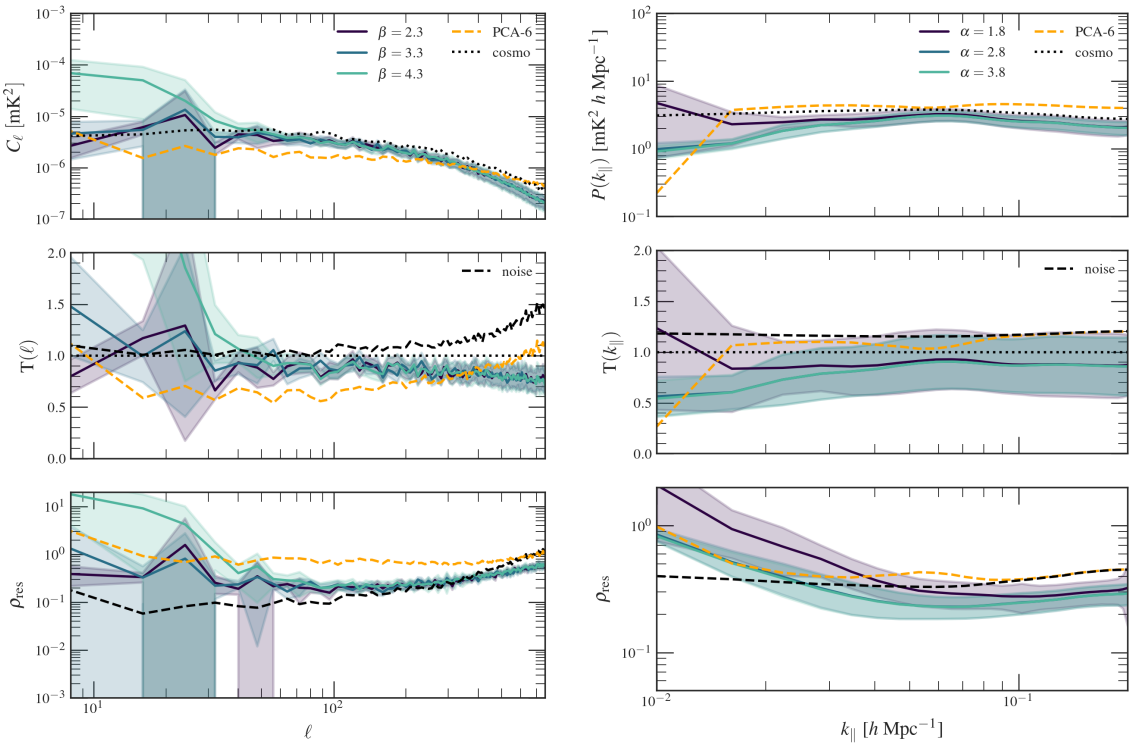
4.6 Generalization to new foreground parameters

Despite training `deep21` on many foreground and cosmological realizations, we assumed the same fiducial model for all training data generation. For a foreground cleaning experiment with real data, one would ideally train `deep21` on a range of foreground and cosmological models. As a final test for our trained UNet ensemble, we task the network, trained on fiducial simulation parameters, with cleaning observed signal generated by different input parameters.

To do this, we chose to alter the galactic synchrotron foregrounds via the `CRIME` simulation package, since these represent the largest of the foreground contaminants. Since the PCA processing removes synchrotron amplitude information, we elected to alter its correlation structure according to Equation 2.3, namely 1) angular scale dependence, 2) frequency dependence and 3) switching on polarization effects. We display recovered power spectra for the correlation structure analysis in Figure 9.

4.6.1 Varying galactic synchrotron angular correlation dependence.

Having trained `deep21` on foregrounds with fiducial galactic synchrotron $\beta_o = 3.3$, we vary the synchrotron ℓ dependence according to Equation 2.3 by $\pm 30\%$ to $\beta = 1.3\beta_o$ and $\beta = 0.7\beta_o$, generating a new full-sky simulation of 192 `HEALPix` voxels for each case. Increasing β yields a smaller foreground correlation in ℓ, reflected in `deep21`'s over-estimate of the power spectrum in Figure 9a. Likewise, the network under-estimates the power spectrum for lower $\beta = 2.3$.



(a) Varying galactic synchrotron ℓ dependence. (b) Varying galactic synchrotron ν dependence.

Figure 9: Generalization testing for `deep21`. We compare two-point statistics on test data by varying foreground parameters, keeping training data and trained network ensembles constant. We vary the galactic synchrotron β and α in Figures 9a and 9b, respectively, according to Equation 2.3. We display power spectra (top), transfer function, i.e. the square root of predicted to true spectrum (middle), and residual spectra (bottom) for both angular (at $\nu = 357$ MHz) and radial statistics.

This indicates that the network has indeed captured the fiducial foreground model in the training data.

4.6.2 Varying galactic synchrotron frequency correlation dependence.

We repeated the same analysis for the α parameter, varying the galactic synchrotron C_ℓ model's dependence on frequency by $\pm 35\%$. We recovered a similar trend in `deep21` performance in Figure 9b. Here, decreasing α results in a smaller correlation amplitude as a function of frequency. `deep21` produces an over-estimate of the radial power spectrum since it is trained on $\alpha = 2.8$. Increasing $\alpha_v = 1.35\alpha_o$ produces little difference in the ensemble estimate, indicating that the model might generalize well in this regime.

In contrast, PCA-6 cleaning (dashed orange) yields almost identical results for changes in the correlation parameters, indicating that the blind method is robust to changes in correlation structure. We display foreground cleaning results at different scales for our test and generalization cases in Table 3.

4.6.3 Polarized foregrounds

We also tested simulations contaminated by a galactic synchrotron polarization leakage of 1%. Polarized foregrounds due to the Milky Way's magnetic effects could wreak a catastrophic effect on signal recovered using blind techniques, since poorly-understood polarization leakage could induce foregrounds to interfere with cosmological signal modes [3, 34]. To motivate a follow-up study, we enabled galactic synchrotron polarization within the `CRIME` simulation package, varying the polarized correlation length $\xi_{\rm polar}$, and cleaned the resulting observed maps with our technique. We recovered substantial differences in performance for PCA-6 and `deep21`, as shown in Figure 10. Reducing the correlation length, $\xi_{\rm polar}$ makes leaked galactic synchrotron foreground emission behave similarly to the cosmological signal in frequency (see lefthand plot for $\xi_{\rm polar} = 0.01$, and Figure 8 in [3]). The PCA subtraction fails since synchrotron foregrounds can no longer be smoothly resolved from the choppy cosmological signal. This is shown formally in [2], Figure 1, where decreasing polarization correlation length spreads foreground contamination across diagonalized PCA eigenvalues, making it more difficult to know when foregrounds have been successfully removed. As expected, `deep21` does not generalize well to polarized foregrounds, since PCA-3 preprocessing is unable to remove the 1% synchrotron leakage. We do, however, note that `deep21` does produce a slight improvement in radial power spectrum recovery (Figure 10, right-hand side).

