# Modelling Cosmological Reionisation and its Observational Signatures

**Michele Bianco**

Submitted for the degree of Doctor of Philosophy
University of Sussex
June 2021

# Declaration

I hereby declare that this thesis has not been and will not be submitted in whole or in part to another University for the award of any other degree. Parts of this thesis have been undertaken in collaboration with other researchers. Where this is the case, we acknowledge their work at the beginning of the relevant chapter.

Signature:

Michele Bianco

UNIVERSITY OF SUSSEX

Michele Bianco, Doctor of Philosophy

Modelling Cosmological Reionization and its Observational Signatures

## Summary

The Epoch of Reionization is an important period in studying structure formation and evolution of our Universe. The first luminous objects, which may have been star-forming galaxies and quasi-stellar objects, influenced later-day structures formation and evolution. These bright objects produced enough ultra-violet radiation to alter the nature of the host and propagated out into the intergalactic medium. These energetic photons transitioned our Universe from a cold and neutral state to ultimately a hot and ionised state. This interesting period is one of the least understood epochs in the Universe evolution due to the lack of direct observations. The redshifted 21-cm signal of neutral hydrogen can be used as an observable sign of reionisation. The upcoming Square Kilometre Array telescope will be sensitive enough to detect the 21-cm signal and produce images of its spatial distribution throughout reionisation.

This research focuses on improving numerical methods and develop new techniques for understanding and interpreting future observational evidence. Our simulations will play a crucial role and provide numerical support for the upcoming experiments. We proposed a new approach that correctly quantifies the effect of local recombinations on the scale below the large numerical simulation resolution. We present a more general model for the sub-grid gas clumping, depending on the local density. I improved the latter method with an empirical stochastic model based on high-resolution N-body simulation results, and the relevant fluctuations are fully resolved. Moreover, we developed a stable and reliable convolutional neural network, which can identify neutral and ionised regions from noisy 21-cm image observations. The network can identify the regions of interest with greater precision and is less sensitive to the limitation of previous methods. We successfully recover the signal for different instrumental noise levels based on the intensity contrast in the 21-cm signal and from ionised regions simulation independent pattern.

*1st September 2021*

# Acknowledgements

During the PhD. I met many people that, in a way or another, shaped my path. I want to start by thanking my supervisor Prof. Ilian T. Iliev, for his constant help and guidance over the years. Thanks for your patience with my (initially) poor English writing and ensuring that I was always granted remarkable amounts of computing time on the best performance computers.

I would also like to thanks Prof. Garrelt Mellema. Despite your injury during my visit to Stockholm University, you did your best to stay in contact and endorsed my work with Sambit, I am deeply thankful. Thank you for welcoming me, even after my departure, to your weekly Journal Club. I always considered these meetings a great resource of wisdom and knowledge. Many thanks also to Dr Benedetta Ciardi. With your persistence and motivation, you keep me involved in your research team despite the strange time during the pandemic. Our regular meeting has helped me cope with the stress due to lockdown and kept my mind busy. It has always been helpful and interesting to get advice from you.

I want to pay special thanks also to Dr Sambit K. Giri. Discussing with you has always brought new ideas, and your genuine interest in your work was of great inspiration for me. Thank you for your help at Stockholm University. I can comfortably say that my research would have suffered dramatically without your support. Thank you also for the fantastic experience in India and the Netherlands.

During my time in Sussex, I met wonderful people. In particular, my office mates; Sunayana Bhargava, Azizah Hosein, Carlos Vergara and Rose Coogan. They reminded me to take some time off and not to work too hard. I sincerely enjoyed our pizza and poker nights. Special mention also goes to Luke Conaboy. You have been very kind and always ready to help. I enjoyed our discussions and our shared love for a good cup of coffee. Despite my absence and the global pandemic, I have been fortunate to have met and kept in contact with many great people from the University of Sussex. I want to thank Aswin Payyoor Vijayan, Chris Lovell, William Roper, Reese Wilkinson, Dimitrios Irodotou, Jussi Kuusisto, Dan Pryer, David Turner, Itzi Aldecoa and Edward Salakpi. Thank you for your

enthusiasm and the multiple occasions that reminded me of the beauty of being part of a community.

