Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **494,** 5761–5774 (2020)
Advance Access publication 2020 May 4

doi:10.1093/mnras/staa1151

Constraining the astrophysics and cosmology from 21 cm tomography using deep learning with the SKA

Sultan Hassan,[1,2★†] Sambatra Andrianomena[2,3] and Caitlin Doughty[1]

[1] *Department of Astronomy, New Mexico State University, Las Cruces, NM 88003, USA*
[2] *Department of Physics and Astronomy, University of the Western Cape, Bellville, Cape Town 7535, South Africa*
[3] *South African Radio Astronomy Observatory (SARAO), Black River Park, Observatory, Cape Town 7925, South Africa*

Accepted 2020 April 20. Received 2020 March 17; in original form 2019 July 17

ABSTRACT

Future Square Kilometre Array (SKA) surveys are expected to generate huge data sets of 21 cm maps on cosmological scales from the Epoch of Reionization. We assess the viability of exploiting machine learning techniques, namely, convolutional neural networks (CNNs), to simultaneously estimate the astrophysical and cosmological parameters from 21 cm maps from seminumerical simulations. We further convert the simulated 21 cm maps into SKA-like mock maps using the detailed SKA antennae distribution, thermal noise, and a recipe for foreground cleaning. We successfully design two CNN architectures (VGGNet-like and ResNet-like) that are both efficiently able to extract simultaneously three astrophysical parameters, namely the photon escape fraction ($f_{\rm esc}$), the ionizing emissivity power dependence on halo mass ($C_{\rm ion}$), and the ionizing emissivity redshift evolution index ($D_{\rm ion}$), and three cosmological parameters, namely the matter density parameter (Ω_m), the dimensionless Hubble constant (h), and the matter fluctuation amplitude (σ_8), from 21 cm maps at several redshifts. With the presence of noise from SKA, our designed CNNs are still able to recover these astrophysical and cosmological parameters with great accuracy ($R^2 > 92$ per cent), improving to $R^2 >$ 99 per cent towards low-redshift and low neutral fraction values. Our results show that future 21 cm observations can play a key role to break degeneracy between models and tightly constrain the astrophysical and cosmological parameters, using only few frequency channels.

Key words: methods: statistical – galaxies: high-redshift – intergalactic medium – cosmological parameters – dark ages, reionization, first stars.

1 INTRODUCTION

The last global phase transition in the Universe, known as the Epoch of Reionization (EoR), marks the time at which the first stars gradually reionized the intergalactic medium (IGM) and the Universe transitioned from highly neutral opaque to a highly ionized-transparent state (for a review, see e.g. Loeb & Barkana 2001). This epoch represents a crucial period in the Universe's history, particularly with regard to the formation and evolution of early galaxies.

Constraining the astrophysical and cosmological parameters has been the focus for most observational and theoretical studies. Several techniques have been developed to constrain the cosmological parameters (e.g. matter density parameter Ω_m and Hubble constant H_0) such as using the cosmic microwave background (CMB) anisotropies measurements (e.g. Hinshaw et al. 2013; Planck Collaboration XIII 2016), Sunyaev–Zel'dovich cluster surveys (e.g. Battye & Weller 2003), galaxy clusters in optical and X-ray bands (e.g. Moscardini, Matarrese & Mo 2001), gamma-ray burst X-ray afterglow light curves (e.g. Cardone et al. 2010), lensed GW+EM signals (e.g. Li, Fan & Gou 2019), Ly-α forest power spectrum and *COBE*-DMR (e.g. Phillips et al. 2001), large-scale clustering of SDSS luminous red galaxies (e.g. Padmanabhan et al. 2007), and a joint CMB and weak lensing analysis (e.g. Contaldi, Hoekstra & Lewis 2003). On the other hand, several works have attempted to constrain the astrophysical parameters (e.g. the photon escape fraction, $f_{\rm esc}$, and ionizing emissivity evolution, $\dot{N}_{\rm ion}$), using Ly-α forest measurements (e.g. Becker & Bolton 2013), Lyman continuum (LyC) radiation from local galaxies (e.g. Leitet et al. 2013), and inferred constraints by tuning different theoretical models to other measurements (e.g. Mitra, Choudhury & Ferrara 2015; Finlator et al. 2015).

While all these methods show different levels of success to place constraints on various parameters, tighter constraints are expected to come from the EoR through measurements of the 21 cm fluctuations on cosmological scales. With its strong dependence on the ionization and density fields, the 21 cm signal carries a

★ E-mail: shassan@nmsu.edu
† Tombaugh Fellow.

wealth of information that is important in order to understand early stages of galaxy formation and evolution. In this light, many radio interferometer experiments, such as the Low Frequency Array (LOFAR; van Haarlem et al. 2013), the Precision Array for Probing the Epoch of Reionization (Parsons et al. 2010), the Murchison Wide field Array (Bowman et al. 2013), the Giant Metrewave Radio Telescope (Paciga et al. 2011), the Hydrogen Epoch of Reionization Array (HERA; DeBoer et al. 2017), and Square Kilometer Array (SKA; Mellema et al. 2013) are devoted to detecting reionization in the near future. These growing observational efforts require equivalent efforts in both the theoretical and statistical sides, in order to prepare for extracting all possible information and constrain the cosmological and astrophysical parameters from future 21 cm surveys.

Several studies have already shown that combining the 21 cm power spectrum with Markov Chain Monte Carlo (MCMC) analysis is a powerful technique to obtain tighter constraints and break degeneracy between models (e.g. Greig & Mesinger 2015; Liu et al. 2016; Pober, Greig & Mesinger 2016; Hassan et al. 2017; Park et al. 2019). Besides the power spectrum, future 21 cm surveys, the SKA in particular, are also expected to generate huge imaging data sets for the 21 cm fluctuations on large scales that will contain more information than the power spectrum. Going beyond the power spectrum has been the target of many studies (e.g. Bharadwaj & Pandey 2005; Barkana & Loeb 2008; Watkinson & Pritchard 2015; Majumdar et al. 2018), in which more information can be obtained through investigating the non-Gaussian nature of the 21 cm signal using higher order statistics such as the bispectrum. To efficiently use the 21 cm information stored in the 2D 21 cm maps, convolutional neural networks (CNNs) have been a very successful deep learning tool to recover the astrophysical parameters during reionization (Gillet et al. 2019), to learn the reionization history (La Plante & Ntampaka 2019; Mangena, Hassan & Santos 2020), to emulate reionization simulations (Chardin et al. 2019), and to identify reionization sources from different models (Hassan et al. 2019). However, the astrophysical parameter recovery by Gillet et al. (2019) ignores the instrumental effects as an initial proof-of-concept study. Accounting for these effects such as the angular resolution, foreground cleaning, and thermal noise, are all crucial in order to add realism to the simulated 21cm images as we prepare for the 21 cm era.

In this work, we take a step further to design two different CNNs to simultaneously estimate several parameters from 21 cm maps at several redshifts and different stages through reionization. We here that assume all observations at different redshifts are performed independently. We simply take maps from different redshifts and apply the instrumental noise directly on each map assuming a single frequency channel of a size ∼ 50 kHz (i.e. simulation resolution). We finally combine the maps from different redshifts to create our training data sets. We note that learning from light-cones is beyond the scope of the current work. Our aim is to provide a network that is able to predict parameters without requiring the redshift nor neutral fraction as inputs, which is a more flexible design. Three astrophysical parameters are evaluated: the photon escape fraction ($f_{\rm esc}$), the ionizing emissivity power dependence on halo mass ($C_{\rm ion}$), and the redshift evolution index ($D_{\rm ion}$). Additionally, we estimate three cosmological parameters: the matter density parameter (Ω_m), the dimensionless Hubble constant (h), and the matter fluctuation amplitude (σ_8). To assess the ability of future 21 cm tomography to constrain these parameters, we follow the recipe presented in Hassan et al. (2019) to add a physically motivated and realistic 21 cm noise to large-scale 21 cm maps that are produced using our seminumerical model, SIMFAST21 (Santos et al. 2008, 2010). This paper is organized as follows: we first describe our suite of simulations of the 21 cm signal and noise in Section 2. We then present the two network designs in Section 3 and the training data set in Section 4. We present the main results in Section 5, and draw our concluding remarks in Section 6.

2 SIMULATIONS

2.1 Seminumerical model, SIMFAST21

We use the Instantaneous model of our seminumerical simulations SIMFAST21, that has been developed in Hassan et al. (2016), to improve over previous implementations of the ionizing source and sink populations in these seminumerical simulations. In addition, it has been recently shown that this model is in a relatively good agreement with predictions from our radiative transfer simulation (ARTIST; Molaro et al. 2019), particularly in terms of the morphology and power spectrum of the ionization and 21 cm fields. However, the reionization history can be quite different for the same photon escape fraction value. This arises from violation of photon conservation which is an intrinsic problem in the use of excursion set formalism (ESF) in seminumerical simulations (Zahn et al. 2007; Paranjape, Choudhury & Padmanabhan 2016; Hassan et al. 2017). As indicated by ARTIST, as a temporary solution all our photon escape fraction predictions can be adjusted by a factor of 20 per cent to account for the photon conservation problem. We here briefly describe the simulation ingredients, and defer to Santos et al. (2010) for the full details of the simulation algorithm, and to Hassan et al. (2016) for the Instantaneous model development.