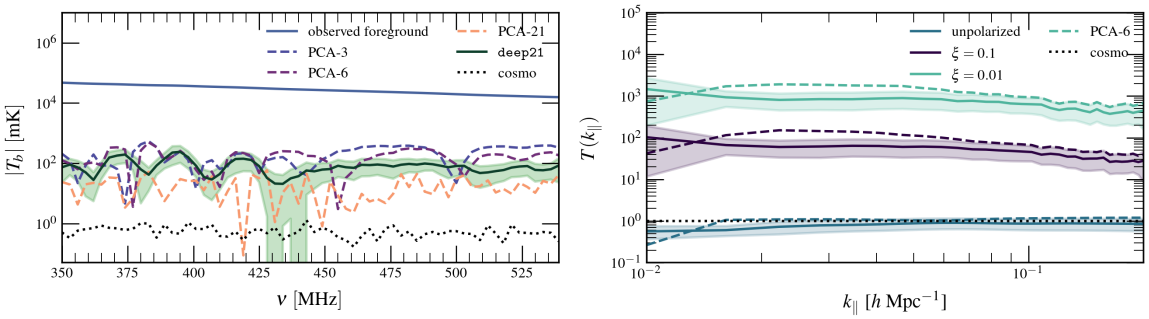


Figure 10: Foreground cleaning errors due to 1% galactic synchrotron polarization leakage. (*Left*) `deep21` temperature recovery compared with several PCA subtractions for $\xi_{\rm polar} = 0.01$. Removing as many as 21 PCA components does not resolve the cosmological signal in the presence of polarization leakage. (*Right*) Transfer function for `deep21` recovery of radial power spectrum recovery for various values of $\xi_{\rm polar}$. Here the PCA-6 subtraction is shown for the given simulation as a dashed line, and shaded contours show $\pm 2\sigma_w$ `deep21` estimates. Shrinking the polarization correlation length makes leaked galactic synchrotron foregrounds behave similarly to the cosmological signal in frequency, rendering blind methods and `deep21` ineffective.

For a more complete understanding of polarized foregrounds and `deep21`'s effectiveness in removing them, a follow-up study is warranted. Existing codes such as `Hammurabi` [74] make use of detailed three-dimensional Milky Way simulations to model the magnetic fields responsible for polarization leakage. Training `deep21` on these detailed foregrounds is a necessary next step for real data-preparedness.

	MSE (global)	T(ℓ) $\ell = 50$	T($k_\parallel$) $k_\parallel = 0.02$	$\rho_{\rm res}(\ell)$ $\ell = 50$	$\rho_{\rm res}(k_\parallel)$ $k_\parallel = 0.02$	T(ℓ) $\ell = 550$	T($k_\parallel$) $k_\parallel = 0.15$	$\rho_{\rm res}(\ell)$ $\ell = 550$	$\rho_{\rm res}(k_\parallel)$ $k_\parallel = 0.15$
test phase									
deep21	0.877±0.156	0.899±0.186	0.728±0.087	3.857±3.178	1.099±0.307	0.848±0.085	0.868±0.088	1.099±0.13	0.648±0.115
PCA-6	17.15	0.611	1.095	8.2	1.184	0.814	1.195	2.417	1.005
$\beta = 2.3$									
deep21	0.872±0.152	0.895±0.212	0.728±0.088	4.545±4.414	1.076±0.3	0.839±0.086	0.867±0.088	1.063±0.135	0.649±0.115
PCA-6	17.15	0.542	1.093	11.49	1.183	0.773	1.194	2.397	1.005
$\beta = 4.3$									
deep21	1.044±0.289	0.977±0.19	0.738±0.081	5.83±5.094	1.297±0.46	0.853±0.114	0.874±0.103	1.344±0.291	0.666±0.132
PCA-6	17.16	0.607	1.092	9.025	1.21	0.763	1.195	2.805	1.006
$\alpha = 1.3$									
deep21	1.764±1.056	1.57±0.505	0.844±0.081	23.761±23.8	1.909±0.785	0.867±0.111	0.87±0.114	1.588±0.321	0.691±0.143
PCA-6	17.15	0.573	1.058	11.35	1.298	0.781	1.195	2.707	1.005
$\alpha = 3.3$									
deep21	0.871±0.152	0.907±0.229	0.727±0.088	3.135±3.146	1.073±0.3	0.823±0.083	0.867±0.088	1.164±0.151	0.649±0.114
PCA-6	17.15	0.534	1.095	7.945	1.186	0.755	1.194	2.536	1.005

Table 3: Summary of foreground cleaning results. All residuals and MSE metrics are normalized to the corresponding statistic computed for observational noise generated with $\alpha_{\rm noise} = 0.25$. Angular power spectra are computed for a slice at $\nu = 357$MHz. Uncertainty intervals for `deep21` were computed for $\pm 2\sigma_w$ for each statistic.

