Deep-Learning Study of the 21cm Differential Brightness Temperature During the Epoch of Reionization

Yungi Kwon

Department of Physics, University of Seoul, 163 Seoulsiripdaero,
Dongdaemun-gu, Seoul 02504, Republic of Korea

Sungwook E. Hong*

Natural Science Research Institute, University of Seoul,
163 Seoulsiripdaero, Dongdaemun-gu, Seoul 02504, Korea

Inkyu Park

Department of Physics, University of Seoul,
163 Seoulsiripdaero, Dongdaemun-gu,
Seoul 02504, Republic of Korea and
Natural Science Research Institute, University of Seoul,
163 Seoulsiripdaero, Dongdaemun-gu, Seoul 02504, Korea

Abstract

We propose a deep learning analyzing technique with convolutional neural network (CNN) to predict the evolutionary track of the Epoch of Reionization (EoR) from the 21-cm differential brightness temperature tomography images. We use 21cmFAST, a fast semi-numerical cosmological 21-cm signal simulator, to produce mock 21-cm maps between $z = 6 \sim 13$. We then apply two observational effects into those 21-cm maps, such as instrumental noise and limit of (spatial and depth) resolution somewhat suitable for realistic choices of the Square Kilometre Array (SKA). We design our deep learning model with CNN to predict the sliced-averaged neutral hydrogen fraction from the given 21-cm map. The estimated neutral fraction from our CNN model has a great agreement with its true value even after coarsely smoothing with broad beamsize and frequency bandwidth, and also heavily covered by noise with narrow. Our results have shown that deep learning analyzing method has a large potential to efficiently reconstruct the EoR history from the 21-cm tomography surveys in future.

PACS numbers: 32.80.Fb, 95.85.Bh, 98.58.Ge, 98.70.Vc, 98.80.Es

Keywords: epoch of reionization, deep learning

---
*E-mail: swhong83@uos.ac.kr

I. INTRODUCTION

About 380,000 years after the Big Bang, the Universe had cooled down enough that the floating free protons and electrons could combine to form the neutral hydrogen atoms (HI). As there were no luminous objects such as stars and galaxies, this period is referred to as the cosmic dark ages. During the cosmic dark ages, most of the hydrogen atoms exist as neutral. Then, by the gravitational collapse of the overdense regions in the intergalactic medium (IGM), they started to form more and more pronounced structures of the Universe. Eventually, the first luminous objects started to form, which leads the epoch of reionization (EoR). During this epoch, the first radiating objects in the Universe heat and re-ionize the surrounding local neutral IGM by energetic radiation.

As the EoR involves the key science to understand the formation of first objects and large-scale structures, numerous studies have performed to understand the epoch. For example, several observations have constrained the duration of the EoR from redshifts $z \simeq 12$ to 6 [1–5]. However, due to the lack of bright objects, the current observational evidences of the EoR are rather limited. For instance, we poorly understand for what is the main source governing the EoR and what properties lead the ionizing process, and how did it intimately affect on subsequent structure formation.

Several observational methods have been suggested to understand the EoR. For example, [6–8] studied how the cosmic microwave background (CMB) anisotropy observations can constrain the evolutionary history of the mean neutral hydrogen fraction ($\bar{x}_{\mathrm{HI}}(z)$). Among various observational methods for studying the EoR, one of the most promising ways is to use the 21-cm wavelength radiation emitted by the hyperfine transition of neutral hydrogen atom, as the hydrogen gas was predominant component of the Universe. Several radio telescope experiments are being planned and accumulated for observing the redshifted 21-cm signal from the EoR: the Murchison Widefield Array (MWA) [9], the Giant Metrewave Radio Telescope (GMRT) [10], the Low Frequency Array (LOFAR) [11], the Precision Array Probing the Epoch of Reionization (PAPER) [12], the Hydrogen Epoch of Reionization Array (HERA) [13], and the Square Kilometre Array (SKA) [14, 15]. Nevertheless, there remains a challenge to extract the proper redshifted 21-cm signal during the EoR from radio observations, due to the foregrounds from the extra-galactic radio sources and diffuse galactic foreground, the ionospheric distortions, and the instrumental noise, etc. Specifically,

the Galactic synchrotron radiation, which has similar wavelength to those of the redshifted 21-cm signal during the EoR, is expected to be much brighter than the cosmological 21-cm signal [16–18].

While numerous methods have suggested on subtracting foregrounds and analyzing the 21-cm signals [19–22], brand-new analyzing techniques have increasingly approved recently. One of them is the deep learning, a subset of machine learning that works with artificial neural networks (ANN), that is designed to imitate how humans think and learn by using the so-called neurons. The neural network accepts a series of data as input and grasps the certain patterns within the data by fitting the weights on the connections between neurons of the network. Deep learning technique has shown to have enormous potential on astrophysics [23–29]. For example, [30] used the deep learning technique to estimate the ionizing efficiency, the minimum virial temperature of halos, and the mean free path of ionizing photons by using the 21-cm power spectra at different redshifts as inputs. Also, [31] used the entire 21-cm lightcone map as the input of the convolutional neural network (CNN) to discriminate the re-ionizing process from the active galactic nuclei (AGNs) and star-forming galaxies. To date, however, a series of studies has proceeded to deal with somewhat ideal simulated data as being poorly considered the actual observational effects.

In this paper, we introduce a novel deep learning-based method by re-constructing $\bar{x}_{\mathrm{HI}}(z)$ during the EoR from mock 21-cm differential brightness temperature maps by considering various observational effects. We simulate 21-cm maps by using a semi-numerical simulation code 21cmFAST [32] at different redshifts. We convolve them with certain beamsize and frequency bandwidth under somewhat realistic observational conditions, and add the corresponding white Gaussian noise to generate the noisy maps. We use those noisy 21-cm maps as inputs of our CNN model and train the model to predict the corresponding value of x_{HI}.

This paper is organized as follows. In Section II. , we describe how to simulate the mock 21-cm differential brightness temperature maps. In Section III. , we introduce our CNN architecture and training method. In Section IV. , we show the performance of our method with various observational conditions. Finally, in Section V. , we summarize our results.

Throughout this paper, we adopt the background cosmological parameters best fit to the standard ΛCDM cosmology from Planck 2018 [33]: matter density parameter $\Omega_{\mathrm{m}} = 0.31$, baryon density parameter $\Omega_{\mathrm{b}} = 0.048$, cosmological constant density parameter $\Omega_{\Lambda} = 0.69$,

spectral index $n_s = 0.97$, rms density fluctuation $\sigma_8 = 0.81$, Hubble parameter $h = 0.68$, and the primordial helium abundance $Y_p = 0.245$.

II. SIMULATION OF NOISY MOCK 21-CM MAPS

1. Simulating Undistorted Redshifted 21-cm Signals

We use the 21cmFAST [32] to generate the 21-cm signal and corresponding neutral fraction with the redshifts $z = 6 \sim 13$, where most of the current studies strongly have constrained. The 21cmFAST is the publicly-available simulation of the cosmological reionization which is self-consistent and semi-numerical. Specifically, this is optimized to generate the 21-cm signal during the epoch. By the combinations of the excursion-set formalism [34,35] which mentioned earlier for identifying ionized hydrogen regions and the first-order perturbation theory [36], it generates the full three-dimensional realizations of the density, ionization field, velocity, and spin temperature and finally the 21-cm differential brightness temperature with a given redshift.

We first generate the density and velocity fields with 50, 100, 200, and 300 Mpc comoving boxsizes and 800^3, 1000^3, and 1250^3 grids between $z = 6 \sim 13$. We use different random seeds to generate the initial conditions for each combination of comoving boxsize and grid number, and, as a result, we get 12 independent simulation sets. For each simulation set, we obtain $18 \sim 35$ snapshots between $z = 6 \sim 13$, where the number of snapshots depends on the number of grids. Note that the spatial resolutions used in this paper ($\Delta x = 0.04 \sim 0.375$ Mpc) are smaller than the length scale where density perturbation from the 21cmFAST agrees well with the N-body simulations [32]. However, since we do not focus on small-scale details and the output 21-cm maps would be smoothed in larger scales, our choice of spatial resolution does not severely affect our result.