The dark matter density is generated in the linear regime from a Gaussian distribution using a Monte Carlo approach. Evolving the density field to non-linear regime is performed through the Zel'dovich (1970) approximation. Halos are then generated using the ESF. Ionized regions are identified using a similar form of the ESF that is based on a direct comparison between the instantaneous rates of ionization $R_{\rm ion}$ and recombination $R_{\rm rec}$ in spherical regions of decreasing sizes as specified by the ESF. Regions are flagged as ionized if:

$$f_{\rm esc}\, R_{\rm ion} \geq R_{\rm rec}\,, \tag{1}$$

where $f_{\rm esc}$ is the escape fraction. The $R_{\rm rec}$ is obtained from a radiative transfer simulation (Finlator et al. 2015), in order to account for the clumping effects below our cell size. The $R_{\rm rec}$ is parametrized as a function of overdensity Δ and redshift z as follows:

$$\frac{R_{\rm rec}}{V} = 9.85 \times 10^{-24}(1+z)^{5.1}\left[\frac{(\Delta/1.76)^{0.82}}{1+(\Delta/1.76)^{0.82}}\right]^4\,, \tag{2}$$

where V refers to the cell volume. The $R_{\rm ion}$ parametrization is derived from a combination of the radiative transfer simulation (Finlator et al. 2015), and a larger hydrodynamic simulation (Davé et al. 2013) that both have been shown to reproduce wide range of observations, including low-z observations. The $R_{\rm ion}$ is parametrized as a function of halo mass $M_{\rm h}$ and redshift z as follows:

$$\frac{R_{\rm ion}}{M_h} = 1.1 \times 10^{40}\,(1+z)^{D_{\rm ion}}\left(\frac{M_h}{9.51 \times 10^7}\right)^{C_{\rm ion}} \times \exp\left(\frac{-9.51 \times 10^7}{M_h}\right)^{3.0}\,, \tag{3}$$

where the best-fitting values of the ionizing emissivity dependence on halo mass $C_{\rm ion}$ and redshift $D_{\rm ion}$ were found to be $C_{\rm ion} = 0.41$ and

Table 1. Summary of our assumed SKA array design.

Array design	866 compact core
Station diameter, D [m]	35
Station area, A [m^2]	$962.11\left(\frac{110}{\nu[\mathrm{MHz}]}\right)^2$
System temperature [K] ($T_{\rm sys} = T_{\rm sky} + T_{\rm rcvr}$)	$1.1\,T_{\rm sky} + 40$
Total observation time $t_{\rm int}$ [h]	1000
Frequency resolution $\Delta\nu$ [kHz]	48
Redshift	10, 9, 8 ,7
Frequency [MHz]	129 , 142, 158, 178
FWHM [arcmin]	1.37, 1.24, 1.12, 0.99
Beam angle θ [rad]	0.066, 0.06, 0.054, 0.048
Default wedge slope m, equation (4)	0.27, 0.23, 0.19, 0.15

$D_{\rm ion} = 2.28$, respectively. Later, we will change these parameters to generate the training and testing data sets. Note that equation (3) shows that $R_{\rm ion}$ scales as $M_{\rm h}^{1.41}$, which is consistent with the SFR$-M_h$ relation previously found by Finlator, Davé & Özel (2011). We defer to Hassan et al. (2016) for the full details on the derivation of the $R_{\rm ion}$ and $R_{\rm rec}$ fitting functions and their effects on several reionization key observables.

2.2 21 cm instrument simulation

We here describe the method used to account for various instrumental effects following the recipe developed in Hassan et al. (2019). We briefly review this method below and refer the interested readers to Hassan et al. (2019) for detailed information and complete steps of how we convert a 21 cm simulated map into a *mock map* according to the assumed array design. In this work, we restrict our analysis to the SKA proposed design and leave a more detailed comparison between different arrays, such as HERA and LOFAR, for future works. The instrumental noise is applied separately on each redshift assuming a single frequency channel corresponding to the map size ($\sim$ 50 kHz). We leave to future works learning from the light-cones by considering many frequency channels over a large bandwidth in the analysis.

The 21 cm Instrument simulation pipeline consists of three parts:

(i) *Foreground cleaning:* foreground contaminated modes of the signal lie inside the foreground wedge in the $k_\perp - k_\shortparallel$ plane. The foreground wedge slope (m) is given by:

$$m = \frac{D\,H_0\,E(z)\,\sin\theta}{c(1+z)}, \tag{4}$$

where H_0 is the Hubble parameter, c is the speed of light, $E(z) \equiv \sqrt{\Omega_m(1+z)^3 + \Omega_\Lambda}$, and θ is the beam angle. To clean foregrounds, we simply zero out all modes within the wedge, satisfying $k_\shortparallel < m\,k_\perp$. For the same experiment, the slope increases with redshift, which means more modes are removed at higher redshifts. We quote exact wedge slope values for the SKA at our redshifts of interest in Table 1. This the first step of the noise pipeline to clean foregrounds from the 3D co-eval cubes.

(ii) *Angular resolution:* we account for the angular resolution of a given array by exploiting its detailed baseline distribution, via the uv-coverage, which is a measure of the baseline intensity observing the signal modes in directions perpendicular to the sightline. The uv-coverage is computed using the 21CMSENSE package[1] from our assumed SKA antennae distribution. We then Fourier transform the simulated 21 cm map and set the signal to zero at $k_\perp$ modes whose uv-coverage is zero.[2] We additionally smooth down the simulated maps using a Gaussian filter whose full width half-maximum (FWHM) is given by: FWHM $= \lambda_{\rm 21cm}\,(1+z)/B$, whereas the maximum baseline length $B = 5834$ m for our assumed SKA design, and $\lambda_{\rm 21cm}$ is the rest-frame wavelength of the 21 cm signal. This sets the minimum angular resolution for our assumed SKA design. For instance, our simulated maps initially have an angular resolution of $\sim$ 0.3 arcmin at $z = 7$, that are smoothed to have a lower angular resolution of $\sim$ 1 arcmin according to the FWHM at this redshift. Exact angular resolution values as a function of redshift are quoted in Table 1. The angular resolution recipe is applied on maps extracted from the 3D foreground filtered boxes from the previous step.

(iii) *Thermal noise:* the thermal noise is uncorrelated between measurements, and can be drawn from a Gaussian distribution of unit mean and standard deviation (Zaldarriaga, Furlanetto & Hernquist 2004) given by:

$$\sqrt{\langle|N|^2\rangle}[\mathrm{Jy}] = \frac{2\,k_{\rm B}\,T_{\rm sys}}{A\,\sqrt{\Delta\nu\,t_{\rm int}}}\,, \tag{5}$$

where $t_{\rm int}$ here is the integration time to observe a single visibility at a frequency resolution $\Delta\nu$, and $k_{\rm B}$ is the Boltzmann constant. The total system temperature $T_{\rm sys}$ and other parameters are summarized in Table 1. Having generated the thermal noise in 2D grid using the above equation in Fourier space, we further suppress the noise by the amount of the uv-coverage N_{uv} by a factor of $\sim 1/\sqrt{N_{uv}}$. We finally inverse Fourier transform the noise map and add it to the angular resolution – foreground filtered signal map to form our mock 21 cm map.

Using this pipeline with parameters listed in Table 1, the rms brightness temperature (noise level) is about $\sim$ 3 mK at $z = 8$, consistent with previous estimates (e.g. see Furlanetto, Oh & Briggs 2006; Kakiichi et al. 2017; Giri, Mellema & Ghara 2018). This pipeline is used to add realism to our simulated training and testing data set, in order to assess the ability of future SKA 21 cm surveys to constrain the astrophysical and cosmological constraints. In Fig. 1, we show an example of four randomly selected 21 cm maps (top) with their mock versions (bottom) from our training data set for different set of astrophysical and cosmological parameters as quoted in the subtitles. These maps are generated from different simulations realizations of a box size of $L = 150$ Mpc, number of cells $N = 200$, resulting in a resolution of 0.75 Mpc. We find that most of the large- and small-scale ionized bubbles are still present after adding the instrumental effects. This is due to the high angular resolution of our assumed SKA design as well as the high uv-coverage that extends down to a very small scales ($\sim 3.5\,h\,{\rm Mpc}^{-1}$) during these epochs.

However, fully ionized ($x_{\rm HI} < 0.01$) and fully neutral maps ($x_{\rm HI} > 0.99$), as described later in Section 4, are already excluded from the training sample, since they are identical for different set of parameters. Distinguishing identical maps is challenging for neural networks, where more information, such as redshift evolution, is required to assist parameter recovery at these extreme limits. When the Universe is highly ionized (e.g. $x_{\rm HI} \sim$ 0.1–0.2, see column 4 in Fig. 1), the noise dominates but nevertheless the residual neutral patches can still be seen and recognized. These residual patches are usually different for different set of parameters, which might help

[1] https://github.com/jpober/21cmSense

[2] Modes with zero uv-coverage lie outside the angular resolution of the experiment.

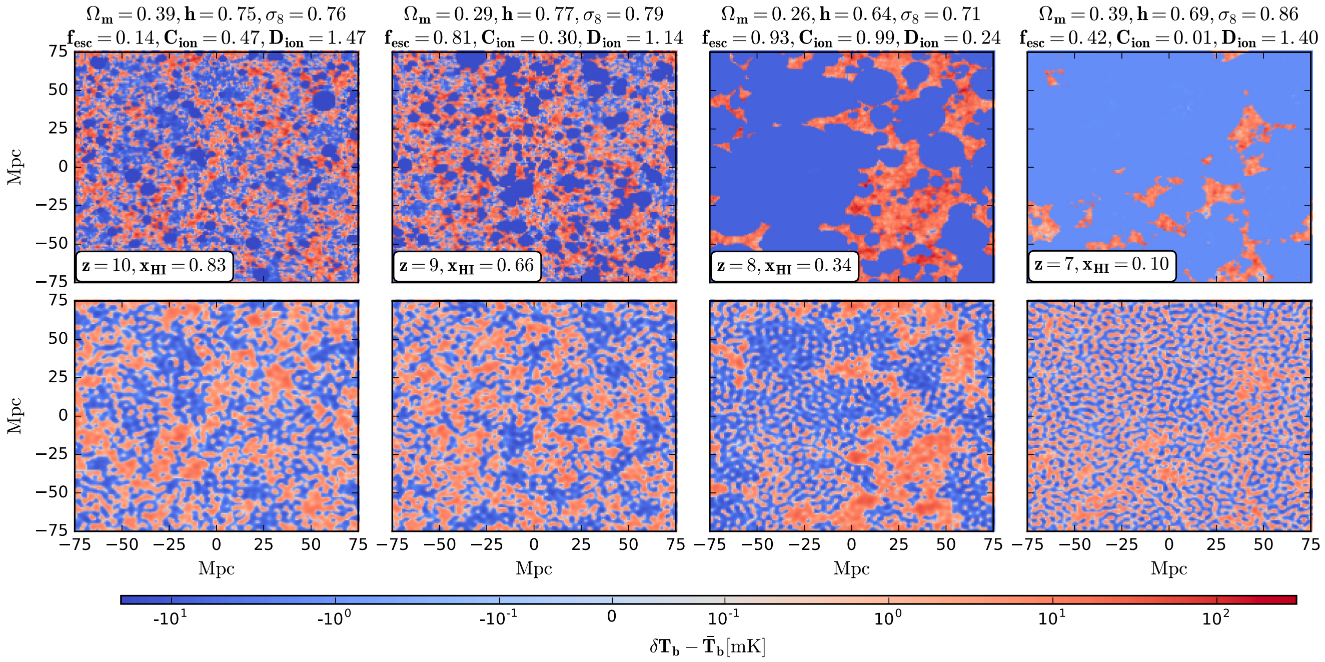

Figure 1. Examples of four randomly selected 21 cm maps (top), from our training data set, with their corresponding mock version (bottom), using our assumed SKA design. Red and blue colour represent neutral and ionized regions, respectively. Subtitles show the astrophysical and cosmological parameters used to generated each map. These parameters are: the photon escape fraction ($f_{\rm esc}$), ionizing emissivity power dependence on halo mass ($C_{\rm ion}$) and ionizing emissivity redshift evolution index ($D_{\rm ion}$), matter density parameter (Ω_m), dimensionless Hubble constant (h), and matter fluctuation amplitude (σ_8). Coloured version is available online.