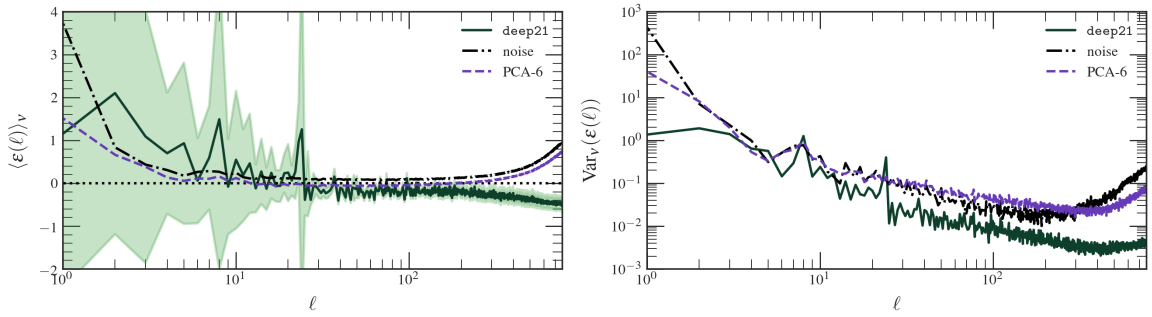


Figure 11: Mean and variance of angular power spectrum accuracy, $\epsilon(\ell)$ over frequencies in the first comoving shell for PCA (purple) and `deep21` methods (green, shaded $\pm 2\sigma_w$). The `deep21` ensemble power spectrum recovery is consistent with zero at all scales and provides a more precise angular spectrum recovery than that of the more variable PCA, especially on the smallest angular scales.

5 Conclusions

In this study we developed a deep learning-based method to improve foreground cleaning techniques for single-dish 21cm cosmology. Our method outputs clean intensity maps instead of derived summary statistics. Training an ensemble of independent UNet architectures on simulated foregrounds and cosmological signal resulted in the improved recovery of intensity maps and angular and radial power spectra (especially at small angular scales, see Figure 11). In addition, the ensemble method provides an estimate of uncertainty for the summary statistics of resulting maps. We show that `deep21` effectively marginalizes out observational noise at small angular and radial scales, demonstrating a marked improvement in the sensitivity limit of foreground subtraction over PCA. We demonstrated that `deep21` is sensitive to foreground physics by changing test data simulation parameters, meaning deep networks trained on more detailed (and varied) simulations will likely be able to effectively remove

foregrounds in absence of a formal foreground likelihood. We also investigated `deep21`'s failure modes, namely in the presence of galactic synchrotron polarization leakage. Improved networks trained on polarized foregrounds (including those which require no PCA preprocessing) will be the subject of a future work to mitigate these effects on radio map retrieval. Our method demonstrates that cosmological analyses on previously irretrievable 21cm intensity maps may be possible in an observational setting.

5.1 Future Work

The methods outlined here utilize a simulation-based deep learning method to retrieve intensity maps for radio cosmology. These techniques pave the way for more fundamental studies of the 21cm signal captured by upcoming SKA experiments, as well as future studies with even higher resolution. The ability to retrieve intensity maps will allow future studies to probe structure formation and EoR physics beyond the power spectrum statistic.

However, before `deep21` can be applied to real 21cm data, several key aspects should be addressed in follow-up studies:

- A more realistic noise model: the white noise model considered here, while frequency-dependent, will not extend to more complicated intensity map datasets, such as those obtained via interferometry [60]. Thus the impact of nontrivial noise correlations on network performance must be considered before real signals can be reliably separated. A `deep21`-like study on systematics would additionally benefit from varying observed sky fraction, as is done in Villaescusa-Navarro et al. [70] for 21cm BAO recovery.

- Varying astrophysical parameters: here we trained `deep21` on thousands of input voxels derived from the same fiducial cosmological and foreground parameters. While the network performed relatively well in some generalization cases, `deep21` would ideally be trained on a range of simulation parameters, like is done by Villanueva-Domingo and Villaescusa-Navarro [72], before asked to clean real data.

- Training on polarized foregrounds: We additionally probed a failure mode of the network in the presence of 1% galactic synchrotron radiation polarization leakage. Here the PCA preprocessing fails to separate the leaked and cosmological signals, making it harder for `deep21` to pick out the cosmological signal. A follow-up study might additionally train on more realistic polarized simulations, such as those produced by `Hammurabi`, as well as bypass the need for a blind preprocessing step.

- Increasing input sizes: `deep21`'s input voxel sizes were limited by available GPU memory. With improved deep learning computational resources (or a detailed tiling strategy such as the one employed by He et al. [28] for N-Body analyses), Increasing input size will likely improve foreground removal on large scales, since the network will have access to a larger context of information. Ideally, entire maps would only be split into a handful of UNet input units. An increase in input volume might also allow for a larger frequency range to be assessed, which would aid the network in distinguishing cosmological and polarized signals, as well as improve radial power spectrum recovery.

Code Availability

The code used for training and generation of results is publicly available at `https://github.com/tlmakinen/deep21`. A browser-based tutorial for the experiment and UNet module is available via the accompanying Colab notebook.

6 Acknowledgements

Many thanks to Nick Carriero and the Flatiron Institute's HPC support team, without whom this work would not be possible. Thank you also to David Alonso for simulation guidance and to Ben Wandelt for helpful discussions. FVN acknowledges funding from the WFIRST program through NNG26PJ30C and NNN12AA01c. The work of SH and DNS has been supported by the Simons Foundation. TLM completed a large portion of this work to satisfy requirements for the Degree of Bachelor of Arts at Princeton University.

References

[1] F. B. Abdalla and S. Rawlings. Probing dark energy with baryonic oscillations and future radio surveys of neutral hydrogen. *Monthly Notices of the Royal Astronomical Society*, 360(1):27–40, Jun 2005. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2005.08650.x. URL http://dx.doi.org/10.1111/j.1365-2966.2005.08650.x.

[2] David Alonso, Philip Bull, Pedro G. Ferreira, and Mário G. Santos. Blind foreground subtraction for intensity mapping experiments. *Monthly Notices of the Royal Astronomical Society*, 447(1):400–416, Dec 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu2474. URL http://dx.doi.org/10.1093/mnras/stu2474.

[3] David Alonso, Pedro G. Ferreira, and Mario G. Santos. Fast simulations for intensity mapping experiments. *Monthly Notices of the Royal Astronomical Society*, 444(4):3183–3197, Sep 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu1666. URL http://dx.doi.org/10.1093/mnras/stu1666.

[4] David J. Bacon, Richard A. Battye, Philip Bull, Stefano Camera, Pedro G. Ferreira, Ian Harrison, David Parkinson, Alkistis Pourtsidou, Mário G. Santos, and et al. Cosmology with phase 1 of the square kilometre array red book 2018: Technical specifications and performance forecasts. *Publications of the Astronomical Society of Australia*, 37, 2020. ISSN 1448-6083. doi: 10.1017/pasa.2019.51. URL http://dx.doi.org/10.1017/pasa.2019.51.