Then we calculate the "undistorted" 21-cm differential brightness temperature, $\delta T_b^0(\mathbf{x}, z)$, by downgrading the resolution to 200^3 grids:

$$\delta T_b^0 \approx (27\,\mathrm{mK}) x_{\mathrm{HI}}(1+\delta) \left(\frac{H}{\mathrm{d}v_r/\mathrm{d}r + H} \right) \left(1 - \frac{T_{\mathrm{CMB}}}{T_S} \right) \left(\frac{1+z}{10} \frac{0.15}{\Omega_m h^2} \right)^{1/2} \left(\frac{\Omega_b h^2}{0.023} \right), \quad (1)$$

where δ, T_{CMB}, and T_S are the gas overdensity, the CMB temperature, and the gas spin temperature, respectively [17]. In 21cmFAST, $T_S(z)$ is calculated from the evolution of the

kinetic gas temperature and the Ly-α background. Also, $x_{\mathrm{HI}}(\mathbf{x}, z)$ is calculated from the number of IGM ionizing photons per baryon [37]

$$n_{\mathrm{ion}} = \bar{\rho}_{\mathrm{b}}^{-1} \int_0^{\infty} \mathrm{d}M_{\mathrm{h}} \frac{\mathrm{d}n(M_{\mathrm{h}}, z)}{\mathrm{d}M_{\mathrm{h}}} f_{\mathrm{duty}} M_{\star} f_{\mathrm{esc}} N_{\gamma/\mathrm{b}} , \tag{2}$$

where $\bar{\rho}_{\mathrm{b}}$, M_{h}, f_{duty}, $M_{\star}$, f_{esc}, $N_{\gamma/\mathrm{b}}$ are the mean baryon density, halo mass, factor related to the suppression of star formation at massive halos, stellar mass, escape fraction, and the number of ionizing photons per stellar baryon, respectively. f_{duty}, $M_{\star}$, and f_{esc} can be expressed as a function of M_{h} as follows:

$$f_{\mathrm{duty}}(M_{\mathrm{h}}) = \exp\left(-\frac{M_{\mathrm{h}}}{M_{\mathrm{turn}}}\right) \tag{3}$$

$$M_{\star}(M_{\mathrm{h}}) = f_{\star} \frac{\Omega_{\mathrm{b}}}{\Omega_{\mathrm{m}}} M_{\mathrm{h}} = f_{\star,10} \left(\frac{M_{\mathrm{h}}}{10^{10}\,\mathrm{M_\odot}}\right)^{\alpha_{\star}} \frac{\Omega_{\mathrm{b}}}{\Omega_{\mathrm{m}}} M_{\mathrm{h}} \tag{4}$$

$$f_{\mathrm{esc}}(M_{\mathrm{h}}) = f_{\mathrm{esc},10} \left(\frac{M_{\mathrm{h}}}{10^{10}\,\mathrm{M_\odot}}\right)^{\alpha_{\mathrm{esc}}} , \tag{5}$$

where M_{turn} is the halo mass threshold for efficient star formation, and $f_{\star}$ is the fraction of galactic gas in stars. In this paper, we adopt a fiducial setup in [37] for reionization parameters: $(N_{\gamma/\mathrm{b}}, f_{\star,10}, \alpha_{\star}, f_{\mathrm{esc},10}, \alpha_{\mathrm{esc}}, M_{\mathrm{turn}}) = (5000, 0.05, 0.5, 0.1, -0.5, 5 \times 10^8\,\mathrm{M_\odot})$. Note that we do not test different values of reionization parameters in this paper, mainly because we focus on the evolution of the mean neutral hydrogen fraction $\bar{x}_{\mathrm{HI}}(z)$ rather than constraining the reionization model.

Figure 1 shows an example of undistorted 21-cm maps in a $50\,\mathrm{cMpc}$-box. As expected in Eq. (1) (if $T_{\mathrm{S}} \gg T_{\mathrm{CMB}}$, which is believed to be satisfied during the main reionization epoch), the undistorted 21-cm maps are proportional to the matter density and follow Gaussian distribution at high $(z, \bar{x}_{\mathrm{HI}})$. On the other hand, at low $(z, \bar{x}_{\mathrm{HI}})$, the overdense region becomes fully ionized and shows no 21-cm signal, and therefore, the distribution is highly non-Gaussian. Note that the estimation of $\bar{x}_{\mathrm{HI}}$ from the undistorted 21-cm maps could be rather straightforward, as it is simply the fraction of area where $x_{\mathrm{HI}}(\mathbf{x})$ is close to zero. However, the estimation of $\bar{x}_{\mathrm{HI}}$ may not be that straightforward if one adds the observational effects to the 21-cm signal, which we will see in the next section.

2. Processing with the Observational Effects

In practice, we expect that the observed signal would be very different with Figure 1, because the actual observational signal will be obtained by integrating the angle, frequency,
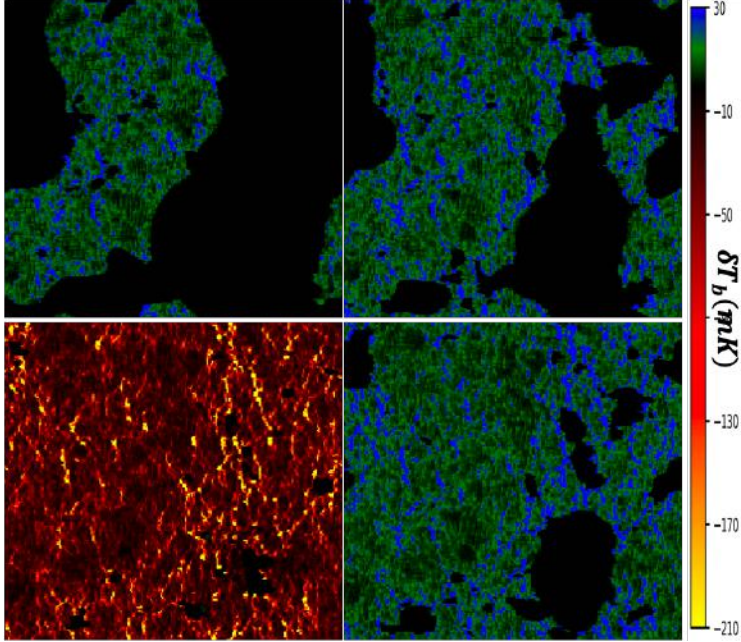
Fig. 1. An example of undistorted mock 21-cm maps with $50\,\mathrm{cMpc}$ length and $0.25\,\mathrm{cMpc}$ thickness. From bottom-left to counter-clockwise direction: $(z, \bar{x}_{\mathrm{HI}}) = (10.66, 0.94)$, $(8.58, 0.76)$, $(7.52, 0.54)$, and $(6.57, 0.25)$.

and observation time. At this point, the size of the angle and frequency are the limit of spatial and depth resolution of the telescope. As the size of the angle or frequency of the telescope becomes larger, the observed signal becomes more smoothed. Then the small-scale information is lost and the estimation of $\bar{x}_{\mathrm{HI}}$ similar to that we mentioned at the previous section may fail. On the other hand, the smaller angle or frequency could contain the small-scale information. However, if the integration time is fixed, the level of noise becomes higher and the estimation of $\bar{x}_{\mathrm{HI}}$ could be also difficult. Therefore, it is essential to include such observational effects, i.e., smoothing and noise, for the performance test of analysis methods such as our CNN model.

We start from smoothing the undistorted 21-cm maps with given angle and frequency setup of radio telescope. Here, we use three choices of the beamsizes and frequency bandwidths—$\Delta\theta = 1'$, $2'$, and $3'$, and $\Delta\nu = 0.2\,\mathrm{MHz}$, $1\,\mathrm{MHz}$, and $2\,\mathrm{MHz}$. The corre-

sponding comoving length scales for given $\Delta\theta$ and $\Delta\nu$ are

$$\Delta L_\perp(\Delta\theta, z) = D_c(z) \times \Delta\theta \tag{6}$$

$$\Delta L_\parallel(\Delta\nu, z) = \frac{c(1+z)^2}{H_0\sqrt{\Omega_m(1+z)^3 + \Omega_\Lambda}}\frac{\Delta\nu}{\nu_0}, \tag{7}$$

where $D_c(z)$ is the comoving distance, and $\nu_0 = 1420\,\text{MHz}$ is the rest-frame frequency of the 21-cm line. Note that the subscripts '$\perp$' and '$\parallel$' emphasize that the beamsize and frequency bandwidth affect the smoothing of δT_b in directions perpendicular and parallel to the line-of-sight (LOS), respectively.

The actual shape of smoothing kernels in directions parallel and perpendicular to the LOS ($W(\nu; \Delta\nu)$ and $W(\theta; \Delta\theta)$, respectively) strongly depends on the configuration of radio telescope, even with the identical choice of $(\Delta\theta, \Delta\nu)$. Especially, the shape of $W(\theta; \Delta\theta)$ (beam shape) could greatly vary depending on the antennae design. It is known that some "dirty" beam shape such as the compensated Gaussian kernel might be possible [38], and in that case the understanding of the $\bar{x}_{HI}(z)$ could be extremely difficult [39]. Nevertheless, for simplicity and because the configuration of future surveys is not fixed, we assume the Gaussian kernel for both beam and frequency smoothing kernels in this paper. We use $\Delta\theta$ and $\Delta\nu$, or identically, $\Delta L_\perp$ and $\Delta L_\parallel$, as the full-width half-maximum of the beam and frequency smoothing kernel.

Also, since interferometric radio observations might be insensitive to the large-scale fluctuation of δT_b, we subtract the mean of the smoothed 21-cm signal from the map so that the average of the subtracted 21-cm map becomes zero. In summary, we obtain the "smoothed" 21-cm map (or, $\delta T_b^s(\boldsymbol{\theta}, \nu)$, where $\nu = \nu_0/(1+z)$) as follows:

$$\delta T_b^s(\boldsymbol{\theta}, \nu) = \delta T_b'(\boldsymbol{\theta}, \nu) - \langle \delta T_b^{s'}(\boldsymbol{\theta'}, \nu)\rangle_{\boldsymbol{\theta'}} \tag{8}$$

$$\delta T_b^{s'}(\boldsymbol{\theta}, \nu) = \int d\nu' \int d\boldsymbol{\theta'}\, \delta T_b^0(\boldsymbol{\theta'}, \nu') W_G(|\boldsymbol{\theta} - \boldsymbol{\theta'}|; \Delta\theta) W_G(\nu - \nu'; \Delta\nu). \tag{9}$$

Here, the subscript 'G' emphasizes the Gaussian kernel.