the network to distinguish between maps and parameters. On the other hand, in the beginning of reionization, the ionized regions are very small due to the small number of sources and ionizing photons. The noise then contaminates and fills these small ionized regions (e.g. $x_{\rm HI} \sim 0.8$–0.9, see column 1 in Fig. 1), and hence maps might look similar to those from a fully neutral Universe. This makes recognizing the prominent signal features more challenging, and many of the reionization realizations for a highly neutral Universe become approximately indistinguishable. This might impact the parameter recovery from a highly neutral IGM, which basically exists at high redshifts where the noise is stronger.

3 NETWORK ARCHITECTURE

We consider two types of network in this work. It is worth reiterating that our main objective is to able to infer both astrophysical $\{f_{\rm esc}$, $C_{\rm ion}$, $D_{\rm ion}\}$ and cosmological $\{\Omega_m$, h, $\sigma_8\}$ parameters simultaneously from their corresponding 21 cm maps. To this end, our main focus is to simply explore different network designs with different layout in width and depth as an attempt to achieve our goal.

The first architecture (*network I*) considered for our investigation is given in Table 2. It is slightly similar to `VGG--Net` (Simonyan & Zisserman 2014) in terms of chaining convolutional layers before downsampling, however the key difference here is that each stage,[3] we have two convolutional layers with same number of feature maps in a row followed by batch normalization and ReLU activation (`Conv+Conv+BN+ReLU`) as shown in Fig. 2. We note that the representation `N x N + M(S)` denotes the kernel size (`N x N`) and the stride (`M`) of a convolutional layer. In total, we

Table 2. The architecture of *network I* for this study.

	Layer	Output shape
1	Input	(1, 200, 200)
2	3 × 3 convolutional layer	(32, 200, 200)
3	3 × 3 convolutional layer	(32, 200, 200)
4	Batch normalization	–
5	ReLU activation	–
6	2 × 2 max pooling	(32, 100, 100)
7	3 × 3 convolutional layer	(64, 100, 100)
8	3 × 3 convolutional layer	(64, 100, 100)
9	Batch normalization	–
10	ReLU activation	–
11	2 × 2 max pooling	(64, 50, 50)
12	3 × 3 convolutional layer	(128, 50, 50)
13	3 × 3 convolutional layer	(128, 50, 50)
14	Batch normalization	–
15	ReLU activation	–
16	2 × 2 max pooling	(128, 25, 25)
17	3 × 3 convolutional layer	(256, 25, 25)
18	3 × 3 convolutional layer	(256, 25, 25)
19	Batch normalization	–
20	ReLU activation	–
21	Fully connected layer	(1024)
22	Batch normalization	–
23	ReLU activation	–
24	Fully connected layer	(1024)
25	Batch normalization	–
26	ReLU activation	–
27	Fully connected layer	(1024)
28	Batch normalization	–
29	ReLU activation	–
30	Fully connected layer	(6)

[3] Which we refer to as mapping the input x without reducing the dimensions (weight × height).

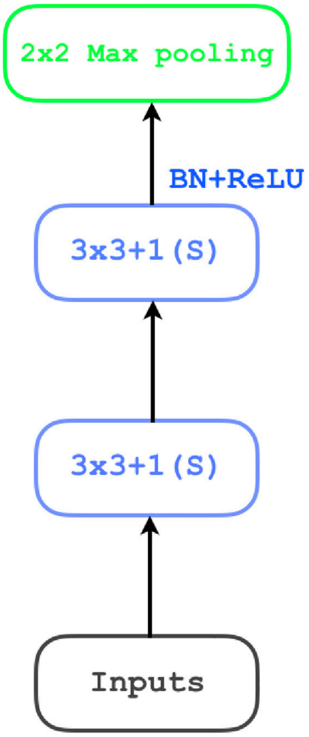

Figure 2. One stage in *network I*. Chaining two convolutional layers with same number of feature maps followed by a batch normalization and ReLU function before a max pooling. Coloured version is available online.

Table 3. The architecture of *network II* for this study.

	Layer	Output shape
1	Input	(1, 200, 200)
2	Convolutional layer	(16, 200, 200)
3	Batch normalization	–
4	ReLU activation	–
5	Residual layer (3 residual blocks)	(16, 100, 100)
6	Residual layer (6 residual blocks)	(32, 50, 50)
7	Residual layer (6 residual blocks)	(64, 25, 25)
8	Residual layer (3 residual blocks)	(128, 13, 13)
9	Inception module	(240, 13, 13)
10	Max pooling	(240, 7, 7)
11	Inception module	(240, 7, 7)
12	Inception module	(256, 7, 7)
13	Inception module	(288, 7, 7)
14	Max pooling	(288, 4, 4)
15	Fully connected layer	(1024)
16	Batch normalization	–
17	ReLU activation	–
18	Fully connected layer	(1024)
19	Batch normalization	–
20	ReLU activation	–
21	Fully connected layer	(1024)
22	Batch normalization	–
23	ReLU activation	–
24	Fully connected layer	(6)

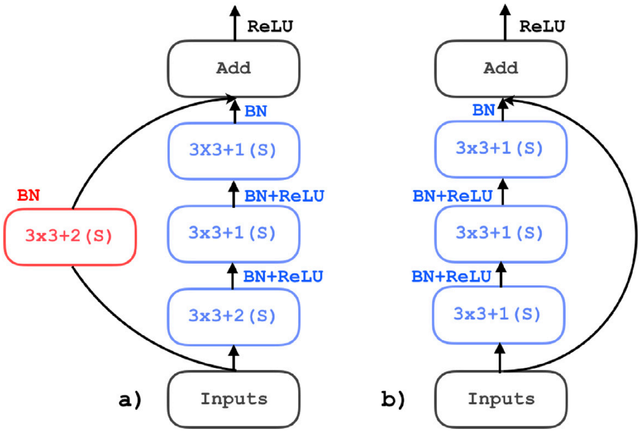

Figure 3. Residual block in *network II*. Left-hand panel: the downsampling only occurs at the first convolutional layer (blue 3x3+2(S)), but the dimension is kept the same at the second convolutional layer (blue 3x3+1(S)). To match the dimensions of the output from the chain of convolutional layers (blue ones), the input is fed to a convolutional layer with strides = 2 (red 3x3+2(S)). Right-hand panel: when there is no downsampling, the input is simply added to the output from the chain of convolutional layers (blue ones). Coloured version is available online.

have four stages, each with a Conv+Conv+BN+ReLU layer followed by a Max Pooling with stride = 2 to reduce the dimensions of the inputs,[4] and four dense layers each with 1024 units with the exception of the output layer, which has only 6 units corresponding to the number of inferred parameters. This network design is also similar to the previous design used in our reionization models classifier (Hassan et al. 2019), except that the convolutional layers used here are wider and no dropout seems to be needed. This is consistent with the disharmony observed between batch normalization and dropout (Li et al. 2018). Similar to our previous works in the classifier, we initialize the network weights using a generalized form of Xavier initializer (Glorot & Bengio 2010) that is also called the Variance Scaling initializer, in which the random numbers are drawn from a zero mean Gaussian distribution whose variance is equal to the inverse of the average of the number of input and output neurons. This initializer ensures that the variance of the input data is preserved as it propagates through the network layers.

Our second architecture, which we simply name *network II*, is based on a combination of residual layers (He et al. 2016) and inception modules (Szegedy et al. 2015) as shown in Table 3. The inputs, as described in Section 2, are first fed into a convolutional layer followed by a batch normalization (Ioffe & Szegedy 2015) before a ReLU activation (Conv+BN+ReLU). This is then followed by four residual layers, each composed of three, six, six, and three residual blocks, respectively. It was shown in He et al. (2016) that the resulting error (both training and testing) of deeper architecture tends to be larger than that of shallower architecture. Therefore, they proposed a residual layer which allowed them to increase the depth of the model in order to gain better performance. In contrast with *network I*, instead of using simple convolutional layers we stack residual blocks, which are achieved with the schematic shown in Fig. 3 (right-hand panel) where the residual learning is constructed using a Conv+BN+ReLU+Conv+BN+ReLU+Conv+BN layer. Depending on whether there is downsampling (Fig. 3 left-hand panel) through the chain of convolutional layers in a residual block, the input needs to be downsampled using a Conv+BN layer to match the dimension of the output of the chain of convolutional layers.

In each residual layer, the downsampling occurs at the first residual block. There are variants of deep residual networks, but in

[4] In other words, the ouputs from the previous stage.