[5] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David Cox. Hyperopt: A python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8:014008, 07 2015. doi: 10.1088/1749-4699/8/1/014008.

[6] G. Bernardi, A. G. de Bruyn, M. A. Brentjens, B. Ciardi, G. Harker, V. Jelić, L. V. E. Koopmans, P. Labropoulos, A. Offringa, V. N. Pandey, and et al. Foregrounds for observations of the cosmological 21 cm line. *Astronomy & Astrophysics*, 500(3):965–979, Jun 2009. ISSN 1432-0746. doi: 10.1051/0004-6361/200911627. URL http://dx.doi.org/10.1051/0004-6361/200911627.

[7] G. Bernardi, A. G. de Bruyn, G. Harker, M. A. Brentjens, B. Ciardi, V. Jelić, L. V. E. Koopmans, P. Labropoulos, A. Offringa, V. N. Pandey, and et al. Foregrounds for observations of the cosmological 21 cm line. *Astronomy & Astrophysics*, 522:A67, Nov 2010. ISSN 1432-0746. doi: 10.1051/0004-6361/200913420. URL http://dx.doi.org/10.1051/0004-6361/200913420.

[8] Philip Bull, Pedro G. Ferreira, Prina Patel, and Mário G. Santos. Late-time Cosmology with 21 cm Intensity Mapping Experiments. ApJ, 803(1):21, April 2015. doi: 10.1088/0004-637X/803/1/21.

[9] Stefano Camera, Mário G. Santos, Pedro G. Ferreira, and Luís Ferramacho. Cosmology on Ultralarge Scales with Intensity Mapping of the Neutral Hydrogen 21 cm Emission: Limits on Primordial Non-Gaussianity. Phys. Rev. Lett., 111(17):171302, October 2013. doi: 10.1103/PhysRevLett.111.171302.

[10] Tzu-Ching Chang, Ue-Li Pen, Kevin Bandura, and Jeffrey B. Peterson. Hydrogen 21-cm Intensity Mapping at redshift 0.8. *arXiv e-prints*, art. arXiv:1007.3709, July 2010.

[11] Jonathan Chardin, Grégoire Uhlrich, Dominique Aubert, Nicolas Deparis, Nicolas Gillet, Pierre Ocvirk, and Joseph Lewis. A deep learning model to emulate simulations of cosmic reionization. *Monthly Notices of the Royal Astronomical Society*, 490(1):1055–1065, Sep 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz2605. URL http://dx.doi.org/10.1093/mnras/stz2605.

[12] Tom Charnock, Laurence Perreault-Levasseur, and François Lanusse. Bayesian neural networks, 2020.

[13] A. S. Cohen, H. J. A. Röttgering, M. J. Jarvis, N. E. Kassim, and T. J. W. Lazio. A Deep, High-Resolution Survey at 74 MHz. ApJS, 150(2):417–430, February 2004. doi: 10.1086/380783.

[14] Steven Cunnington, Laura Wolz, Alkistis Pourtsidou, and David Bacon. Impact of foregrounds on h i intensity mapping cross-correlations with optical surveys. *Monthly Notices of the Royal Astronomical Society*, 488(4):5452–5472, Jul 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz1916. URL http://dx.doi.org/10.1093/mnras/stz1916.

[15] A. de Bruyn, P. Katgert, Marijke Haverkorn, and D. Schnitzeler. Radio polarization and rm structure at high galactic latitudes. *Astronomische Nachrichten*, 327:487 – 490, 06 2006. doi: 10.1002/asna.200610566.

[16] Angélica de Oliveira-Costa, Max Tegmark, B. M. Gaensler, Justin Jonas, T. L. Landecker, and Patricia Reich. A model of diffuse Galactic radio emission from 10 MHz to 100 GHz. MNRAS, 388(1):247–260, July 2008. doi: 10.1111/j.1365-2966.2008.13376.x.

[17] George B. Field. Excitation of the Hydrogen 21-CM Line. *Proceedings of the IRE*, 46:240–250, January 1958. doi: 10.1109/JRPROC.1958.286741.

[18] George B. Field. The Spin Temperature of Intergalactic Neutral Hydrogen. ApJ, 129:536, May 1959. doi: 10.1086/146653.

[19] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep Ensembles: A Loss Landscape Perspective. *arXiv e-prints*, art. arXiv:1912.02757, December 2019.

[20] Steven R. Furlanetto, S. Peng Oh, and Frank H. Briggs. Cosmology at low frequencies: The 21 cm transition and the high-redshift Universe. Phys. Rep., 433(4-6):181–301, October 2006. doi: 10.1016/j.physrep.2006.08.002.

[21] Nicolas Gillet, Andrei Mesinger, Bradley Greig, Adrian Liu, and Graziano Ucci. Deep learning from 21-cm tomography of the cosmic dawn and reionization. *Monthly Notices of the Royal Astronomical Society*, Jan 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz010. URL http://dx.doi.org/10.1093/mnras/stz010.

[22] Liron Gleser, Adi Nusser, and Andrew J. Benson. Decontamination of cosmological 21-cm maps. *Monthly Notices of the Royal Astronomical Society*, 391(1):383–398, Nov 2008. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2008.13897.x. URL http://dx.doi.org/10.1111/j.1365-2966.2008.13897.x.

[23] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*, 2010.

[24] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[25] K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. ApJ, 622:759–771, April 2005. doi: 10.1086/427976.

[26] Alex Hall, Camille Bonvin, and Anthony Challinor. Testing general relativity with 21-cm intensity mapping. Phys. Rev. D, 87(6):064026, March 2013. doi: 10.1103/PhysRevD.87.064026.

[27] C. G. T. Haslam, C. J. Salter, H. Stoffel, and W. E. Wilson. A 408 MHz all-sky continuum survey. II. The atlas of contour maps. A&AS, 47:1–143, January 1982.

[28] Siyu He, Yin Li, Yu Feng, Shirley Ho, Siamak Ravanbakhsh, Wei Chen, and Barnabás Póczos. Learning to predict the cosmological structure formation. *Proceedings of the National Academy of Sciences*, 116(28):13825–13832, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1821458116. URL https://www.pnas.org/content/116/28/13825.