Figure 2 shows an example of the smoothed mock 21-cm map in a $300\,\text{cMpc}$-box at $(z, \bar{x}_{HI}) = (12.73, 0.97)$. Compared to the undistorted 21-cm maps in Figure 1, there exists two key features in smoothed maps that make the estimation of $\bar{x}_{HI}(z)$ more difficult. First of all, due to the mean subtraction after smoothing, we cannot use the mean value ($\langle \delta T_b^s(\boldsymbol{\theta}, \nu)\rangle_{\boldsymbol{\theta}}$) as an indicator of $\bar{x}_{HI}(z)$, especially for high-z cases (see the bottom-left panel
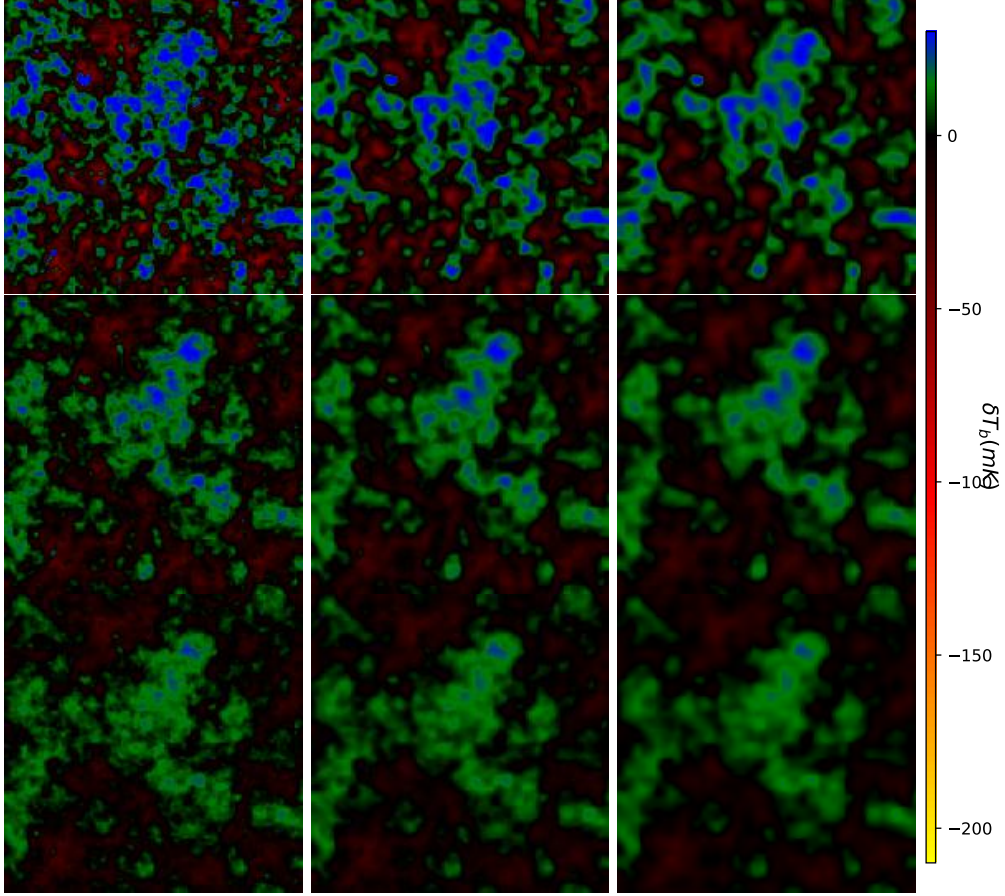
Fig. 2. An example of the "smoothed" mock 21-cm map in a $300\,\mathrm{cMpc}$-box at $(z, \bar{x}_{\mathrm{HI}}) = (12.73, 0.97)$. From left to right : beamsize $\Delta\theta = 1'$, $2'$, and $3'$. From top to bottom : frequency bandwidth $\Delta\nu = 0.2\,\mathrm{MHz}$, $1\,\mathrm{MHz}$, and $2\,\mathrm{MHz}$. We assume the Gaussian kernel for both beam and frequency smoothing.

of Figure 1 as a comparison). Additionally, smoothing in both parallel and perpendicular directions to the LOS wipes out the small-scale features. As a result, without a prior knowledge, the heavily smoothed 21-cm map at the lower-right panel of Figure 2 might be confused to be at the middle of EoR (e.g., $\bar{x}_{\mathrm{HI}}(z) \sim 0.5$). Note that, however, the spatial distribution of the smoothed 21-cm maps is Gaussian and highly non-Gaussian at $\bar{x}_{\mathrm{HI}} \simeq 1$ and $1 - \bar{x}_{\mathrm{HI}} \gg 0$, respectively. Therefore, it may be possible to distinguish between the smoothed 21-cm maps before and the middle of EoR by using n-point correlation functions or Minkowski functionals [39, 40], while such methods usually require precise measurement of δT_{b}.

After smoothing, we add the noise to the mock 21-cm maps, whose level depends on

the integration time and the configuration of radio telescope. Here, we adopt the sensitivity limit at [39], which uses the sensitivity calculation from [41] by assuming the SKA with core size of $\sim 1\,\mathrm{km}$. While the spatial and intensity distribution of noise and how it convolves with the signal also strongly depends on the configuration of radio telescope [42], here we adopt a somewhat simple model as follows. We calculate the root-mean-square (RMS) of noise level ($\sigma(\Delta T; \Delta\theta, \Delta\nu)$) by using 10-, 100-, 1000-, and 10000-hour integration times and produce the Gaussian white noise with given RMS level and random seeds (ϵ). Then we obtain the "noisy" 21-cm map (or, $\delta T_{\mathrm{b}}^{\mathrm{n}}(\boldsymbol{\theta}, \nu)$) by adding the noise to the smoothed 21-cm map,

$$\delta T_{\mathrm{b}}^{\mathrm{n}}(\boldsymbol{\theta}, \nu) = \delta T_{\mathrm{b}}^{\mathrm{s}}(\boldsymbol{\theta}, \nu) + \mathcal{W}_{\mathrm{G}}(\boldsymbol{\theta}; \sigma(\Delta T; \Delta\theta, \Delta\nu), \epsilon) \,. \tag{10}$$

Note that the mean value of the noisy 21-cm map remains zero.

Figure 3 shows the comparison between our sensitivity limit estimation and the RMS value of the smoothed mock 21-cm maps ($\delta T_{\mathrm{b,rms}}$) with various configurations of beamsize, frequency bandwidth, and integration time. Here, $\delta T_{\mathrm{b,rms}}/\sigma$ can be regarded as a proxy of the signal-to-noise ratio (SNR). As commented earlier, to fully utilize the advantages of some n-points or topological analysis methods, one needs to achieve SNR $\gg 1$. By assuming marginal choices of beamsize and frequency bandwidth, e.g., $\Delta\theta = 1' \sim 2'$ and $\Delta\nu = 1 \sim 2\,\mathrm{MHz}$, such SNR is possible only with $\gtrsim$ 10000-hour SKA integration time [39].

Figure 4 further shows how adding noise affects the same mock 21-cm maps to Figure 2, by assuming 1000-hour SKA integration time. At large beamsize and frequency bandwidth (e.g., bottom-right panel of Figure 4), the RMS value of the smoothed mock 21-cm maps is higher than the sensitivity limit. Also, the overall distribution of the noisy 21-cm map is similar to the smoothed 21-cm map. On the other hand, at small beamsize and frequency bandwidth (e.g., top-left panel of Figure 4), the RMS value is about an order of magnitude smaller than the sensitivity limit. As a result, sophisticated analysis methods that rely the detailed spatial distribution of $\delta T_{\mathrm{b}}^{\mathrm{n}}$ might find difficulties in such maps without an additional denoising process.