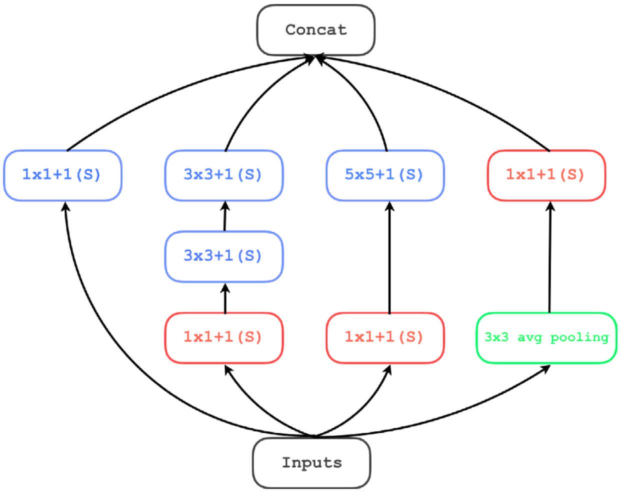

Figure 4. Inception module considered in this study. The red convolutional layers (`1x1+1(S)`) are used for dimensionality reduction. Coloured version is available online.

Table 4. The hyperparameters and optimizers used to train the algorithms.

	Optimizer	Learning rate	Batch	Cost function
I	Nesterov	0.005	128	ℓ_1 norm
II	Adam	0.01	128	rmse

essence what we consider here is such that the network performance is optimized for our specific task.

As proposed by Szegedy et al. (2015), in order to improve the recognition of more complex features at the higher levels of the network, we make use of four inception modules after the residual layers. The prescription suggested in Szegedy et al. (2015) is to deal with the computational complexity related to the depth of the network, that is increasing the size of the network while maintaining the computational cost. The inception module used in this network design is shown in Fig. 4. The idea behind convolving the inputs with a 1 × 1 filter before the convolutional layers with 3 × 3 and 5 × 5 kernels is to reduce the number of feature maps from the inputs as computations are more expensive with larger kernel size. The features at different scales – captured by different kernel sizes 1 × 1, 3 × 3 and 5 × 5 – can be learned *simultaneously* (Szegedy et al. 2015). It is worth noting that we opt for He initialization (He et al. 2016) for all layers in *network II*.

For training, as usual for any machine learning tasks, one needs to fine-tune the hyperparameters, summarized in Table 4, in order for the algorithm to generalize well and hence achieve the best possible performance, where the distance between the ground truth and network predictions is minimum. To that end, as shown in Table B1 in Appendix B, we use two completely different approaches in terms of optimization for the two architectures. Although the two networks produce similar results, as will be presented in Section 5, *network I* converges faster. This can be explained by the capacity of *network I* with its number of trainable parameters of about 167 millions which translates to ∼2.05 × 10^9 floating point operation per second (flops) at inference time, whereas *network II* has ∼10 millions of trainable parameters corresponding to ∼1.39 × 10^9 flops at inference time.

For reproducibility, we have used TENSORFLOW package (Abadi et al. 2016) to develop *network I* which has been trained for 50 000 training steps (about 100 training steps per epoch) which spend ∼ 15 h on a single GPU. Each training step with a batch size of 128 images takes about ∼ 1 s. The network converges from the first few epochs but reaches minimum (RMSE∼ 0.001) at epoch = 40 (see Fig. A1 in Appendix A). This indicates that same results can be obtained in about 6 h with a single GPU. For *network II*, we have used PYTORCH (Paszke et al. 2019) resorting to three GPUs to speed up the convergence. Each epoch, in which each GPU processes in parallel a batch of 128 images at a time, takes about 2 min which is translated to 40 h for 1200 epochs.

4 TRAINING DATA SET

We generate the training data set from a large simulation box of a size $L = 150$ Mpc with $N = 200^3$ cells. We run 1000 different reionization simulations realizations with 1000 different random seeds for the initial density field fluctuations. The prior range assumed to our parameters of interest is as follows:

(i) *Cosmology*:

(a) Matter density parameter: $0.2 \le \Omega_m \le 0.4$.
(b) Hubble constant: $0.6 \le h \le 0.8$.
(c) Matter fluctuation amplitude: $0.7 \le \sigma_8 \le 0.9$.

(ii) *Astrophysics*:

(a) Photon escape fraction: $0.01 \le f_{\rm esc} \le 1$.
(b) $R_{\rm ion}$–M_h power dependence: $0 \le C_{\rm ion} \le 1$.
(c) $R_{\rm ion}$ redshift evolution index: $0 \le D_{\rm ion} \le 2$.

The ranges considered for the astrophysical parameters are motivated from our previous MCMC estimates to reproduce various reionization observables (Hassan et al. 2017), and those of the cosmology are inspired by the recent parameters estimates from the Planck Collaboration 2018 (Aghanim et al. 2020). From these priors, we select 1000 values for each parameter using Latin Hypercube Sampling in order to efficiently explore our 6D parameter space and ensuring that the simulation does not run twice using the same set of parameters. From each simulation run, we store the 21 cm brightness temperature at several redshifts $z = 10$, 9, 8, and 7 in order to have a sufficiently large number of maps to ensure training. Balancing the training data set is important to ensure equal learning at all redshifts and all neutral fraction values. This can be achieved by flattening the distribution according to the neutral fraction at each redshift. Flattening the distributions has previously been used in learning cluster masses (Ntampaka et al. 2015) to reduce the bias towards low-mass clusters. To flatten our distribution, we take the following steps: first, we ignore highly ionized ($x_{\rm HI} < 0.01$) and highly neutral ($x_{\rm HI} > 0.99$) 21 cm boxes. Second, at each redshift, we bin the boxes according to their neutral fraction. Since the neutral fraction changes strongly with different parameters at different redshifts, the number of boxes in each neutral fraction bin is also different. One has to choose a fixed number to select boxes from bins to flatten the distribution. We here choose 20 boxes. These 20 boxes are randomly selected from each bin. If the neutral fraction bin has less than 20 boxes, then we consider all boxes in this specific bin. If all neutral fraction bins have 20 boxes at the four different redshifts, then the total number of all boxes is 800. However, few bins have less than 20 boxes which reduces the total number of boxes to 763. Each selected box has 200 different maps along each of x-, y-, and z-directions. This means each redshift has 200 × 3 × 1000 = 600 000 possible different 21 cm maps. However, close maps in the same box would contain similar features, and from our own experience (e.g.

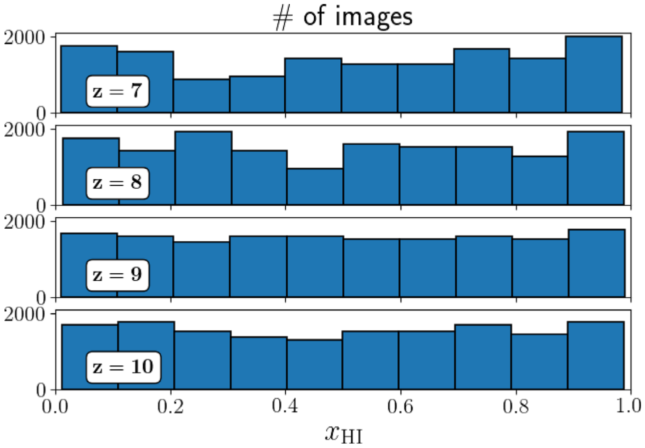

Figure 5. The distribution of training sample at $z = 7$, 8, 9, and 10 (top to bottom) as a function of neutral fraction. We intentionally ignore all current reionization history constraints to check the CNNs' viability to recover parameters without imposing any priors. Coloured version is available online.

Hassan et al. 2019), we have found that ~ 2 Mpc separation between maps is sufficient to obtain distinct maps. To be more conservative, we consider $\sim$ 4 Mpc separation between maps to only select 40 slices along two directions (e.g. x, y) for training, and take 10 slices on the third direction (e.g. z) for testing and validation each. This results in $763 \times (40 \times 2 + 10 + 10) = 76\,300$ total number of images, in which 80 per cent is used for training, 10 per cent for validation, and the remaining 10 per cent for testing. In Fig. 5, we show the histogram of the training data set as a function of neutral fraction at each redshift. The distribution is approximately flat by construction and includes all possible neutral fraction values at each redshift. We here ignore all current reionization constraints to allow specific neutral fraction values at each redshift such as allowing only high neutral fractions at high redshifts and vice versa. This is an important initial test when constraining parameters, which is to verify the method viability to recover these parameters without imposing any priors and constraints. It is worthwhile to mention that the time dependence between maps is included through the following:

(i) Each set of the six parameters corresponds to four boxes at redshifts $z = 10$, 9, 8, and 7. This shows that the network sees the same six parameters for four different maps from four different redshifts.

(ii) The redshift information is encoded in each box through the density field contribution on small scales. This shows that the network sees four different levels of density field contribution in the neutral regions in all maps.

However, an explicit inclusion of this effect is through creating light-cones for each set of the six parameters to account for ionized bubbles growth along the sightline (i.e. the $k_{\parallel}$ modes), redshift-space distortion and angular scales. However, we here assume that all observations at different redshifts are performed independently and apply the noise using a single frequency channel (resolution) corresponding to the map size. The light-cone is more relevant when the full bandwidth is considered. This study sets the baseline for a more detailed analysis to compare the results from 2D maps (single frequency channel) versus 3D light-cones (full bandwidth). Indeed, it is expected that more information exists in the reionization window (e.g. Liu, Parsons & Trott 2014) which contains the bubble evolution along the frequency axis through constructing the light-cone. This we leave for future works as its beyond the scope of the current paper.

While our box size, 150 Mpc, might be small to capture the large-scale fluctuations and cosmic variance (Iliev et al. 2014), we have previously found that our simulation produces a convergent 21 cm power spectrum with respect to the volume (see fig. 8 in Hassan et al. 2016), such that the 150 Mpc volume produces similar power to that from 300 Mpc volume. This indicates the ionized bubbles distribution in 150 Mpc volumes is similar, on average, to those in large volumes. In addition, it has also been found that such a volume produces a convergent reionization history (Iliev et al. 2014). However, the resolution (number of cells) is more important for the neural network performance, since higher resolution maps contain more information and structures. Our maps are composed of 200 x 200 pixels which are able to resolve the small- and large-scale fluctuations reasonably well. We leave investigating the network performance in terms of box size and resolution for future works.