Many other people have helped in my research with helpful and delightful discussions; Raghu Ghara, Ivelin Georgiev, Hannah Ross, Enrico Garaldi, Leon Koopmans, Keri Dixon, Anastasia Fialkov, Yi Mao, Hyunbae Park, Kyungjin Ahn and Paul Shapiro to name a few.

Un ringraziamento speciale va alla mia famiglia. A mio padre, Domenico, che mi ha sempre insegnato l'importanza di avere un lato pratico delle cose e il sacrificio verso i propri cari. A mia madre, Paola, che mi ha sempre assistito negli studi e nella vita, insengnandomi la virtù della pazienza dell'amore materno. Ai miei nonni di Napoli, Anna e Michele, che mi hanno insegnato l'amore incondizionato per la famiglia e l'importanza di fare sempre la cosa giusta, mantenendo la parola data. Ai mei nonni svizzeri, Maria e Luciano, che mi hanno accudito in modo da crescere sano e forte, insegnadomi la virtù dell'umiltà e l'importanza di studiare le lingue.

J'aimerais remercier aussi ma copine Estelle Louet. Partir pour l'Angleterre et vivre à distance a été la chose la plus difficile pendant ces trois années. Sans ta patience, ton humour et ton soutien quotidien je n'aurais pas pu réussir dans mes études et comme être humain. J'attends maintenant avec joie de commencer notre vie ensemble.

*” The effort to understand the universe is one of the very few things which lifts human life a little above the level of farce and gives it some of the grace of tragedy. ”*

**Steven Weinberg (1993), The First Three Minutes**

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Introduction to Cosmology

The expansion of our Universe was first observed by Hubble (1929), which determined the velocity-distance relation of extra-galactic nebulae and deduced the nowadays know Hubble constant $H_0$, the rate at which galaxies are receding from one another (Hubble & Humason, 1931). It was proven by comparing the observed wavelength $\lambda_{\text{obs}}$ of the spectral lines emitted by these distant objects with the corresponding rest-frame wavelength $\lambda_{\text{ref}}$.

$$1 + z = \frac{\lambda_{\text{obs}}}{\lambda_{\text{ref}}} \tag{1.1.1}$$

This effect is known as the *cosmological Doppler redshift*, and the redshift $z$ is a proportional factor that quantifies the line of sight velocity of the receding objects. This expansion means that early in the history of the Universe, the distances between galaxies were smaller than nowadays. Therefore, based on this evidence, at the time, it was concluded that our Universe started from a single point of infinity density, an initial singularity known as the Big Bang. However, later observations (e.g. Riess et al., 1998; Perlmutter et al., 1999; Riess et al., 2007) indicated that the Universe underwent a fast exponential expansion during its early stage, referred to as the cosmic inflation. This accelerated expansion of the initial singularity ultimately lays the basis of density perturbation theory and the subsequent structure formation in our Universe. According to the standard theory of the Big Bang, the early Universe was extremely hot and dense such that photons were constantly scattered and remitted by the matter that was kept completely ionised. Nevertheless, the most significant evidence for the theory of the Big Bang was firstly detected by Penzias & Wilson (1965). By measuring the noise temperature at the zenith with the 6 metres long Holmdel Horn antenna, they measured the residual radiation from the last scattering surface of the primordial soup. At that stage, the Universe cooled down and