Note that, however, one can see a visual pattern of patches with generally positive and negative values of $\delta T_{\mathrm{b}}^{\mathrm{n}}$ at the top-left panel of Figure 4, even before any denoising process. Also, one can see that such visual pattern has similarity to the clearer patterns shown at other panels (i.e., noisy 21-cm maps with larger beamsize and frequency bandwidth).
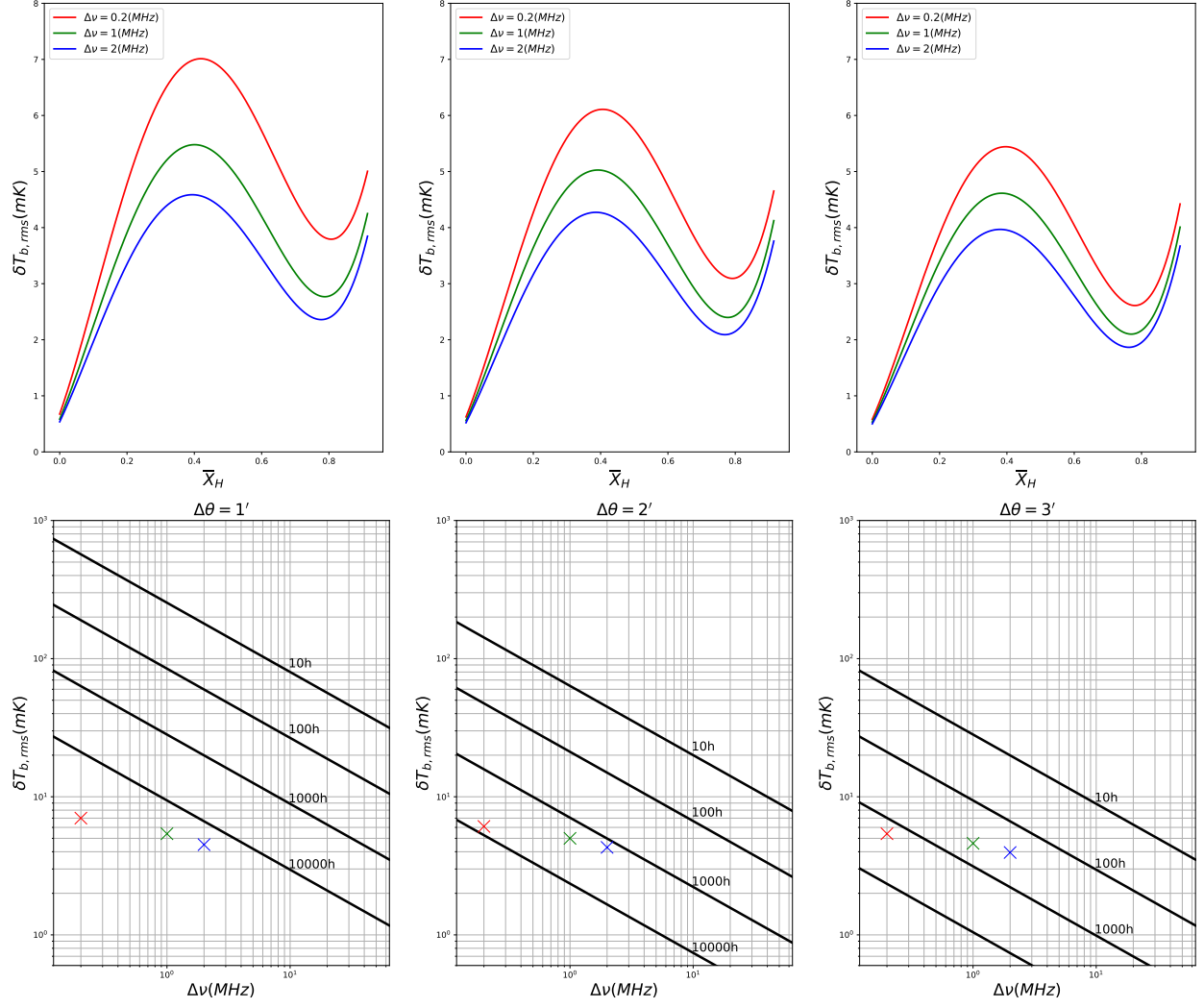
9

Fig. 3. Top panel: RMS of smoothed mock 21-cm maps as a function of mean neutral hydrogen fraction. Bottom panel: sensitivity limits with 10-, 100-, 1000-, and 10000-hour integration time of SKA (solid lines; from top to botom; [39]), compared to the maximum RMS of the smoothed mock 21-cm maps over time (peak value at the top panel; crosses). From left to right: $\Delta\theta = 1'$, $2'$, and $3'$. Colors: $\Delta\nu = 0.2\,\text{MHz}$(red), $1\,\text{MHz}$(green), and $2\,\text{MHz}$(blue).

However, although there exist some candidates such as binning or Gaussian smoothing, the mathematical definition of such "visual pattern" may not be straightforward. Note that such recognition of mathematically unclear visual patterns is one of the strongest specialties of deep learning technique, which we will see in the next section.
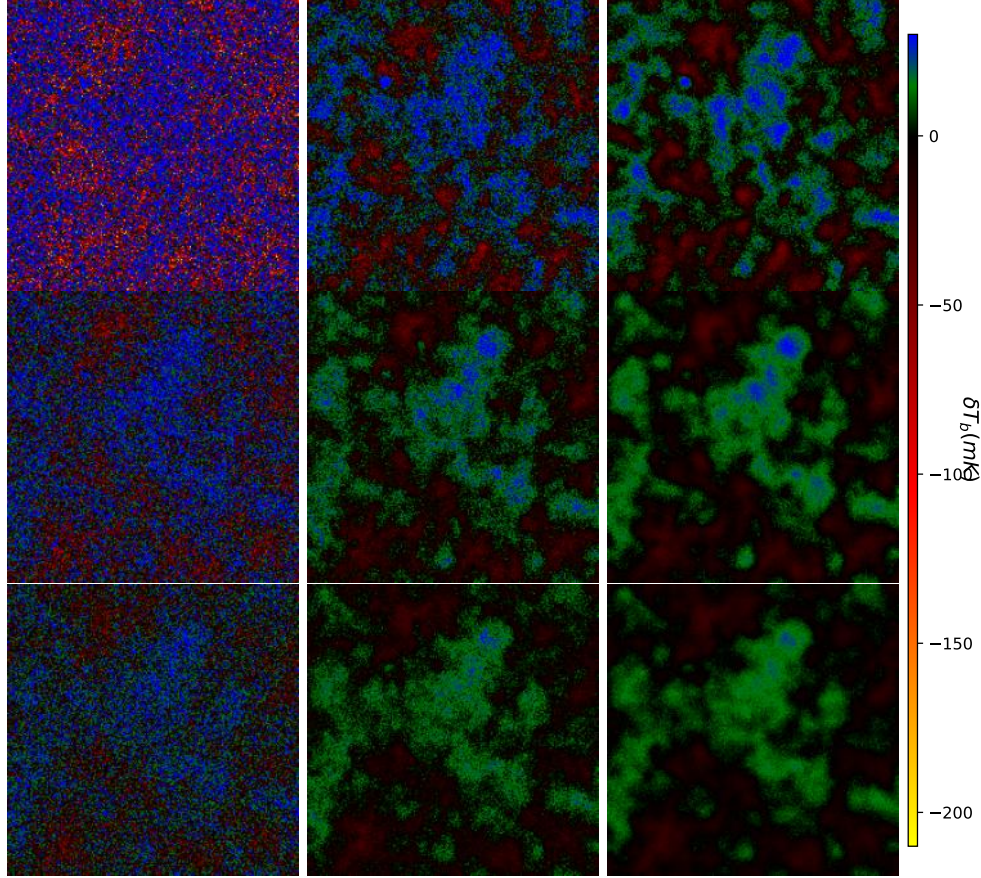
Fig. 4. Same as Figure 2, but for the "noisy" 21-cm maps after adding white Gaussian noise by assuming 1000-hour integration time of SKA.

III. DEEP LEARNING PREDICTION OF MEAN NEUTRAL HYDROGEN FRACTION

1. Deep Learning Architecture

We use the convolutional neural network (CNN) optimized to recognize visual patterns within the image dataset with efficiently lessening computational cost for expressing complex non-linear function for multi-dimensional input data. Figure 5 and Table 1 show our CNN architecture design. The entire network is divided into two parts, feature extraction and (fully-connected; FC) linear regression. All the filters in the first part of the architecture have the same size, $(3, 3)$, and are employed to extract the feature of the processed 21-cm map through a series of convolutions. Before each convolution layer, we use the max pooling (MaxPool) layer. The MaxPool layer helps to ease operations needed over the training by reducing the features resulting from convolution one, and can improve the result by keeping
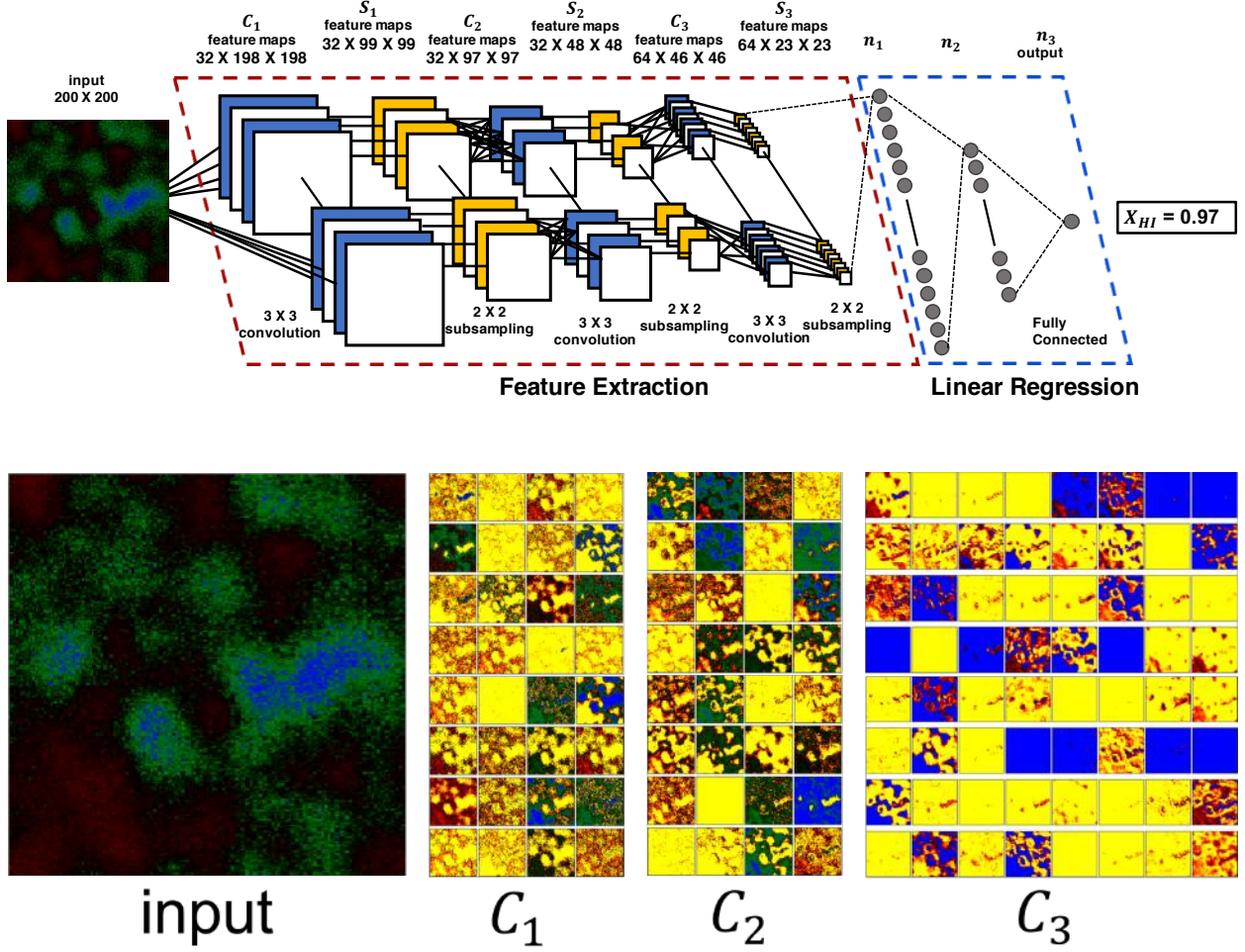
Fig. 5. Top panel: schematic architecture of convolutional neural network used for training. Bottom panel: layer-by-layer outputs of an example noisy 21-cm map input with $(z, \bar{x}_{\mathrm{HI}}, \Delta\theta, \Delta\nu, \Delta T) = (11.93, 0.97, 3', 0.2\,\mathrm{MHz}, 1000\,\mathrm{hours})$ during the feature extraction.