5 RESULTS

To assess how well the algorithms perform in terms of predicting the parameters from learning the input features, we use the coefficient of determination, also known as R^2 score, which is given by

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{6}$$

where $\hat{y}_i$, y_i, and $\bar{y}$ are the predicted value, the actual value and the average of all the actual values in the test sample, respectively. The numerator of the second term in equation (6) – residual sum of squares – quantifies the variation of the predicted values $\hat{y}_i$ around the actual values y_i, and the denominator accounts for the variation of actual values y_i around their mean $\bar{y}$. This metric quantifies the strength of the correlation between the inferred and true values of the parameters, in other words unity R^2 indicates that the network predictions are identical to the ground truth. The R^2 also quantifies the fraction by which the predicted variance is less than the true variance. For each architecture, we carry out two types of training depending on the input features that the regressors are meant to learn from in order to infer our astrophysical and cosmological parameters ($f_{\rm esc}$, $C_{\rm ion}$, $D_{\rm ion}$, $\Omega_{\rm m}$, h, σ_8):

(i) feature extraction from a simulated 2D 21 cm map (*clean/noiseless map* hereafter), this involves training and testing using *clean maps*

(ii) feature extraction from a simulated 2D 21 cm map which was convolved with a simulated SKA like noise (*noisy/mock map* hereafter, see Section 2), this consists of training and testing the networks using *noisy maps*.

It is worthwhile to mention that we train our networks to predict standardized parameters, meaning that we first subtract the mean and divide by the standard deviation for each array of the parameters. After training, we scale back the predictions to the prior range. Standardizing parameters is important, particularly when the parameters range is different, to prevent the highest parameter range from dominating the loss function. This step is commonly used in multiparameters regression deep learning tasks.

5.1 Learning from *clean maps*

The top two rows in Fig. 6 show the test results when training the networks with the *clean maps*. The red and blue areas encompass

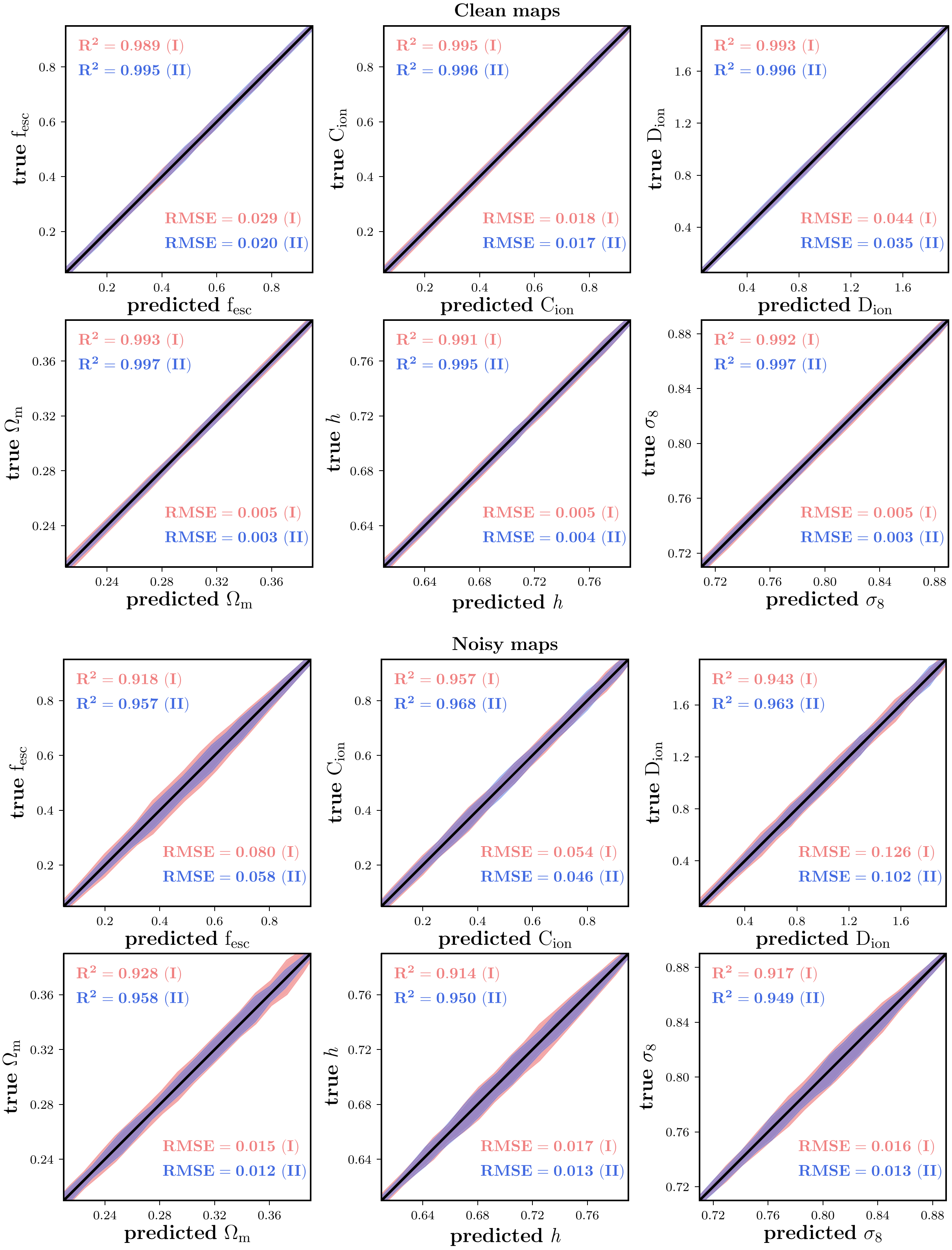

Figure 6. Correlation between the true and predicted parameters. On the top two rows, the networks have been trained with the maps without noise, whereas on the bottom two rows, simulated SKA like noise has been injected into the maps which have been used to train the networks. Red and blue shaded areas encompass the 15.9 per cent and 84.1 per cent percentiles (i.e. $\sim 1\sigma$ level) of the true values given the predictions from *network I* and *network II*, respectively. Solid black lines represent the identity line, that is true parameters versus true parameters. In all cases, the astrophysical parameters recovery is better than those of the cosmology. Adding the noise reduces the accuracy but the parameter recovery is still promising. The *network II* outperforms *network I*, particularly with the mock images, since more complex architecture seems to be needed to extract more information. Large fluctuations are due to low number statistics. Coloured version is available online.

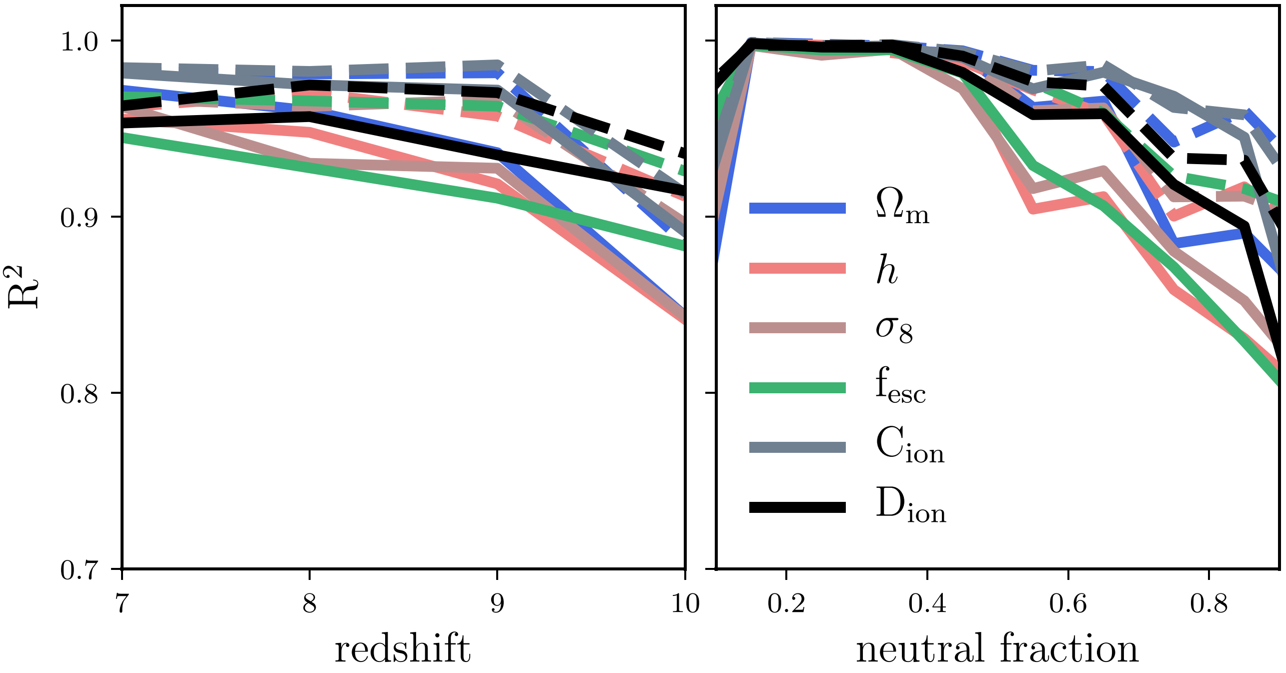

Figure 7. Variation of the resulting coefficient of determination R^2 as a function of redshift (left-hand panel) and neutral fraction (right-hand panel). Solid and dashed lines correspond to *network I* and *network II,* respectively. The accuracy of parameter recovery increases slowly towards low redshift, where the noise is smaller, and rapidly towards low neutral fraction values, where the images features can still be recognized (see Fig. 1). Coloured version is available online.

the 15.9 per cent and 84.1 per cent percentiles (i.e. $\sim 1\sigma$ level) of the true values given the predictions at each bin for *network I* and *network II,* respectively. Overall, the constraints on each parameter are very tight. The high value of the R^2 score (≥ 99 per cent for both *network I* and *network II*) corresponding to each parameter fitting denotes very strong correlation between the true and inferred parameter, suggesting that the algorithms are able to learn the salient features from the data. On comparing the performance of the two architectures, their R^2 score for each fitting suggests that they are in a fairly good agreement, and hence perform equally well.