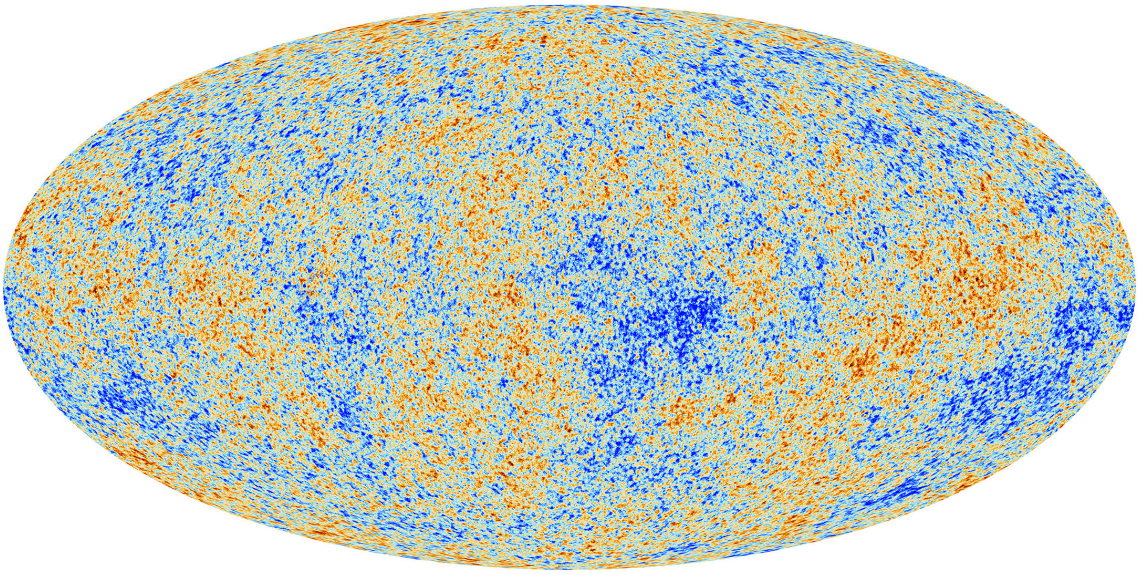

Figure 1.1: The Cosmic Microwave Background observed with the Planck telescope that shows the full sky temperature fluctuations. *Image credits: ESA/Planck (2013)*

therefore expanded enough, allowing photons to escape. This period is often called the *recombination era*, it occurred approximately at $z \sim 1100$, and the relic radiation is referred to as the Cosmic Microwave Background (CMB). The discovery is conceivably the most important revelation of modern cosmology, and yet it was detected almost by accident. The CMB is isotropic and has a black body spectrum that peak at a temperature today of about $T_{\rm CMB} = 2.726\,{\rm K}$. Recent observations improved the first estimation and were able to detect local fluctuations in the order of $\delta T/T \approx 10^{-5}$. In figure 1.1 we show the high-resolution full-sky CMB anisotropies provided by Planck space telescope (Ade et al., 2014), where red and blue areas indicate the over- and under-dense regions, respectively. The detection of the CMB proved the validity of the *cosmological principle* on a sufficiently large scale of a few hundred Megaparsecs[1]. This assumption considers our Universe as a perfect fluid, homogeneous (uniform in composition) and isotropic (uniformity in all directions), with a given density $\rho$ and pressure $p$.

### 1.1.1 ΛCDM Cosmology Model

In the second decade of the twenty century, the work of Einstein (1916) provides the foundation for general relativity and it explained the commutation between mass and gravity (e.g. Weinberg, 1972; Wald, 1984; Hartle, 2003). Under the cosmological principle assump-

[1] In cosmology, the standard unit length is the parsec (pc). It is defined as the distance between the Earth and an astronomical object with a parallax of one arcsecond, $1\,{\rm pc} = 1\,{\rm au}/tan(1'') = 3.09 \times 10^{16}\,{\rm m}$.

tion, it was formulated that the Friedmann-Lemaître-Robertson-Walker (FLRW) metric provides an exact solution to Einstein general relativity field equations (Friedmann, 1922). The FLRW metric considers that the physical distance between two points separated by a distance $ds$, in hyper-spherical coordinates, is given by (e.g. Peacock, 1998)

$$ds^2 = -c^2 dt^2 + a^2(t)\,\left(dr^2 + S_{\rm k}^2(r)\,d\Omega^2\right) \tag{1.1.2}$$

Here, $r$ and $t$ are the radial comoving coordinates and time coordinates, respectively. Therefore, the positive term on the right-hand side is the three-dimensional metric and $d\Omega^2 = d\theta^2 + \sin^2(\theta)\,d\phi^2$ the differential solid angle. Here $a \equiv \frac{1}{1+z}$ is the *cosmic scale factor*, and it is a time-dependent factor that characterises the universe expansion after the Big Bang. By definition, the scale factor today is normalised, $a(t_0) \equiv a_0 = 1$. The factor $S_{\rm k}$ takes into account the spatial curvature contribution, with $\rm k = 0$ if flat and $\pm 1$ for open or closed space, respectively.

$$S_{\rm k}(r) = \begin{cases} \sqrt{k}^{-1}\,\sin(r\sqrt{k}) & , \ \rm k > 0 \\ r & , \ \rm k = 0 \\ \sqrt{|k|}^{-1}\,\sin(r\sqrt{|k|}) & , \ \rm k < 0 \end{cases} \tag{1.1.3}$$