the network from getting over-fitted with a represented value (c.f. maximum or mean) of particular features over from the local regions, without numerous feature maps more than necessary. Finally, after the convolution and pooling layers, it is followed by a linear regression part, which consists of three FC layers. The FC layers employ the high-level features from the first part of the CNN and help to learn the network be close to have a desired output value through the last fully-connected layer activated by a linear function.

We apply the rectified linear unit (ReLU), which is defined as $\mathrm{ReLU}(s) = \max(0, s)$, as the activation functions of all convolutional and FC layers [43, 44]. The weights on the fully-connected layers are initialized by applying He uniform initialization [45], which

Table 1. Outline of convolutional neural network used for training.

Layer	# Filters	Filter Size, Stride	Padding	Output Dimension
Input	-	-	-	(3, 200, 200)
Conv2D-1	32	(3,3), (1,1)	valid	(32, 198, 198)
BatchNorm-1 (ReLU)	-	-	-	(32, 198, 198)
MaxPool2D-1	1	(2,2), (2,2)	valid	(32, 99, 99)
Conv2D-2	32	(3,3), (1,1)	valid	(32, 97, 97)
BatchNorm-2 (ReLU)	-	-	-	(32, 97, 97)
MaxPool2D-2	1	(2,2), (2,2)	valid	(32, 48, 48)
Conv2D-3	64	(3,3), (1,1)	valid	(64, 46, 46)
BatchNorm-3 (ReLU)	-	-	-	(64, 46, 46)
MaxPool2D-3	1	(2,2), (2,2)	valid	(64, 23, 23)
Flatten	-	-	-	33856
FC-1 (He-uniform)	-	-	-	64
BatchNorm-4 (ReLU)	-	-	-	64
FC-2 (He-uniform)	-	-	-	32
BatchNorm-5 (ReLU)	-	-	-	32
FC-3 (linear)	-	-	-	1

extract n_{in} samples from weights initialized from a uniform distribution between $-\sqrt{6/n_{\text{in}}}$ and $+\sqrt{6/n_{\text{in}}}$. The He uniform initialization for weights technique helps in reaching a global minimum of the cost function faster and more effectively.

Before each ReLU activation, we also add batch normalization (BatchNorm) layers [46]. For the situation where the inputs are fed with forwarding to the output layer, the distributions of inputs keep changing with each iterations because the weights of each layers are changed over the training of network. So the intermediate layers are hard to adapt its changing distributions of inputs. To prevent such problems, BatchNorm renormalize a mini-batch of inputs (x_i) into y_i as:

$$y_i = \gamma \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \,, \tag{11}$$

where μ and σ are the mean and standard deviation of x_i, ϵ is a small constant value intro-

Table 2. Colormap scheme used to generate input images from the noisy mock 21-cm maps. Colors for $\delta T_{\mathrm{b}}^{\mathrm{n}}$ values in between are determined with the linear interpolation.

$\delta T_{\mathrm{b}}^{\mathrm{n}}$ [mK]	Color Name	Red [0 ∼ 1]	Green [0 ∼ 1]	Blue [0 ∼ 1]
−210	Yellow	1	1	0
−105	Red	1	0	0
0	Black	0	0	0
15	Green	0	0.502	0
30	Blue	0	0	1

duced to avoid divergence, and γ and β are two additional learnable parameters introduced in the BatchNorm process. Not only BatchNorm can effectively complement some chronic problems that deep neural network has, such as gradient vanishing problem, but allow us to use much higher learning rate and perform as a regularizer for a part.

2. Training

For the CNN training, first of all, we produce the samples by splitting the noisy 21-cm maps with 200^3 grids into 67 slices of 200×200 pixels along the LOS direction. We calculate the mean neutral hydrogen fraction for each slice. Note that, while we only use $18 \sim 35$ snapshots between $z = 6 \sim 13$, $\bar{x}_{\mathrm{HI}}$ estimated from different slices at the same redshift may vary. As a result, for a given $(\Delta\theta, \Delta\nu, \Delta T)$, we have $1206 \sim 2345$ samples for each boxsize and high-resolution grid setup. Note that the difference of angular field of view from a fixed boxsize between $z = 6 \sim 13$ is less than 13%, which we will consider as nearly identical. We have also found that the performance is similar even if one uses 10318 samples from all 9 boxsize- and high-resolution-grid-setups, and therefore, we use all 10318 samples hereafter.

We produce the inputs of our CNN architecture by converting each 200×200 pixels of $\delta T_{\mathrm{b}}^{\mathrm{n}}$ slice into a 3-channel RGB image (i.e., an array with $(3, 200, 200)$ size) by applying a custom colormap scheme (see Table 2 for definition and Figure 6 for examples). We have also tested using the 200×200 slice directly as input and found no notable difference in performance.

We then split our samples into three sets: training set used for training the model pa-

rameters, validation set used for checking the training process, and test set for evaluating the performance of the trained model. For a given $(\Delta\theta, \Delta\nu, \Delta T)$, we split 10318 samples to 7428 training samples, 826 validation samples, and 2064 test samples. We randomly split the samples into three sets while keeping the distribution of $\bar{x}_{\mathrm{HI}}$ similar.

We compile our CNN model with Adam optimizer [47] with a learning rate 10^{-3}. Also, we adopt the mean squared error (MSE) as the loss function,

$$\mathrm{MSE} = \frac{1}{m} \sum_{i=1}^{m} \left(\bar{x}_{\mathrm{HI,pred}}^{i} - \bar{x}_{\mathrm{HI,true}}^{i} \right)^2 , \tag{12}$$

Here, m is the minibatch size for training, which is set to 8. We have also tested several minibatch sizes between 4 and 64 and found no notable difference in performance for $m \geq 8$.

We define an epoch as an iteration of fitting the model parameters from 7428 training samples and evaluating its training loss as well as validation loss from 826 validation samples. Since we set the minibatch size to 8, the number of minibatches per epoch is 929. For a single run of CNN training, we set the maximum number of epochs as 500. However, for saving the time and preventing the overfitting, we stop the training when the validation loss is not improved for 30 epochs, and the typical number of epochs becomes ~ 200. We then use the model at the epoch where the validation loss is minimized. We perform our training with Keras [48] with Tensorflow GPU version as a backend [49], and each run takes ~ 2 hours with a single NVIDIA V100 GPU card in a NVIDIA DGX-1 GPU platform.

Furthermore, we make use of the cross validation (CV) technique to test the stability of a certain deep learning model. In CV technique, the entire dataset is split into K-folds, and multiple CNN models are trained by using a certain fold as validation set and the remaining $(K-1)$-folds as training set. We use the shuffled 10-folded CV so that we train our network 5 times with the number of epoch and data having different distribution for each time. We have found that the validation loss from different models agrees within less than 75% deviation from their average.