5.2 Learning from *noisy maps*

The test results after training the algorithms on the *noisy maps* are presented in the bottom two rows in Fig. 6. The constraints on all parameters are slightly weaker as compared to those obtained from training the fitters using the *clean maps* . The overall decrease in performance denoted by the lower values of R^2 score corroborates that finding. This can be accounted for by the fact that the relevant features are in this case convolved with noise, therefore extracting them is a bit more challenging.

Despite being convolved with noise, which essentially causes the quality of their features to degrade, all parameters are successfully recovered with an accuracy of $R^2 \geq 92$ per cent for *network I* and $R^2 \geq 95$ per cent for *network II*, which is remarkably promising. In contrast to training the algorithms with the *clean maps*, it can be noticed that, overall, *network II* outperforms *network I*, as demonstrated by the R^2 scores of the former, which are a bit higher than those of the latter on all parameters.

5.3 Dependence on redshift and neutral fraction

In real observations, both foreground contamination and the thermal noise become stronger with increasing redshift. One would then expect some form of dependence of the constraints on redshift. To investigate that possibility, we bin the maps according to their redshift in the test sample and do the predictions by considering each bin separately using the regressors trained with the *noisy maps*. The results presented in the left-hand panel of Fig. 7 suggest the parameter recovery improves with decreasing redshift. While *network II* tends to have a slightly higher accuracy for each parameter as a function of redshift than *network I*, the dependence on redshift is fairly mild. This weak dependence is due to the fact that there are all possible neutral fraction values at each redshift, without imposing any prior knowledge to the training data set by allowing certain neutral fraction values for each redshift, following the current reionization history constraints.

As mentioned and seen earlier, the observed features in a 21 cm map are more dependent on $x_{\rm HI}$. To address this effect on the performance of the algorithms, we now bin the slices according their value of $x_{\rm HI}$. It is noticeable in Fig. 7 (right-hand panel) that the performance of each fitter on all parameters declines with increasing value of the neutral fraction. This is expected, as previously seen in Fig. 1, the noise always dominates the ionized regions. When the Universe is highly ionized, the prominent features, which are probably the recombining clumps of the remaining dense gas, can still be seen in the presence of noise. This is in contrast to the case where the Universe is highly neutral, and the bubbles are small. Here, the ionized bubbles extend to much smaller scales where the noise dominates, and hence recognition of the bubbles becomes challenging. At this limit, different realizations (with different sets of parameters) of a highly neutral Universe would look similar. This also explains the rapid increase of the accuracy of parameter recovery towards low neutral fraction values. Similar trends have been recently found with using deep learning to constrain the reionization history (e.g. Mangena et al. 2020). It is worthwhile mentioning that this interesting dependence on the neutral fraction cannot be used in future observations since the exact neutral fraction is not known prior observations, although some constraints can be obtained independently from Lyα forest observations (e.g. Fan, Carilli & Keating 2006). However, it is beyond the scope of current networks to use this dependence to constrain parameters. It is rather an interesting theoretical finding and consistent with redshift dependence trends since the Universe is highly ionized at low redshifts.

Having established that the constraints are tighter at lower redshift and lower neutral fraction, that is ionized case, we now apply some conditions on the test sample as follows

(i) select examples with $x_{\rm HI} < 0.5$,

(ii) select examples with lower neutral fraction at lower redshift, $x_{\rm HI} < 0.5$ and $z < 9$.

We show in the top two rows of Fig. 8 the resulting constraints on all parameters when considering maps with $x_{\rm HI} < 0.5$. The results show how the constraints greatly improve by selecting examples with lower neutral fraction from the test sample. For this specific case, the coefficient of determination value $\geq$0.94 for *network I* and $\geq$0.96 for *network II* on all parameters indicates that the performance of the algorithms, despite considering *mock maps* for their training, is comparable to their performance when trained with *noiseless maps*.

Restricting ourselves to lower z, together with only selecting maps with low neutral fraction, in the test data set further improves the predictions on all parameters as indicated by $R^2 \geq 0.99$ for both *network I* and *network II* (see Fig. 8, bottom two rows). Inferring parameters from *noisy maps* at higher redshift is more challenging, since the noise is stronger (see Fig. 7, left-hand panel). Therefore one would expect further improvement of the predictions by combining the two criteria $x_{\rm HI} < 0.5$ and $z < 0.9$. This is an exciting result for future 21 cm surveys that tighter constraints can be obtained from low-redshift observations ($z \sim 6$ and 7), where the Universe is highly ionized. This finding is supported by the fact that the noise is higher at high redshifts, and further confirmed by our additional tests in Appendix B. Using this technique, the SKA will be able to place their first constraints on the astrophysical and cosmological parameters in the near future and from the first cycle of imaging.

5.4 Generalization error

For the sake of completeness and in order to able to compare our results to other similar studies, we compute, for each parameter, the resulting root mean square error, RMSE, as follows:

$$\text{rmse} = \sqrt{\frac{1}{N}\sum(y_{\rm predicted} - y_{\rm true})^2}\,, \qquad (7)$$

where the summation runs through the whole test data set. This metric, among others, tells us about the generalization error inherent to our parameter estimation, in other words the level of accuracy,[5] the fitters can achieve on average when estimating the parameters from encoding the inputs. We show the RMSE values obtained for each parameter when considering both *noiseless* and *noisy maps* in Fig. 6 (two top and two bottom rows, respectively). By comparing the RMSE values resulting from training on *clean maps* and those obtained from training on *mock maps*, we find that in the idealized scenario the prediction is subject to smaller average error for each parameter in contrast to the realistic one. This finding is expected and consistent with our results based on the R^2 metric in that the inference is more challenging for each parameter when the data considered for training/testing are contaminated by noise. Although the results based on the two metrics are found to be consistent, it is tempting to expect that for any two different parameters, irrespective of the case (*noisy* or *clean maps*), if the R^2 score of one of them is higher than that of the other, it implies that its RMSE must be lower. This trend is seen for all parameters as quoted in the legends.

Gupta et al. (2018), by training a CNN with ~26 millions parameters to predict cosmological parameters from simulated noiseless convergence maps, arrived at a generalization error of 35×10^{-3} for Ω_m and 40.3×10^{-3} for σ_8. Ribli, Pataki & Csabai (2018) improved the constraints with a different neural network architecture of about ~1.4 millions parameters, by also using simulated lensing maps, achieving RMSE $= 5.5 \times 10^{-3}$, 13.5×10^{-3} for Ω_m and σ_8, respectively. In terms of encoding features from a 2D map using CNN to infer cosmological parameters, our results – RMSE $= 5 \times 10^{-3}$ and 3×10^{-3} on Ω_m and σ_8 from *network I* and *network II,* respectively – are comparable to those obtained in these previous works. More importantly, our results corresponding to the realistic case, with/without imposing constraints (see Fig. 6 bottom row and Fig. 8), where the input maps are *noisy* are very promising and exciting for future 21 cm surveys.

6 CONCLUSIONS

We have demonstrated in this work the feasibility of simultaneously inferring both astrophysical and cosmological parameters ($f_{\rm esc}$, $C_{\rm ion}$, $D_{\rm ion}$, Ω_m, h, σ_8) using 21 cm maps from the EoR, considering future H I surveys with the SKA. To this end, we have generated thousands of realizations each with a different set of parameters using SIMFAST21, then compiled a data set composed of 2D maps (see Section 4 for details). The approach is to train our two proposed algorithms – CNN-based – to extract the features from the maps in order to predict the underlying astrophysical and cosmological parameters. We have considered an optimistic case where we train the networks with *noiseless* simulated maps and a real world-mimicking scenario in which the networks are trained with simulated maps contaminated by simulated SKA-like noise. We have used R^2 – coefficient of determination – as a performance metric.

We summarize our findings as follows:

(i) The overall results for the idealized case, with $R^2 \geq$ 99 per cent for both *network I* and *network II* on all parameters, suggest that the algorithms considered in this work are capable of learning the salient features from the maps in order to constrain the corresponding parameters with a remarkably excellent accuracy.

(ii) In a more realistic setup, where maps from observations are subject to noise contamination, the constraints on all parameters are slightly weaker, with an accuracy of $R^2 \geq 92$ per cent for *network I* and ≥ 95 per cent for *network II*. This is expected since disentangling the relevant information from noise is more challenging. It has been found that *network II*, leveraging the combination of residual layers at lower level and inception module at higher level of the architecture, outperforms *network I* despite the former's lower capacity. This then points towards deploying similar architectures to *network II* in a real world scenario.

(iii) In the case of learning from the *noisy maps*, the predictions are dependent on both the underlying neutral fraction of the map and its distance from an observer. The performance of the methods improves with decreasing neutral fraction and, as foreground

[5] Not to be confused with accuracy, the metric used in classification tasks.