With the FLRW metric as the solution of general relativity field equations, we can derive the *Friedman Equation* for modelling the evolution of an expanding, homogeneous and isotropic universe based on its constituents and the scale factor $a$ (e.g. Dodelson, 2003)

$$H^2(t) = \left[\frac{\dot{a}(t)}{a(t)}\right]^2 = \frac{8\pi G\,\rho + \Lambda\,c^2}{3} - \frac{k\,c^2}{a^2} \tag{1.1.4}$$

Here $H \equiv H(t)$ is known as the Hubble parameter, $G$ is the gravitational constant, $\rho$ indicates the energy density of the different components, and $\Lambda$ is the cosmological constant that raises from the general relativity equations. The latter variable gauges the vacuum pressure, $p_\Lambda = -c^2\,\rho_\Lambda$, and it implies the presence of *dark energy*. Initially, the cosmological constant was introduced to counterbalance gravitation in order to obtain an everlasting Universe. However, only long after Hubble's discovery, in the late 90s, after precise measurements of supernovae luminosity distance that it started to be considered as a non-zero value (Riess et al., 1998; Schmidt et al., 1998; Perlmutter et al., 1999). For a flat geometric universe $k = 0$, we can define the critical density of the Universe as a function of the evolving cosmic time, as

$$\rho_{\rm c}(t) = \frac{3\,H^2(t)}{8\pi\,G} \tag{1.1.5}$$

It is convenient to express the relative energy density $\rho_{\rm i}$, for the constituent $i$, as a ratio of the critical density, equation (1.1.5). Therefore, we can define the dimensionless density

Table 1.1: Cosmological parameter measured from four different observation experiments.

| Parameters | WMAP5 | Planck (2018) | Planck+SNeIa | DES3 |
|---|---|---|---|---|
| $H_0$ [km s$^{-1}$ Mpc$^{-1}$] | $71.90^{+2.60}_{-2.70}$ | $67.66 \pm 0.42$ | $74.20 \pm 1.40$ | $68.10^{+0.40}_{-0.30}$ |
| $\Omega_\mathrm{b}$ | $0.0441 \pm 0.0030$ | $0.0489 \pm 0.0003$ | $0.0402 \pm 0.0004$ | $0.0487^{+0.0005}_{-0.0004}$ |
| $\Omega_\mathrm{m}$ | $0.256 \pm 0.012$ | $0.311 \pm 0.005$ | $0.252 \pm 0.003$ | $0.306^{+0.004}_{-0.005}$ |
| $\Omega_\Lambda$ | $0.742 \pm 0.030$ | $0.698 \pm 0.006$ | $0.748 \pm 0.003$ | $0.715^{+0.004}_{-0.005}$ |

parameters $\Omega_\mathrm{i} \equiv \rho_\mathrm{i}/\rho_\mathrm{crit}$ such that the total energy density is constant and $\sum_i \Omega_i = 1$. The standard $\Lambda$CDM cosmological model considers three main components that drive the Universe evolution. These constituents are the matter density $\Omega_\mathrm{m} = \Omega_\mathrm{b} + \Omega_\mathrm{cdm}$, that itself is composed of baryonic and collisionless cold dark matter. The radiation component $\Omega_\mathrm{r}$, that consider the contribution from photons and relativistic fermions and the dark energy $\Omega_\Lambda$ contribution, that takes into account the vacuum energy. In cosmology is also convenient to express Friedman equation in term of redshift $z \equiv \frac{1}{a} - 1$. Thus, the equation (1.1.4) is redefined for a flat universe $k = 0$ as

$$H(z) = H_0\sqrt{\Omega_r(1+z)^4 + \Omega_m(1+z)^3 + \Omega_\Lambda} \tag{1.1.6}$$

Here $H(z = 0) = H_0$ is the Hubble constant. In equation (1.1.6) the $\Omega_\mathrm{i} \equiv \Omega_{i,0}$ terms indicate respectively the present day radiation, matter or $\Lambda$ density parameter. From various past and present high-redshift galaxy survey (e.g. Colless et al., 2001), large scale clusters measurements (e.g. Wong et al., 2019; Abbott et al., 2021) or combination of CMB and Supernovae observation (Bennett et al., 2013; Riess et al., 2019), it is possible to derive the density parameters of the Universe today. In table 1.1, we list four different estimations for the cosmological parameters. From the left column to the far-right we have the five years observation with the Wilkinson Microwave Anisotropy Probe (WMAP) (Komatsu et al., 2009), results from the Planck satellite telescope (Aghanim et al., 2020), a combination between the previous measurements and local cosmology observations by Riess et al. (2019) and the three years result of the Dark Energy Survey (Abbott et al., 2021).