[29] V. Jelić, S. Zaroubi, P. Labropoulos, R. M. Thomas, G. Bernardi, M. A. Brentjens, A. G. de Bruyn, B. Ciardi, G. Harker, L. V. E. Koopmans, and et al. Foreground simulations for the lofar-epoch of reionization experiment. *Monthly Notices of the Royal Astronomical Society*, 389 (3):1319–1335, Sep 2008. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2008.13634.x. URL http://dx.doi.org/10.1111/j.1365-2966.2008.13634.x.

[30] Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A Probabilistic U-Net for Segmentation of Ambiguous Images. *arXiv e-prints*, art. arXiv:1806.05034, June 2018.

[31] Yungi Kwon, Sungwook E. Hong, and Inkyu Park. Deep-learning study of the 21-cm differential brightness temperature during the epoch of reionization. *Journal of the Korean Physical Society*, 77(1):49–59, Jul 2020. ISSN 1976-8524. doi: 10.3938/jkps.77.49. URL http://dx.doi.org/10.3938/jkps.77.49.

[32] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc., 2017.

[33] Weitian Li, Haiguang Xu, Zhixian Ma, Ruimin Zhu, Dan Hu, Zhenghao Zhu, Junhua Gu, Chenxi Shan, Jie Zhu, and Xiang-Ping Wu. Separating the eor signal with a convolutional denoising autoencoder: a deep-learning-based method. *Monthly Notices of the Royal Astronomical Society*, 485(2):2628–2637, Feb 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz582. URL http://dx.doi.org/10.1093/mnras/stz582.

[34] Adrian Liu and J. Richard Shaw. Data Analysis for Precision 21 cm Cosmology. PASP, 132 (1012):062001, June 2020. doi: 10.1088/1538-3873/ab5bfd.

[35] Adrian Liu and Max Tegmark. A method for 21 cm power spectrum estimation in the presence of foregrounds. *Physical Review D*, 83(10), May 2011. ISSN 1550-2368. doi: 10.1103/physrevd.83.103006. URL http://dx.doi.org/10.1103/PhysRevD.83.103006.

[36] Adrian Liu and Max Tegmark. How well can we measure and understand foregrounds with 21-cm experiments? *Monthly Notices of the Royal Astronomical Society*, 419(4):3491–3504, 01 2012. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2011.19989.x. URL https://doi.org/10.1111/j.1365-2966.2011.19989.x.

[37] Adrian Liu, Max Tegmark, Judd Bowman, Jacqueline Hewitt, and Matias Zaldarriaga. An improved method for 21-cm foreground removal. *Monthly Notices of the Royal Astronomical Society*, 398(1):401–406, Sep 2009. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2009.15156.x. URL http://dx.doi.org/10.1111/j.1365-2966.2009.15156.x.

[38] Piero Madau, Avery Meiksin, and Martin J. Rees. 21 Centimeter Tomography of the Intergalactic Medium at High Redshift. ApJ, 475(2):429–444, February 1997. doi: 10.1086/303549.

[39] Tumelo Mangena, Sultan Hassan, and Mario G. Santos. Constraining the reionization history using deep learning from 21cm tomography with the Square Kilometre Array. MNRAS, March 2020. doi: 10.1093/mnras/staa750.

[40] Tumelo Mangena, Sultan Hassan, and Mario G Santos. Constraining the reionization history using deep learning from 21-cm tomography with the square kilometre array. *Monthly Notices of the Royal Astronomical Society*, 494(1):600–606, Mar 2020. ISSN 1365-2966. doi: 10.1093/mnras/staa750. URL http://dx.doi.org/10.1093/mnras/staa750.

[41] Tumelo Mangena, Sultan Hassan, and Mario G Santos. Constraining the reionization history using deep learning from 21-cm tomography with the square kilometre array. *Monthly Notices of the Royal Astronomical Society*, 494(1):600–606, Mar 2020. ISSN 1365-2966. doi: 10.1093/mnras/staa750. URL http://dx.doi.org/10.1093/mnras/staa750.

[42] K. W. Masui, E. R. Switzer, N. Banavar, K. Bandura, C. Blake, L.-M. Calin, T.-C. Chang, X. Chen, Y.-C. Li, Y.-W. Liao, and et al. Measurement of 21 cm brightness fluctuations atvz 0.8 in cross-correlation. *The Astrophysical Journal*, 763(1):L20, Jan 2013. ISSN 2041-8213. doi: 10.1088/2041-8205/763/1/l20. URL http://dx.doi.org/10.1088/2041-8205/763/1/L20.

[43] Tiziana Di Matteo, Rosalba Perna, Tom Abel, and Martin J. Rees. Radio foregrounds for the 21 centimeter tomography of the neutral intergalactic medium at high redshifts. *The Astrophysical Journal*, 564(2):576–580, jan 2002. doi: 10.1086/324293. URL https://doi.org/10.1086%2F324293.

[44] Hrushikesh Mhaskar and Tomaso Poggio. Deep vs. shallow networks : An approximation theory perspective, 2016.

[45] Raul A. Monsalve, Alan E. E. Rogers, Judd D. Bowman, and Thomas J. Mozdzen. Results from EDGES High-band. I. Constraints on Phenomenological Models for the Global 21 cm Signal. ApJ, 847(1):64, September 2017. doi: 10.3847/1538-4357/aa88d1.

[46] Raul A. Monsalve, Bradley Greig, Judd D. Bowman, Andrei Mesinger, Alan E. E. Rogers, Thomas J. Mozdzen, Nicholas S. Kern, and Nivedita Mahesh. Results from EDGES High-band. II. Constraints on Parameters of Early Galaxies. ApJ, 863(1):11, August 2018. doi: 10.3847/1538-4357/aace54.

[47] David F. Moore, James E. Aguirre, Aaron R. Parsons, Daniel C. Jacobs, and Jonathan C. Pober. The effects of polarized foregrounds on 21 cm epoch of reionization power spectrum measurements. *The Astrophysical Journal*, 769(2):154, May 2013. ISSN 1538-4357. doi: 10.1088/0004-637x/769/2/154. URL http://dx.doi.org/10.1088/0004-637X/769/2/154.

[48] Miguel F. Morales and J. Stuart B. Wyithe. Reionization and Cosmology with 21-cm Fluctuations. ARA&A, 48:127–171, September 2010. doi: 10.1146/annurev-astro-081309-130936.