IV. RESULTS

The bottom panel of Figure 5 shows how our CNN model processes the input noisy 21-cm map during the feature extraction. Numerous feature maps are similar to the black-and-white segmented images of the input map with different intensity thresholds and small-scale
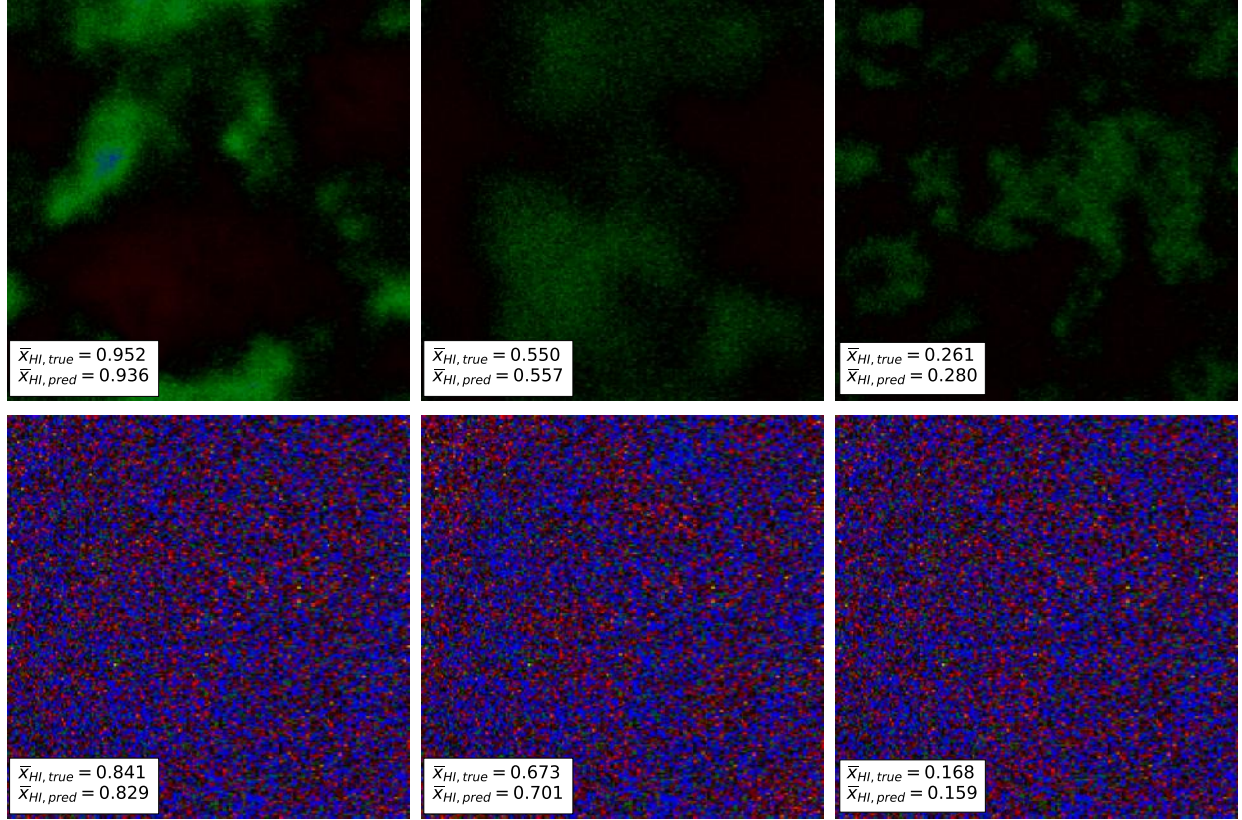
Fig. 6. The noisy 21-cm map, mean neutral hydrogen fraction, and its prediction from our CNN model for the two extreme choices of beamsize and frequency bandwidth. Top panel: extremely smoothed signal with $(\Delta\theta, \Delta\nu) = (3', 2\,\text{MHz})$. Bottom panel: extremely noisy signal with $(\Delta\theta, \Delta\nu) = (1', 0.2\,\text{MHz})$. We assume 1000-hour integration time of SKA.

noises. It means that our CNN model allows a similar strategy to the most straightforward estimation method of $\bar{x}_{\text{HI}}$, i.e., measuring the area fraction of zero $\delta T_{\text{b}}^0(\boldsymbol{\theta})$. Allowing different intensity thresholds is mainly because the value of $\delta T_{\text{b}}^{\text{s}}(\boldsymbol{\theta})$ at $\delta T_{\text{b}}^0(\boldsymbol{\theta}) = 0$ is uncertain, due to the renormalization of the mean value. Also, allowing different small-scale noises may be related to denoising the noisy 21-cm map in several ways. Note that the variations of intensity thresholds and small-scale noises depend on $(\Delta\theta, \Delta\nu, \Delta T)$.

Figure 6 shows the noisy mock 21-cm maps for the two extreme choices of beamsize and frequency bandwidth. The top panel of Figure 6 corresponds to large values of beamsize and frequency bandwidth. Although the observational noise is relatively small, the entire 21-cm map is too smoothed so that the 21-cm maps from three characteristic epochs ($\bar{x}_{\text{HI}} \sim 1$, 0.5, and 0) looks similar. On the other hand, the bottom panel that comes from small
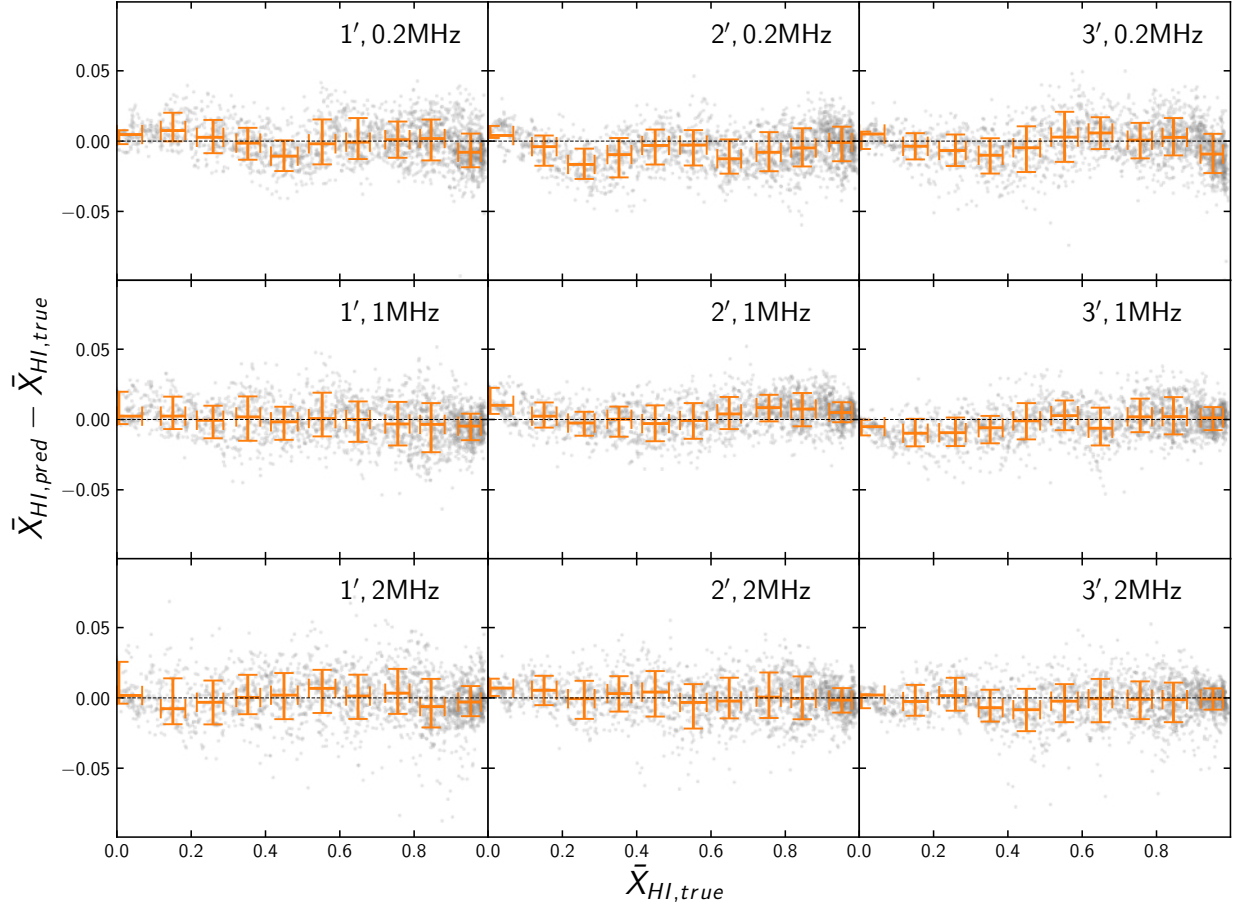
Fig. 7. Prediction error of the mean neutral hydrogen fraction from our CNN method as a function of its truth value. Dots: 2064 test samples. Error bars: median and 68% certainty level from 10 equally spaced bins of $\bar{x}_{\mathrm{HI,true}}$. From left to right: $\Delta\theta = 1'$, $2'$, and $3'$. From top to bottom: $\Delta\nu = 0.2\,\mathrm{MHz}$, $1\,\mathrm{MHz}$, and $2\,\mathrm{MHz}$. We assume 1000-hour integration time of SKA.

$(\Delta\theta, \Delta\nu)$ has more observational noise than the signal itself, which also makes the 21-cm maps from three epochs hard to visually distinguish. Interestingly, our CNN method can clearly distinguish such noisy 21-cm maps at different epochs with the prediction error of $\bar{x}_{\mathrm{HI}}$ less than ~ 0.03.