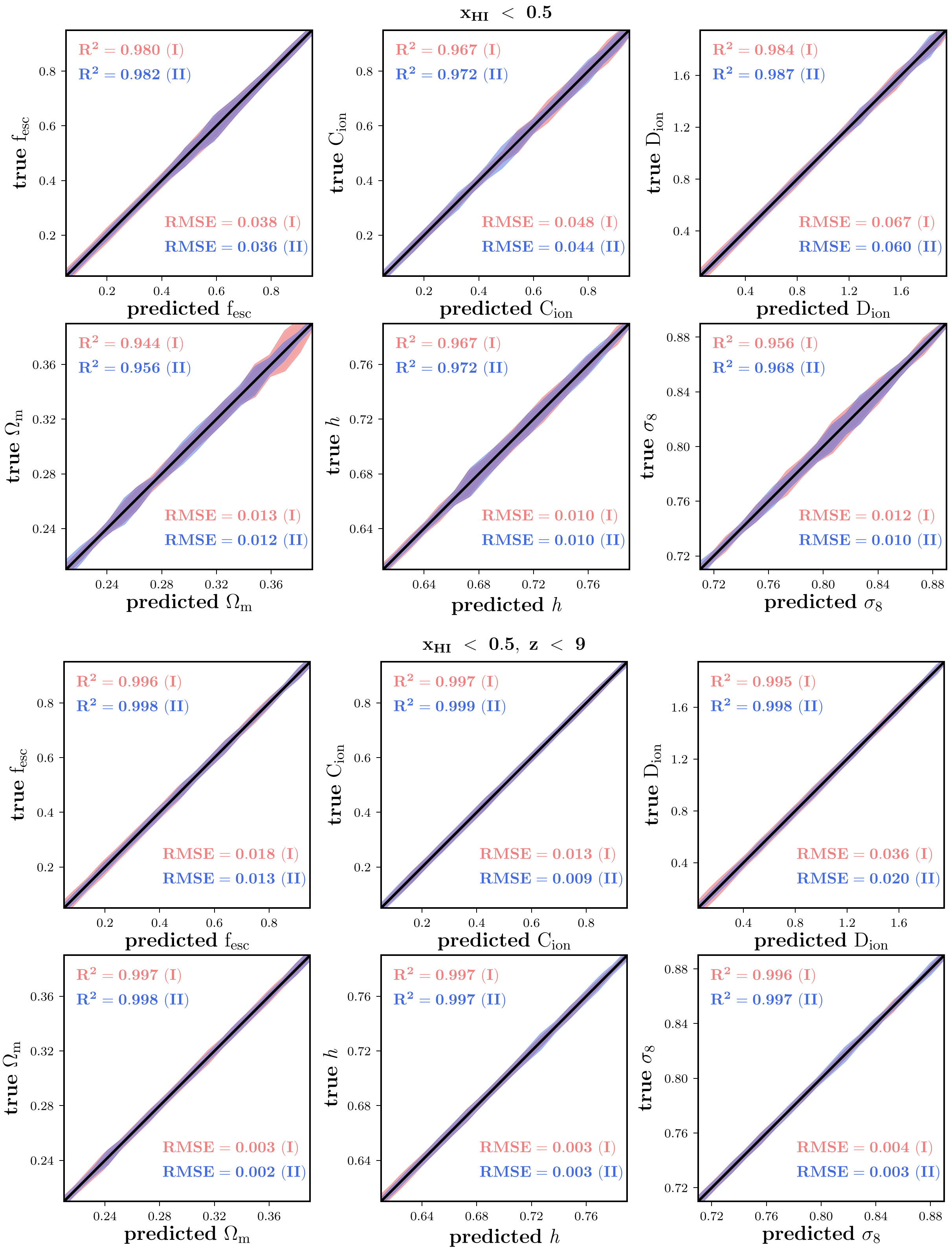

Figure 8. Correlation between the actual and the predicted parameters using the validation sample. On the top two rows, the networks have been trained with the *noisy maps* but a test sample with a neutral fraction <0.5 has been used for predictions. On the bottom two rows, the same *noisy* data have been used to train the networks but some cut on both the neutral fraction <0.5 and redshift $z < 9$ have been applied on the test sample. Red and blue shaded areas encompass the 15.9 per cent and 84.1 per cent percentiles (i.e. $\sim 1\sigma$ level) of the true values given the predictions from *network I* and *network II*, respectively. Solid black line represents the identity line, that is true parameters versus true parameters. Imposing constraints on the neutral fraction and redshift of the testing sample increases the accuracy and performance is comparable to the case without including any instrumental effects as seen in top rows in Fig. 6. Large fluctuations are due to low number statistics. Coloured version is available online.

contamination is more important at higher redshift, the constraints are tighter at lower z. The results obtained from the test sample is $R^2 \geq 94$ per cent with *network I* and $R^2 \geq 96$ per cent with *network II* when only selecting maps with $x_{\rm HI} < 0.5$. Recovery improves with imposing constraints on both neutral fraction and redshift ($x_{\rm HI} < 0.5$ at lower z <9), resulting in $R^2 > 99$ per cent with both *network I* and *network II*. This indicates that even in the presence of *noise* in maps, our methods can still estimate the relevant parameters to an excellent level of precision, which is indeed quite promising.

(iv) We have computed the prediction error on average – RMSE – on each parameter in both optimistic and realistic cases. It has been found that the RMSE is smaller in the former, in good agreement with the results when using the coefficient of determination as a performance metric. Compared to other previous works, our approach has also shown a great potential for inferring the underlying parameters of what is observed in future cosmological experiments, such as H I intensity mapping.

We here considered a redshift range that is consistent with an early reionization scenario, which has been increasingly favoured by galaxy-dominated models of reionization, although more recent work by Kulkarni et al. (2019) shows that galaxies can produce low optical depth and a late reionization scenario. However, late reionization as usually favoured by active galactic nucleus-dominated scenarios is currently disfavoured (e.g. see Qin et al. 2017; Hassan et al. 2018; Mitra, Choudhury & Ferrara 2018; Parsa, Dunlop & McLure 2018). Regardless of the redshift range, the main result, that the accuracy increases with decreasing redshift and neutral fraction, would qualitatively remain valid if lower redshifts are included in this study, such as $z = 5$ and 6, since the instrumental effects are always higher at a higher redshift.

In our analysis, we have generated our training samples based on 1000 different reionization simulations to constrain six parameters. For instance, Gupta et al. (2018), La Plante & Ntampaka (2019), and Gillet et al. (2019) have used 96, 1000, and 10 000 model evaluations to constrain 2, 1, and 3 parameters. Schmit & Pritchard (2018) further have shown that 100 model evaluations is sufficient to constrain three parameters. In comparison to these works, the number of simulations used in this study is low, which limits the presented results. The prior range assumed in this study is also small (i.e. 0.2–2) which places additional limitation to our results. Higher accuracy than reported in this study is expected with larger training samples and more model evaluations, which we will explore in future works.

Our results are entirely limited to the set of assumptions and approximation implemented in our 21 cm instrument simulation. A more refined and sophisticated recipe to account for all of the implemented instrumental effects, such as the angular resolution, foreground cleaning and thermal noise, might alter our concluding remarks. The approximation and assumptions implemented in the seminumerical simulations, through the use of the ESF to identify the ionized regions, as well as the choice of our dynamic range and resolution, place additional limitations to the presented results. While limited to the SKA, our analysis can be easily extended to include instrumental effects from other 21 cm surveys such as HERA and LOFAR, which we leave for future works to perform a detailed comparison between different array designs and different observing strategies. Inferring parameters from the 3D light-cones might improve recovery in the presence of noise without the need to impose constraints on the neutral fraction or redshift. Our analysis also can be easily extended to include all of the astrophysical parameters from the source and sink models, and all cosmological parameters, which we leave for future works.

This study has not only highlighted the constraining power of our methods, probing deep into EoR in the near future with the arrival of more advanced H I instruments like SKA, but also shown how future 21 cm surveys and H I intensity mapping can help break the degeneracy between models by combining them with other experiments, such as *Planck*, to better the constraints on cosmological parameters in an era of precision cosmology.

ACKNOWLEDGEMENTS

The authors acknowledge helpful discussions with Laura Boucheron, Kristian Finlator, Jon Holtzman, Tumelo Mangena, James McAteer, Mario Santos, and Rene Walterbos. We particularly thank the anonymous referee for their constructive comments which have improved the paper quality significantly. Simulations and analysis were performed at UWC's PUMBAA, IDIA/ILIFU cloud computing facilities, and NMSU's DISCOVERY supercomputers. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562, and computational resources (Bridges) provided through the allocation AST190003P. SA acknowledges financial support from the *South African Radio Astronomy Observatory* (SARAO) and is grateful to Sean February and Martin Slabber from the *Science Data Processing* (SDP) team at SARAO for their strong technical support with regards to the computing resources. CCD thanks the Large Synoptic Survey Telescope Corporation (LSSTC) Data Science Fellowship Program, which is funded by LSSTC, NSF Cybertraining grant number 1829740, the Brinson Foundation, and the Moore Foundation; their participation in the program has benefitted this work. CCD acknowledges funding from the New Mexico Space Grant Consortium grant number NNX15AL51H.

REFERENCES

Abadi M. et al., 2016, preprint (arXiv:1603.04467)
Aghanim N. et al., 2020, A&A, preprint (arXiv:1807.06209)
Barkana R., Loeb A., 2008, MNRAS, 384, 1069
Battye R. A., Weller J., 2003, Phys. Rev. D, 68, 083506
Becker G. D., Bolton J. S., 2013, MNRAS, 436, 1023
Bharadwaj S., Pandey S. K., 2005, MNRAS, 358, 968
Bowman J. D. et al., 2013, PASA, 30, e031
Cardone V. F., Dainotti M. G., Capozziello S., Willingale R., 2010, MNRAS, 408, 1181
Chardin J., Uhlrich G., Aubert D., Deparis N., Gillet N., Ocvirk P., Lewis J., 2019, MNRAS, 490, 1055
Contaldi C. R., Hoekstra H., Lewis A., 2003, Phys. Rev. Lett., 90, 221303
Davé R., Katz N., Oppenheimer B. D., Kollmeier J. A., Weinberg D. H., 2013, MNRAS, 434, 2645
DeBoer D. R. et al., 2017, PASP, 129, 045001
Fan X., Carilli C. L., Keating B., 2006, ARA&A, 44, 415
Finlator K., Davé R., Özel F., 2011, ApJ, 743, 169
Finlator K., Thompson R., Huang S., Davé R., Zackrisson E., Oppenheimer B. D., 2015, MNRAS, 447, 2526
Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Phys. Rep., 433, 181
Gillet N., Mesinger A., Greig B., Liu A., Ucci G., 2019, MNRAS, 484, 282
Giri S. K., Mellema G., Ghara R., 2018, MNRAS, 479, 5596
Glorot X., Bengio Y., 2010, Proc. 13th Int. Conf. Vol. 9, Artificial Intelligence and Statistics. PLMR, p. 249
Greig B., Mesinger A., 2015, MNRAS, 449, 4246
Gupta A., Matilla J. M. Z., Hsu D., Haiman Z., 2018, Phys. Rev. D, 97, 103515