[49] Miguel F. Morales, Judd D. Bowman, and Jacqueline N. Hewitt. Improving foreground subtraction in statistical observations of 21 cm emission from the epoch of reionization. *The Astrophysical Journal*, 648(2):767–773, Sep 2006. ISSN 1538-4357. doi: 10.1086/506135. URL http://dx.doi.org/10.1086/506135.

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[51] S. Peng Oh and Katherine J. Mack. Foregrounds for 21-cm observations of neutral gas at high redshift. *Monthly Notices of the Royal Astronomical Society*, 346(3):871–877, Dec 2003. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2003.07133.x. URL http://dx.doi.org/10.1111/j.1365-2966.2003.07133.x.

[52] Planck Collaboration, R. Adam, P. A. R. Ade, N. Aghanim, Y. Akrami, M. I. R. Alves, F. Argüeso, M. Arnaud, F. Arroja, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Band ay, R. B. Barreiro, J. G. Bartlett, N. Bartolo, S. Basak, P. Battaglia, E. Battaner, R. Battye, K. Benabed, A. Benoît, A. Benoit-Lévy, J. P. Bernard, M. Bersanelli, B. Bertincourt, P. Bielewicz, I. Bikmaev, J. J. Bock, H. Böhringer, A. Bonaldi, L. Bonavera, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, R. Burenin, C. Burigana, R. C. Butler, E. Calabrese, J. F. Cardoso, P. Carvalho, B. Casaponsa, G. Castex, A. Catalano, A. Challinor, A. Chamballu, R. R. Chary, H. C. Chiang, J. Chluba, G. Chon, P. R. Christensen, S. Church, M. Clemens, D. L. Clements, S. Colombi, L. P. L. Colombo, C. Combet, B. Comis, D. Contreras, F. Couchot, A. Coulais, B. P. Crill, M. Cruz, A. Curto, F. Cuttaia, L. Danese, R. D. Davies, R. J. Davis, P. de Bernardis, A. de Rosa, G. de Zotti, J. Delabrouille, J. M. Delouis, F. X. Désert, E. Di Valentino, C. Dickinson, J. M. Diego, K. Dolag, H. Dole, S. Donzelli, O. Doré, M. Douspis, A. Ducout, J. Dunkley, X. Dupac, G. Efstathiou, P. R. M. Eisenhardt, F. Elsner, T. A. Enßlin, H. K. Eriksen, E. Falgarone, Y. Fantaye, M. Farhang, S. Feeney, J. Fergusson, R. Fernandez-Cobos, F. Feroz, F. Finelli, E. Florido, O. Forni, M. Frailis, A. A. Fraisse, C. Franceschet, E. Franceschi, A. Frejsel, A. Frolov, S. Galeotta, S. Galli, K. Ganga, C. Gauthier, R. T. Génova-Santos, M. Gerbino, T. Ghosh, M. Giard, Y. Giraud-Héraud, E. Giusarma, E. Gjerløw, J. González-Nuevo, K. M. Górski, K. J. B. Grainge, S. Gratton, A. Gregorio, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, D. Hanson, D. L. Harrison, A. Heavens, G. Helou, S. Henrot-Versillé, C. Hernández-Monteagudo, D. Herranz, S. R. Hildebrandt, E. Hivon, M. Hobson, W. A. Holmes, A. Hornstrup, W. Hovest, Z. Huang, K. M. Huffenberger, G. Hurier, S. Ilić, A. H. Jaffe, T. R. Jaffe, T. Jin, W. C. Jones, M. Juvela, A. Karakci, E. Keihänen, R. Keskitalo, I. Khamitov, K. Kiiveri, J. Kim, T. S. Kisner, R. Kneissl, J. Knoche, L. Knox, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, F. Lacasa, G. Lagache, A. Lähteenmäki, J. M. Lamarre, M. Langer, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, J. P. Leahy, E. Lellouch, R. Leonardi, J. León-Tavares, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Lilley, M. Linden-Vørnle, V. Lindholm, H. Liu, M. López-Caniego, P. M. Lubin, Y. Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, D. S. Y. Mak, N. Mandolesi, A. Mangilli, A. Marchini, A. Marcos-Caballero, D. Marinucci, M. Maris, D. J. Marshall, P. G. Martin, M. Martinelli, E. Martínez-González, S. Masi, S. Matarrese, P. Mazzotta, J. D. McEwen, P. McGehee, S. Mei, P. R. Meinhold, A. Melchiorri, J. B. Melin, L. Mendes, A. Mennella, M. Migliaccio, K. Mikkelsen, M. Millea, S. Mitra, M. A. Miville-Deschênes, D. Molinari, A. Moneti, L. Montier, R. Moreno, G. Morgante, D. Mortlock, A. Moss, S. Mottet, M. Münchmeyer, D. Munshi, J. A. Murphy, A. Narimani, P. Naselsky, A. Nastasi, F. Nati, P. Natoli, M. Negrello, C. B. Netterfield, H. U. Nørgaard-Nielsen, F. Noviello, D. Novikov, I. Novikov, M. Olamaie, N. Oppermann, E. Orlando, C. A. Oxborrow, F. Paci, L. Pagano, F. Pajot, R. Paladini, S. Pandolfi, D. Paoletti, B. Partridge, F. Pasian, G. Patanchon, T. J. Pearson, M. Peel, H. V. Peiris, V. M. Pelkonen, O. Perdereau, L. Perotto, Y. C. Perrott, F. Perrotta, V. Pettorino, F. Piacentini, M. Piat, E. Pierpaoli, D. Pietrobon, S. Plaszczynski, D. Pogosyan, E. Pointecouteau, G. Polenta, L. Popa, G. W. Pratt, G. Prézeau, S. Prunet, J. L. Puget, J. P. Rachen, B. Racine, W. T. Reach, R. Rebolo, M. Reinecke, M. Remazeilles, C. Renault, A. Renzi, I. Ristorcelli, G. Rocha, M. Roman, E. Romelli, C. Rosset, M. Rossetti, A. Rotti, G. Roudier, B. Rouillé d'Orfeuil, M. Rowan-Robinson, J. A. Rubiño-Martín, B. Ruiz-Granados, C. Rumsey, B. Rusholme, N. Said, V. Salvatelli, L. Salvati, M. Sandri, H. S. Sanghera, D. Santos, R. D. E. Saunders, A. Sauvé, M. Savelainen, G. Savini, B. M. Schaefer, M. P. Schammel, D. Scott, M. D. Seiffert, P. Serra, E. P. S. Shellard, T. W. Shimwell, M. Shiraishi, K. Smith, T. Souradeep, L. D. Spencer, M. Spinelli, S. A. Stanford, D. Stern, V. Stolyarov, R. Stompor, A. W. Strong, R. Sudiwala, R. Sunyaev, P. Sutter, D. Sutton, A. S. Suur-Uski, J. F. Sygnet, J. A. Tauber, D. Tavagnacco, L. Terenzi, D. Texier, L. Toffolatti, M. Tomasi, M. Tornikoski, D. Tramonte, M. Tristram, A. Troja, T. Trombetti, M. Tucci, J. Tuovinen, M. Türler, G. Umana, L. Valenziano, J. Valiviita, F. Van Tent, T. Vassallo, L. Vibert, M. Vidal, M. Viel, P. Vielva, F. Villa, L. A. Wade, B. Walter, B. D. Wand elt,