Figure 7 and Table 3 show the prediction power of our CNN method for different choices of beamsize and frequency bandwidth by assuming 1000-hour integration time of SKA. Overall, the predicted mean neutral hydrogen fraction has a good agreement of its truth value. In the case $(\Delta\theta, \Delta\nu) = (2', 1\,\mathrm{MHz})$, the median value of the absolute prediction error $|\bar{x}_{\mathrm{HI,pred}} - \bar{x}_{\mathrm{HI,true}}|$ is about 0.008, while the upper bound of its 95% certainty level is

17

Table 3. Summary of the absolute prediction error $|\bar{x}_{\mathrm{HI,pred}} - \bar{x}_{\mathrm{HI,true}}|$ from our CNN method. We assume 1000-hour integration time of SKA. The best and worst cases are marked as **bolded** and <u>underlined</u>, respectively.

Beamsize	Bandwidth	Median	1σ-high	2σ-high	Maximum
	0.2 MHz	0.0090	0.0198	0.0409	0.0956
$1'$	1 MHz	0.0087	0.0197	0.0474	0.0633
	2 MHz	<u>0.0100</u>	<u>0.0227</u>	<u>0.0687</u>	0.0871
	0.2 MHz	0.0092	0.0202	0.0490	0.0590
$2'$	1 MHz	0.0082	0.0169	**0.0349**	**0.0517**
	2 MHz	0.0082	0.0184	0.0544	0.0643
	0.2 MHz	0.0091	0.0212	0.0504	<u>0.0980</u>
$3'$	1 MHz	0.0075	**0.0162**	0.0429	0.0524
	2 MHz	**0.0075**	0.0178	0.0616	0.0772

about 0.05. Note that such accuracy on reconstructing history of $\bar{x}_{\mathrm{HI}}(z)$ is comparable to the forecast presented in [37, 50]. Therefore, although it is beyond our scope, one could expect that our CNN method might provide a similar constraint power to the reionization parameters.

While our CNN method can predict $\bar{x}_{\mathrm{HI}}(z)$ reasonably well with various choices of beamsize and frequency bandwidth in overall, there exists non-negligible difference of performance between them. For example, configurations with small beamsize and frequency, especially those with $\Delta\nu = 0.2\,\mathrm{MHz}$, tend to have a systematic bias with a wavy shape. This might mean that, due to the significant noise level, the CNN model tends to behave a bit similar to classification rather than pure regression. On the other hand, those with large beamsize and frequency, especially with $\Delta\nu = 2\,\mathrm{MHz}$, tend to have a large scatter from zero at the tail of probability distribution. This could be possible if such outlier maps contain many small ionized bubbles (or the small tip of bubbles) that contribute non-negligible fraction to $\bar{x}_{\mathrm{HI}}$.

In Figure 8 and Table 4, we further test the prediction power of our CNN method with 100-hour integration time of SKA. As the noise level is about an order of magnitude higher than the actual signal, there exists a significant amount of deviation even in large beamsize
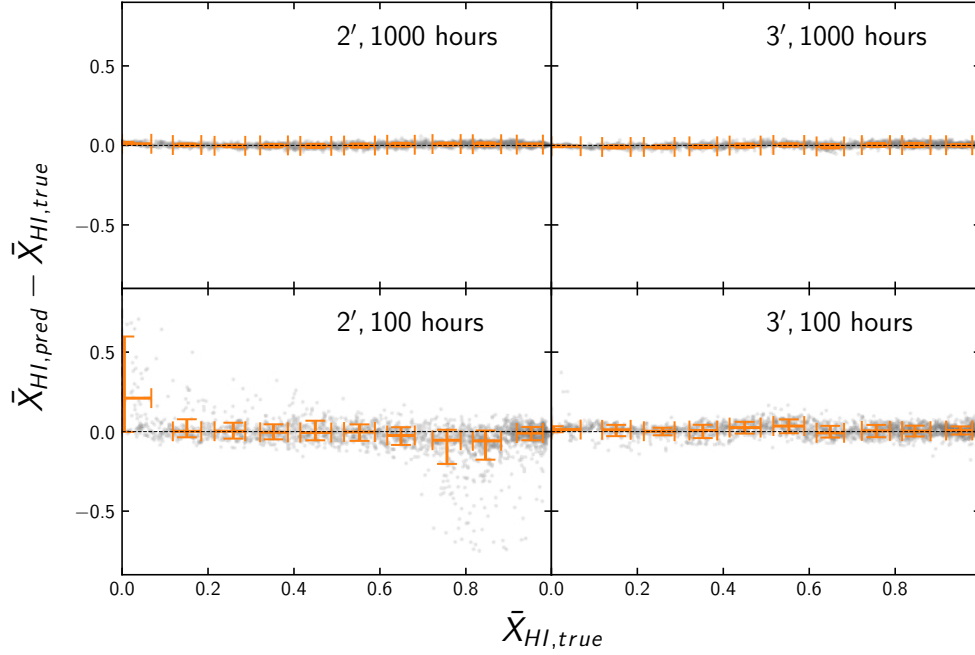
Fig. 8. Same as Figure 7, but by varying $(\Delta\theta, \Delta T)$ while fixing $\Delta\nu = 1\,\mathrm{MHz}$. From left to right: $\Delta\theta = 2'$ and $3'$. From top to bottom: $\Delta T = 1000$ hours and 100 hours.

Table 4. Same as Table 3, but by varying $(\Delta\theta, \Delta T)$ while fixing $\Delta\nu = 1\,\mathrm{MHz}$.

Beamsize	Integration Time	Median	1σ-high	2σ-high	Maximum
2′	1000 hours	0.0082	0.0169	**0.0349**	**0.0517**
	100 hours	<u>0.0416</u>	<u>0.118</u>	<u>0.746</u>	<u>0.821</u>
3′	1000 hours	**0.0075**	**0.0162**	0.0429	0.0524
	100 hours	0.0227	0.0530	0.163	0.367

$\Delta\theta \gtrsim 3'$. Furthermore, for $\Delta\theta \lesssim 2'$ it clearly shows two populations with $(\bar{x}_{\mathrm{HI,true}}, \bar{x}_{\mathrm{HI,pred}}) \approx (0, 1)$ and vice versa. It means that the CNN model fails to distinguish the 21-cm maps before and the late stage of EoR — in other words, the number of "classification" category that our CNN model with large noise level tends to find becomes extremely low. Note that the SNR at $(\Delta\theta, \Delta\nu, \Delta T) = (2', 1\,\mathrm{MHz}, 100\,\mathrm{hours})$ is similar to that at $(1', 1\,\mathrm{MHz}, 1000\,\mathrm{hours})$, where the prediction accuracy of $\bar{x}_{\mathrm{HI}}$ is reasonably good. It emphasizes that, since the distribution of 21-cm map is highly non-Gaussian in general, the SNR (or, $\delta T_{\mathrm{b,rms}}/\sigma$) alone cannot fully determine the prediction power of our CNN method. Further study on the deep learning architecture or the combination with other analysis methods might be helpful to enhance

19

the prediction power of $\bar{x}_{\mathrm{HI}}$, which is beyond our scope.

V. CONCLUSIONS

In this paper, we introduced a novel convolutional neural network (CNN)-based deep learning technique for the prediction of the mean neutral hydrogen fraction ($\bar{x}_{\mathrm{HI}}$) during the Epoch of Reionization (EoR) from the two-dimensional tomography of the redshifted 21-cm maps of differential brightness temperature. We first simulated the undistorted 21-cm maps of 200^3 uniform grids with $50 \sim 300\,\mathrm{cMpc}$ boxsize between $z = 6 \sim 13$ by using a semi-numerical simulation code 21cmFAST. We then applied various instrumental conditions to the dataset by controlling beamsize and frequency bandwidth suitable for the upcoming Square Kilometre Array (SKA) to produce the noisy mock 21-cm maps. After converting the noisy mock 21-cm maps into RGB images, we applied them as inputs of our CNN architecture to predict the corresponding values of $\bar{x}_{\mathrm{HI}}$.

The main results of this paper can be summarized as follows.

1. Our CNN method has a capability to predict $\bar{x}_{\mathrm{HI}}$ from the raw noisy 21-cm maps even when the overall signal-to-noise ratio (SNR) is less than unity, depending on the radio survey configuration. It is because the 21-cm maps during the EoR is highly non-Gaussian in general, so that there exists many additional features than just SNR that the CNN method can utilize.

   In a similar reason, the overall performance of our CNN method depends on the combination of radio survey configuration, such as beamsize, frequency bandwidth, and telescope integration time, rather than just SNR.

2. For survey configuration with low-SNR (e.g., SNR $\lesssim 1$), there exists a systematic bias with a wavy shape between the CNN prediction of $\bar{x}_{\mathrm{HI}}$ and its truth value. This might mean that the CNN method becomes closer to the classifier of discrete categories rather than regressor of continuous value, mainly due to the high noise level.

   On the other hand, survey configuration with large smoothing (e.g., large beamsize and frequency bandwidth) may allow large scatter of prediction error, partly due to the slices containing large number of small ionized bubbles.

3. If one uses 1000-hour integration of SKA, beamsize and frequency bandwidth $\Delta\theta \simeq 2' \sim 3'$ and $\Delta\nu \simeq 1\,\mathrm{MHz}$ can be regarded as an optimal configuration for the prediction of $\bar{x}_{\mathrm{HI}}$ with the CNN method. With such configuration, median and the 2σ-upper bound of the absolute prediction error $|\bar{x}_{\mathrm{HI,pred}} - \bar{x}_{\mathrm{HI,true}}|$ are ~ 0.008 and 0.04, respectively.