Hassan S., Davé R., Finlator K., Santos M. G., 2016, MNRAS, 457, 1550
Hassan S., Davé R., Finlator K., Santos M. G., 2017, MNRAS, 468, 122
Hassan S., Davé R., Mitra S., Finlator K., Ciardi B., Santos M. G., 2018, MNRAS, 473, 227
Hassan S., Liu A., Kohn S., La Plante P., 2019, MNRAS, 483, 2524
He K., Zhang X., Ren S., Sun J., 2016, Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). p. 770
Hinshaw G. et al., 2013, ApJS, 208, 19
Iliev I. T., Mellema G., Ahn K., Shapiro P. R., Mao Y., Pen U.-L., 2014, MNRAS, 439, 725
Ioffe S., Szegedy C., 2015, in Bach F., Blei D., eds, Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. JMLR.org, ICML'15
Kakiichi K. et al., 2017, MNRAS, 471, 1936
Kulkarni G., Keating L. C., Haehnelt M. G., Bosman S. E. I., Puchwein E., Chardin J., Aubert D., 2019, MNRAS, 485, L24
La Plante P., Ntampaka M., 2019, ApJ, 880, 110
Leitet E., Bergvall N., Hayes M., Linné S., Zackrisson E., 2013, A&A, 553, A106
Li X., Chen S., Hu X., Yang J., 2018, preprint (arXiv:1801.05134)
Li Y., Fan X., Gou L., 2019, ApJ, 873, 37
Liu A., Parsons A. R., Trott C. M., 2014, Phys. Rev. D, 90, 023018
Liu A., Pritchard J. R., Allison R., Parsons A. R., Seljak U., Sherwin B. D., 2016, Phys. Rev. D, 93, 043013
Loeb A., Barkana R., 2001, ARA&A, 39, 19
Majumdar S., Pritchard J. R., Mondal R., Watkinson C. A., Bharadwaj S., Mellema G., 2018, MNRAS, 476, 4007
Mangena T., Hassan S., Santos M. G., 2020, MNRAS, 494, 600
Mellema G. et al., 2013, Exper. Astron., 36, 235
Mitra S., Choudhury T. R., Ferrara A., 2015, MNRAS, 454, L76
Mitra S., Choudhury T. R., Ferrara A., 2018, MNRAS, 473, 1416
Molaro M., Davé R., Hassan S., Santos M. G., Finlator K., 2019, MNRAS, 489, 5594
Moscardini L., Matarrese S., Mo H., 2001, MNRAS, 327, 422
Ntampaka M., Trac H., Sutherland D. J., Battaglia N., Póczos B., Schneider J., 2015, ApJ, 803, 50
Paciga G. et al., 2011, MNRAS, 413, 1174
Padmanabhan N. et al., 2007, MNRAS, 378, 852
Paranjape A., Choudhury T. R., Padmanabhan H., 2016, MNRAS, 460, 1801
Park J., Mesinger A., Greig B., Gillet N., 2019, MNRAS, 484, 933
Parsa S., Dunlop J. S., McLure R. J., 2018, MNRAS, 474, 2904
Parsons A. R. et al., 2010, AJ, 139, 1468
Paszke A. et al., 2019, in Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., eds, Advances in Neural Information Processing Systems 32. Curran Associates, Inc., New York, p. 8024
Phillips J., Weinberg D. H., Croft R. A. C., Hernquist L., Katz N., Pettini M., 2001, ApJ, 560, 15
Planck Collaboration XIII, 2016, A&A, 594, A13
Pober J. C., Greig B., Mesinger A., 2016, MNRAS, 463, L56
Qin Y. et al., 2017, MNRAS, 472, 2009
Ribli D., Pataki B. Á., Csabai I., 2018, Nat. Astron., 3, 93
Santos M. G., Amblard A., Pritchard J., Trac H., Cen R., Cooray A., 2008, ApJ, 689, 1
Santos M., Ferramacho L., Silva M., Amblard A., Cooray A., 2010, MNRAS, 406, 2421
Schmit C. J., Pritchard J. R., 2018, MNRAS, 475, 1213
Simonyan K., Zisserman A., 2014, preprint (arXiv:1409.1556)
Szegedy C. et al., 2015, Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Going Deeper with Convolutions. Boston, MA, p. 1
van Haarlem M. P. et al., 2013, A&A, 556, A2
Watkinson C. A., Pritchard J. R., 2015, MNRAS, 454, 1416
Zahn O., Lidz A., McQuinn M., Dutta S., Hernquist L., Zaldarriaga M., Furlanetto S. R., 2007, ApJ, 654, 12
Zaldarriaga M., Furlanetto S. R., Hernquist L., 2004, ApJ, 608, 622
Zel'dovich Y. B., 1970, A&A, 5, 84

APPENDIX A: THE LOSS FUNCTION EVOLUTION DURING TRAINING

Fig. A1 shows the loss evolution of *networks I* (left) and *II* (right) for training (blue) and validation (red) samples as a function of training epoch for the case of training on the noisy data set. In both cases, the loss is decreasing as training progresses, which indicates a reduction in the error rate and predictions are approaching the target labels. The fluctuations in the training and validation curves are due to random selection of batches during training. Regardless

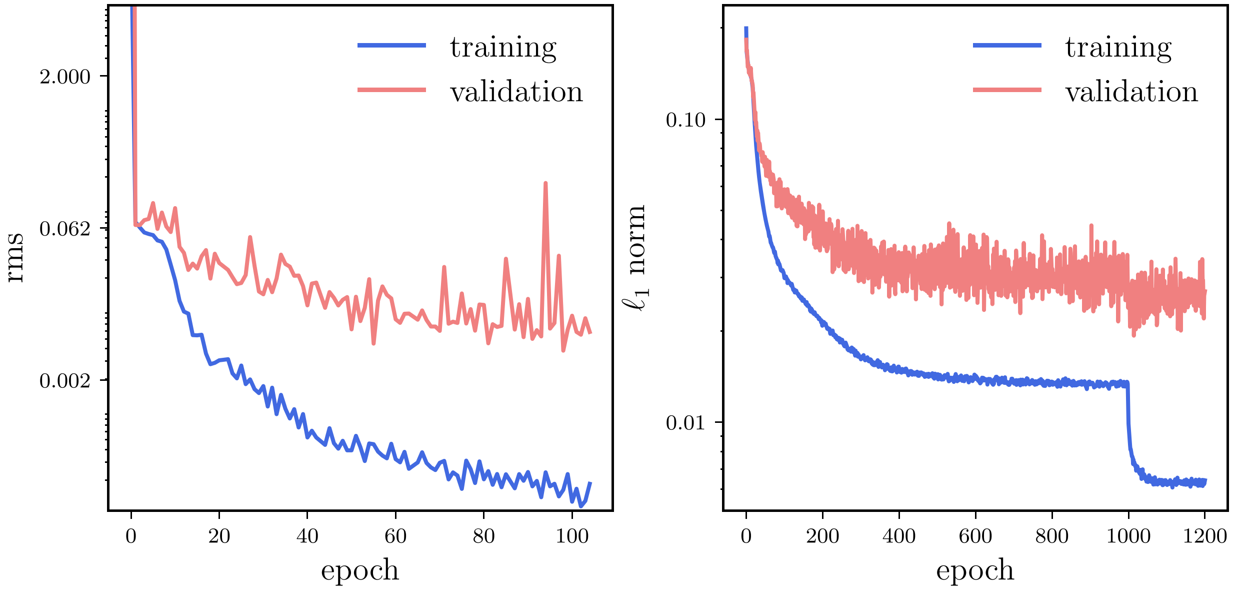

Figure A1. Left-hand panel: progression of the training of *network I*, where the loss function RMSE varies as a function of training epoch. Right-hand panel: progression of the training of *network II* showing the loss function ℓ_1 norm as a function of number of epochs. Coloured version is available online. The two plots are related to the training on the noisy data.

of these fluctuations, the loss evolution for validation converges and stays constant on average, which indicates that the networks are not overfitting. It is worth noting that the sudden drop in training/validation error while training *network II* is owing to the fact that the learning rate is updated to 10 per cent of its initial value in order to escape the plateau.

APPENDIX B: REDSHIFT EVOLUTION

Our main result which suggests that the accuracy increases with decreasing redshift has been derived from a model trained on mixed maps from all redshifts. To confirm whether accuracy increases towards low redshift, we here perform additional learnings by restricting the training sample to have maps only from the minimum ($z = 7$) or maximum ($z = 10$) redshifts considered in this study (referred to as only). We also compare with predictions at these redshifts from training with the whole data set, including all other redshifts (referred to as whole), for the case of noisy maps as reported in Table B1. In all cases, we find that the accuracy at $z = 7$ is always higher than that of at $z = 10$. This shows, regardless of training with whole mixed maps or maps at a given redshift, the qualitative trend that the accuracy increases towards low redshifts is still seen as summarized in Table B1. In addition, the quantitative results are also similar with a minimal difference of about $\lesssim 2$ per cent of accuracy for some parameters as summarized in Table B1 for training with maps at individual redshifts (only) versus those derived using a trained model on mixed maps (whole). Such a minimal difference is expected due to the different number of samples used in the case of ‘only’ versus ‘whole’ tests. This shows that our networks are successful to recover the same qualitative and quantitative results without explicitly including the redshift information as an input to the network (e.g. fitting parameters to four maps from the four different redshifts, $z = 10$–7).

Table B1. Networks accuracy comparison between training only with data set from $z = 7$ and 10 (referred to as only) versus predicting at these redshifts from training with whole data set (including all other redshifts, referred to as whole), for the case of noisy maps. For all parameters with all networks, accuracy increases towards low redshift.

	Network I				*Network II*			
	$z = 10$ (only)	$z = 10$ (whole)	$z = 7$ (only)	$z = 7$ (whole)	$z = 10$ (only)	$z = 10$ (whole)	$z = 7$ (only)	$z = 7$ (whole)
Ω_m	0.86	0.84	0.97	0.97	0.86	0.88	0.96	0.98
h	0.87	0.84	0.95	0.95	0.88	0.91	0.95	0.96
σ_8	0.86	0.84	0.95	0.96	0.88	0.89	0.96	0.96
$f_{\rm esc}$	0.90	0.88	0.95	0.94	0.90	0.92	0.95	0.96
$C_{\rm ion}$	0.91	0.89	0.98	0.98	0.90	0.91	0.98	0.98
$D_{\rm ion}$	0.93	0.91	0.95	0.95	0.92	0.93	0.95	0.96

This paper has been typeset from a TEX/LATEX file prepared by the author.