R. Watson, I. K. Wehus, N. Welikala, J. Weller, M. White, S. D. M. White, A. Wilkinson, D. Yvon, A. Zacchei, J. P. Zibin, and A. Zonca. Planck 2015 results. I. Overview of products and scientific results. A&A, 594:A1, September 2016. doi: 10.1051/0004-6361/201527101.

[53] Jonathan R Pritchard and Abraham Loeb. 21 cm cosmology in the 21st century. *Reports on Progress in Physics*, 75(8):086901, Jul 2012. ISSN 1361-6633. doi: 10.1088/0034-4885/75/8/086901. URL http://dx.doi.org/10.1088/0034-4885/75/8/086901.

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

[55] George B. Rybicki and Alan P. Lightman. *Radiative Processes in Astrophysics*. 1986.

[56] Mário G. Santos, Asantha Cooray, and Lloyd Knox. Multifrequency Analysis of 21 Centimeter Fluctuations from the Era of Reionization. ApJ, 625(2):575–587, June 2005. doi: 10.1086/429857.

[57] D. H. F. M. Schnitzeler, P. Katgert, and A. G. de Bruyn. WSRT Faraday tomography of the Galactic ISM at λ ~0.86 m. I. The GEMINI data set at (l, b) = (181°, 20°). A&A, 494(2): 611–622, February 2009. doi: 10.1051/0004-6361:20078912.

[58] D. Scott and M. J. Rees. The 21-cm line at high redshift: a diagnostic for the origin of large scale structure. MNRAS, 247:510, December 1990.

[59] Toyokazu Sekiguchi, Tomo Takahashi, Hiroyuki Tashiro, and Shuichiro Yokoyama. Probing primordial non-gaussianity with 21 cm fluctuations from minihalos. *Journal of Cosmology and Astroparticle Physics*, 2019(02):033–033, Feb 2019. ISSN 1475-7516. doi: 10.1088/1475-7516/2019/02/033. URL http://dx.doi.org/10.1088/1475-7516/2019/02/033.

[60] J. Richard Shaw, Kris Sigurdson, Ue-Li Pen, Albert Stebbins, and Michael Sitwell. All-sky interferometry with spherical harmonic transit telescopes. *The Astrophysical Journal*, 781(2): 57, Jan 2014. ISSN 1538-4357. doi: 10.1088/0004-637x/781/2/57. URL http://dx.doi.org/10.1088/0004-637X/781/2/57.

[61] J. Richard Shaw, Kris Sigurdson, Michael Sitwell, Albert Stebbins, and Ue-Li Pen. Coaxing cosmic 21 cm fluctuations from the polarized sky usingm-mode analysis. *Physical Review D*, 91 (8), Apr 2015. ISSN 1550-2368. doi: 10.1103/physrevd.91.083514. URL http://dx.doi.org/10.1103/PhysRevD.91.083514.

[62] J. Richard Shaw, Kris Sigurdson, Michael Sitwell, Albert Stebbins, and Ue-Li Pen. Coaxing cosmic 21 cm fluctuations from the polarized sky usingm-mode analysis. *Physical Review D*, 91 (8), Apr 2015. ISSN 1550-2368. doi: 10.1103/physrevd.91.083514. URL http://dx.doi.org/10.1103/PhysRevD.91.083514.

[63] Anze Slosar, Zeeshan Ahmed, David Alonso, Mustafa A. Amin, Evan J. Arena, Kevin Bandura, Nicholas Battaglia, Jonathan Blazek, Philip Bull, Emanuele Castorina, Tzu-Ching Chang, Liam Connor, Romeel Davé, Cora Dvorkin, Alexander van Engelen, Simone Ferraro, Raphael Flauger, Simon Foreman, Josef Frisch, Daniel Green, Gilbert Holder, Daniel Jacobs, Matthew C. Johnson, Joshua S. Dillon, Dionysios Karagiannis, Alexander A. Kaurov, Lloyd Knox, Adrian Liu, Marilena Loverde, Yin-Zhe Ma, Kiyoshi W. Masui, Thomas McClintock, Kavilan Moodley, Moritz Munchmeyer, Laura B. Newburgh, Cherry Ng, Andrei Nomerotski, Paul O'Connor, Andrej Obuljen, Hamsa Padmanabhan, David Parkinson, J. Xavier Prochaska, Surjeet Rajendran, David Rapetti, Benjamin Saliwanchik, Emmanuel Schaan, Neelima Sehgal, J. Richard Shaw, Chris Sheehy, Erin Sheldon, Raphael Shirley, Eva Silverstein, Tracy Slatyer, Anze Slosar, Paul Stankus, Albert Stebbins, Peter T. Timbie, Gregory S. Tucker, William Tyndall, Francisco Villaescusa Navarro, Benjamin Wallisch, and Martin White. Packed

Ultra-wideband Mapping Array (PUMA): A Radio Telescope for Cosmology and Transients. In *Bulletin of the American Astronomical Society*, volume 51, page 53, September 2019.

[64] Marta Spinelli, Gianni Bernardi, and Mario G Santos. On the contamination of the global 21 cm signal from polarized foregrounds. *Monthly Notices of the Royal Astronomical Society*, Sep 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz2425. URL http://dx.doi.org/10.1093/mnras/stz2425.