Although our work can successfully reconstruct the reionization history during the EoR by incorporating some of the difficulties from actual radio observations, we have found there exists plenty of rooms for improvement. For example, for more realistic performance test, it might be useful to adopt a detailed information of beam shape and noise addition from future radio surveys, as well as Galactic and extragalactic HI foreground [42]. Also, including 21-cm maps with varying cosmological and reionization parameters would be helpful to understand how our reconstruction of $\bar{x}_{\mathrm{HI}}(z)$ depends on such parameters.

REFERENCES

[1] J. Miralda-Escudé, Science **300**, 1904 (2003), astro-ph/0307396.

[2] S. R. Furlanetto and S. P. Oh, Mon. Not. R. Astron. Soc. **363**, 1031 (2005), astro-ph/0505065.

[3] X. Fan, M. A. Strauss, R. H. Becker, R. L. White, J. E. Gunn, G. R. Knapp, G. T. Richards, D. P. Schneider, J. Brinkmann, and M. Fukugita, Astron. J. **132**, 117 (2006), astro-ph/0512082.

[4] O. Zahn, C. L. Reichardt, L. Shaw, A. Lidz, K. A. Aird, B. A. Benson, L. E. Bleem, J. E. Carlstrom, C. L. Chang, H. M. Cho, et al., Astrophys. J. **756**, 65 (2012), 1111.6386.

[5] Planck Collaboration, R. Adam, N. Aghanim, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, et al., Astron. Astrophys. **596**, A108 (2016), 1605.03507.

[6] M. J. Mortonson and W. Hu, Astrophys. J. **672**, 737 (2008), 0705.1132.

[7] K. Ahn, I. T. Iliev, P. R. Shapiro, G. Mellema, J. Koda, and Y. Mao, Astrophys. J. Lett. **756**, L16 (2012), 1206.5007.

[8] C. H. Heinrich, V. Miranda, and W. Hu, Phys. Rev. D **95**, 023513 (2017), 1609.04788.

[9] S. J. Tingay, R. Goeke, J. D. Bowman, D. Emrich, S. M. Ord, D. A. Mitchell, M. F. Morales, T. Booler, B. Crosse, R. B. Wayth, et al., Publ. Astron. Soc. Aust. **30**, e007 (2013), 1206.6945.

[10] G. Paciga, T.-C. Chang, Y. Gupta, R. Nityanada, J. Odegova, U.-L. Pen, J. B. Peterson, J. Roy, and K. Sigurdson, Mon. Not. R. Astron. Soc. **413**, 1174 (2011), 1006.1351.

[11] M. P. van Haarlem, M. W. Wise, A. W. Gunst, G. Heald, J. P. McKean, J. W. T. Hessels, A. G. de Bruyn, R. Nijboer, J. Swinbank, R. Fallows, et al., Astron. Astrophys. **556**, A2 (2013), 1305.3550.

[12] A. R. Parsons, D. C. Backer, G. S. Foster, M. C. H. Wright, R. F. Bradley, N. E. Gugliucci, C. R. Parashare, E. E. Benoit, J. E. Aguirre, D. C. Jacobs, et al., Astron. J. **139**, 1468 (2010), 0904.2334.

[13] D. R. DeBoer, A. R. Parsons, J. E. Aguirre, P. Alexander, Z. S. Ali, A. P. Beardsley, G. Bernardi, J. D. Bowman, R. F. Bradley, C. L. Carilli, et al., Publ. Astron. Soc. Pac. **129**, 045001 (2017), 1606.07473.

[14] P. E. Dewdney, P. J. Hall, R. T. Schilizzi, and T. J. L. W. Lazio, IEEE Proceedings **97**, 1482 (2009).

[15] L. Koopmans, J. Pritchard, G. Mellema, J. Aguirre, K. Ahn, R. Barkana, I. van Bemmel, G. Bernardi, A. Bonaldi, F. Briggs, et al., in *Advancing Astrophysics with the Square Kilometre Array (AASKA14)* (2015), p. 1, 1505.07568.

[16] S. R. Furlanetto, S. P. Oh, and F. H. Briggs, Physics Reports **433**, 181 (2006), astro-ph/0608032.

[17] M. F. Morales and J. S. B. Wyithe, Annu. Rev. Astron. Astrophys. **48**, 127 (2010), 0910.3010.

[18] S. Zaroubi, *The Epoch of Reionization* (2013), vol. 396 of *Astrophysics and Space Science Library*, p. 45.

[19] T. Di Matteo, R. Perna, T. Abel, and M. J. Rees, Astrophys. J. **564**, 576 (2002), astro-

ph/0109241.

[20] M. Zaldarriaga, S. R. Furlanetto, and L. Hernquist, Astrophys. J. **608**, 622 (2004), astro-ph/0311514.

[21] F. H. Briggs and J. Kocz, Radio Science **40**, RS5S02 (2005).

[22] X. Wang, M. Tegmark, M. G. Santos, and L. Knox, Astrophys. J. **650**, 529 (2006), astro-ph/0501081.

[23] C. Schaefer, M. Geiger, T. Kuntzer, and J. P. Kneib, Astron. Astrophys. **611**, A2 (2018), 1705.07132.

[24] N. Gillet, A. Mesinger, B. Greig, A. Liu, and G. Ucci, Mon. Not. R. Astron. Soc. **484**, 282 (2019), 1805.02699.

[25] J. Seiler, A. Hutter, M. Sinha, and D. Croton, Mon. Not. R. Astron. Soc. **487**, 5739 (2019), 1902.01611.

[26] J. Chardin, G. Uhlrich, D. Aubert, N. Deparis, N. Gillet, P. Ocvirk, and J. Lewis, Mon. Not. R. Astron. Soc. **490**, 1055 (2019), 1905.06958.

[27] F. List and G. F. Lewis, Mon. Not. R. Astron. Soc. **493**, 5913 (2020), 2002.07940.

[28] H. J. Hortúa, L. Malago, and R. Volpi, arXiv e-prints arXiv:2005.07694 (2020), 2005.07694.

[29] P. La Plante and M. Ntampaka, Astrophys. J. **880**, 110 (2019), 1810.08211.

[30] H. Shimabukuro and B. Semelin, Mon. Not. R. Astron. Soc. **468**, 3869 (2017), 1701.07026.

[31] S. Hassan, A. Liu, S. Kohn, and P. La Plante, in *The 34th Annual New Mexico Symposium*, edited by A. D. Kapinska (2018), p. 7.

[32] A. Mesinger, S. Furlanetto, and R. Cen, Mon. Not. R. Astron. Soc. **411**, 955 (2011), 1003.3878.

[33] Planck Collaboration, N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, et al., arXiv e-prints arXiv:1807.06209 (2018), 1807.06209.

[34] J. R. Bond, S. Cole, G. Efstathiou, and N. Kaiser, Astrophys. J. **379**, 440 (1991).

[35] S. R. Furlanetto, M. Zaldarriaga, and L. Hernquist, Astrophys. J. **613**, 1 (2004), astro-ph/0403697.

[36] Y. B. Zel'Dovich, Astron. Astrophys. **500**, 13 (1970).

[37] J. Park, A. Mesinger, B. Greig, and N. Gillet, Mon. Not. R. Astron. Soc. **484**, 933 (2019), 1809.08995.

[38] G. Mellema, I. T. Iliev, U.-L. Pen, and P. R. Shapiro, Mon. Not. R. Astron. Soc. **372**, 679

(2006), astro-ph/0603518.

[39] S. E. Hong, K. Ahn, C. Park, J. Kim, I. T. Iliev, and G. Mellema, Journal of Korean Astronomical Society **47**, 49 (2014), 1008.3914.

[40] Y. Wang, C. Park, Y. Xu, X. Chen, and J. Kim, Astrophys. J. **814**, 6 (2015), 1510.01404.

[41] I. T. Iliev, E. Scannapieco, H. Martel, and P. R. Shapiro, Mon. Not. R. Astron. Soc. **341**, 81 (2003), astro-ph/0209216.

[42] J. Asorey, D. Parkinson, F. Shi, Y.-S. Song, K. Ahn, J. Kim, J. Yao, L. Zhang, and S. Zuo, arXiv e-prints arXiv:2001.00833 (2020), 2001.00833.

[43] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, Nature **405**, 947 (2000).

[44] X. Glorot, A. Bordes, and Y. Bengio, in *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (2011), pp. 315–323.

[45] K. He, X. Zhang, S. Ren, and J. Sun, arXiv e-prints arXiv:1502.01852 (2015), 1502.01852.

[46] S. Ioffe and C. Szegedy, arXiv e-prints arXiv:1502.03167 (2015), 1502.03167.

[47] D. P. Kingma and J. Ba, arXiv e-prints arXiv:1412.6980 (2014), 1412.6980.

[48] F. Chollet et al., *Keras*, https://keras.io (2015).

[49] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., *TensorFlow: Large-scale machine learning on heterogeneous systems* (2015), software available from tensorflow.org, URL https://www.tensorflow.org/.

[50] A. Liu and A. R. Parsons, Mon. Not. R. Astron. Soc. **457**, 1864 (2016), 1510.08815.