[65] E. R. Switzer, T.-C. Chang, K. W. Masui, U.-L. Pen, and T. C. Voytek. INTERPRETING THE UNRESOLVED INTENSITY OF COSMOLOGICALLY REDSHIFTED LINE RADIATION. *The Astrophysical Journal*, 815(1):51, dec 2015. doi: 10.1088/0004-637x/815/1/51. URL https://doi.org/10.1088%2F0004-637x%2F815%2F1%2F51.

[66] Max Tegmark, Daniel J. Eisenstein, Wayne Hu, and Angelica de Oliveira-Costa. Foregrounds and forecasts for the cosmic microwave background. *The Astrophysical Journal*, 530(1): 133–165, feb 2000. doi: 10.1086/308348. URL https://doi.org/10.1086%2F308348.

[67] Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622, 1999. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/2680726.

[68] Paolo Tozzi, Piero Madau, Avery Meiksin, and Martin J. Rees. Radio Signatures of H I at High Redshift: Mapping the End of the "Dark Ages". ApJ, 528(2):597–606, January 2000. doi: 10.1086/308196.

[69] Francisco Villaescusa-Navarro, Matteo Viel, David Alonso, Kanan K. Datta, Philip Bull, and Mario G. Santos. Cross-correlating 21cm intensity maps with lyman break galaxies in the post-reionization era, 2014.

[70] Francisco Villaescusa-Navarro, David Alonso, and Matteo Viel. Baryonic acoustic oscillations from 21 cm intensity mapping: the Square Kilometre Array case. MNRAS, 466(3):2736–2751, April 2017. doi: 10.1093/mnras/stw3224.

[71] Francisco Villaescusa-Navarro, Shy Genel, Emanuele Castorina, Andrej Obuljen, David N. Spergel, Lars Hernquist, Dylan Nelson, Isabella P. Carucci, Annalisa Pillepich, Federico Marinacci, Benedikt Diemer, Mark Vogelsberger, Rainer Weinberger, and Rüdiger Pakmor. Ingredients for 21 cm Intensity Mapping. ApJ, 866(2):135, October 2018. doi: 10.3847/1538-4357/aadba0.

[72] Pablo Villanueva-Domingo and Francisco Villaescusa-Navarro. Removing Astrophysics in 21 cm maps with Neural Networks. *arXiv e-prints*, art. arXiv:2006.14305, June 2020.

[73] Digvijay Wadekar, Francisco Villaescusa-Navarro, Shirley Ho, and Laurence Perreault-Levasseur. HInet: Generating neutral hydrogen from dark matter with neural networks. *arXiv e-prints*, art. arXiv:2007.10340, July 2020.

[74] A. Waelkens, T. Jaffe, M. Reinecke, F. S. Kitaura, and T. A. Enßlin. Simulating polarized Galactic synchrotron emission at all frequencies. The Hammurabi code. A&A, 495(2):697–706, February 2009. doi: 10.1051/0004-6361:200810564.

[75] Xiaomin Wang, Max Tegmark, Mário G. Santos, and Lloyd Knox. 21 cm Tomography with Foregrounds. ApJ, 650(2):529–537, October 2006. doi: 10.1086/506597.

[76] A. Weltman, P. Bull, S. Camera, K. Kelley, H. Padmanabhan, J. Pritchard, A. Raccanelli, S. Riemer-Sørensen, L. Shao, S. Andrianomena, and et al. Fundamental physics with the square kilometre array. *Publications of the Astronomical Society of Australia*, 37, 2020. ISSN 1448-6083. doi: 10.1017/pasa.2019.42. URL http://dx.doi.org/10.1017/pasa.2019.42.

[77] T. L. Wilson. Techniques of radio astronomy, 2011.

[78] M. Wolleben, T. L. Landecker, W. Reich, and R. Wielebinski. An absolutely calibrated survey of polarized emission from the northern sky at 1.4 GHz. Observations and data reduction. A&A, 448(1):411–424, March 2006. doi: 10.1051/0004-6361:20053851.

[79] L. Wolz, F. B. Abdalla, C. Blake, J. R. Shaw, E. Chapman, and S. Rawlings. The effect of foreground subtraction on cosmological measurements from intensity mapping. *Monthly Notices of the Royal Astronomical Society*, 441(4):3271–3283, May 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu792. URL http://dx.doi.org/10.1093/mnras/stu792.

[80] Jian Yao. *International Symposium on Cosmology and Ali CMB Polarization Telescope*. Sep 2018. URL https://indico.leeinst.sjtu.edu.cn/event/44/attachments/128/345/Foreground_removal_GAN.pdf.

[81] Le Zhang, Emory F. Bunn, Ata Karakci, Andrei Korotkov, P. M. Sutter, Peter T. Timbie, Gregory S. Tucker, and Benjamin D. Wandelt. Bayesian Semi-blind Component Separation for Foreground Removal in Interferometric 21 cm Observations. ApJS, 222(1):3, January 2016. doi: 10.3847/0067-0049/222/1/3.

[82] Andrea Zonca, Leo Singer, Daniel Lenz, Martin Reinecke, Cyrille Rosset, Eric Hivon, and Krzysztof Gorski. healpy: equal area pixelization and spherical harmonics transforms for data on the sphere in python. *Journal of Open Source Software*, 4(35):1298, March 2019. doi: 10.21105/joss.01298. URL https://doi.org/10.21105/joss.01298.

A Hyperparameter choices

Symbol	Prior Distribution	Optimum	Description
`conv ND`	[2,3]	3	convolution filter dimensions
`h`	disc $\mathcal{U}(2,6)$	6	no. of down-convolutions
`w`	disc $\mathcal{U}(1,6)$	3	no. of convolutions for each conv. block
`batchnorm`	[0,1]	`True`*	batch normalization for given layer
`batch size`	$\log\mathcal{U}(4,24)$	48	no. of samples per gradient descent step
$\mathtt{n}_{\mathrm{filters}}$	[8,16,32]	32	initial number of conv. filters
β_{mom}	$\log\mathcal{U}(0.001,0.75)$	0.05	batch normalization momentum
λ	$\log\mathcal{N}(-8.5,2.5)$	0.0002	learning rate for `Adam` optimizer

Table 4: Table of UNet parameters varied in architecture design, with optimal values (per GPU processor) shown. The algorithm `HyperOpt` was used in conjunction with the `Adam` optimizer tuning according to [5]. The discrete uniform distribution is denoted by disc $\mathcal{U}(\cdot)$. *Batch normalization adopted for encoder layers.