If we consider a photon originated at redshift $z$ propagating at the speed of light $c$, during the time interval $dt$, it would have covered the proper distance $a(t)\,dr$, considering the Universe expansion. Thus, we obtain the comoving distance between us and the astronomical event by integrating these contributors. With the help of the Hubble parameter equation (1.1.6), we have

$$D_\mathrm{C}(z) = c\int_0^z \frac{dz'}{H(z')} \tag{1.1.7}$$

This relation provides the relative distance between us and the occurring event, at one given redshift $z$, considering that the two points in scape were (co)moving with the Hubble flow.

### 1.1.2 Structure Formation

The CMB provides the initial condition for the cosmological matter density field. The measurement conducted by the Cosmic Background Explorer COBE[2] first showed that the CMB radiation was not perfectly isotropic (Mather et al., 1994). Variation in the density distribution, velocity field and gravitational potential at the last-scattering surface are among the effects contributing to the origin of these anisotropies. These primordial inhomogeneities imprinted during the recombination era eventfully amplify and grow under the gravitational pull, forming the first collapsed structure that eventually hosts the first luminous astronomical objects. The overdensity $\delta$ is opportunely employed to describe the evolution of the density fluctuations.

$$\delta(\mathbf{x}, t) = \frac{\rho(\mathbf{x}, t) - \bar{\rho}}{\bar{\rho}} \tag{1.1.8}$$

Here $\bar{\rho}$ denotes the average matter density of the Universe. At first, the fluctuations in the underlying matter distribution are small $|\delta| \ll 1$. Therefore, we can use the linear perturbation theory applied to the Eulerian and Poisson equations (the former from fluid mechanics) for an expanding perfect fluid to describe the evolution of small density fluctuations during the cosmic expansion. The combination of the first-order perturbation equations then yield to the differential equation for density fluctuations evolution with time

$$\frac{\partial^2 \delta(\mathbf{x}, t)}{\partial t^2} + 2\, H(t)\, \frac{\partial \delta(\mathbf{x}, t)}{\partial t} = 4\pi G\, \bar{\rho}\, \delta(\mathbf{x}, t) \tag{1.1.9}$$

Here $H$ is the Hubble parameter, equation (1.1.4). The solution to equation (1.1.9) is a combination of spatial components and time-dependent growing and decaying modes. In the case of structure formation the former is dominating on the other one and the solution is proportional to the growth factor $D(a)$. Therefore, the linear evolution of the matter overdensity is given as (e.g. Heath, 1977; Peebles, 1980)

$$\delta(\mathbf{x}, t) \propto \frac{D(a)}{a} = \frac{5\, \Omega_\mathrm{m}\, H_0^3}{2} \int_0^1 \frac{da}{(a \cdot H(a))^3} \tag{1.1.10}$$

Here the growth factor is normalised such that $D(a_0) = 1$. Once the overdensity approaches unity $\delta \approx 1$, this solution would break down and require a different treatment to

[2] https://science.nasa.gov/missions/cobe

further extend the solution to the non-linear regime. For this reason, Zel'Dovich (1970) developed a solution that instead of considering the Universe as a cosmic fluid, it is represented as a collection of particles, and their physical trajectory results in (e.g. Bartelmann, 2015)

$$\mathbf{r}(t) = a(t)\,[\mathbf{x} + D(a)\cdot\nabla\phi(\mathbf{x})] \tag{1.1.11}$$

The first and second term on the right-hand side indicates the displacement respectively due to the cosmic expansion and the scalar gravitational potential. This solution is called the Zel'Dovich approximation. Unfortunately, notwithstanding the improvement from linear theory, the approximation fails when particle trajectories cross, as the approach ignores gravitational interaction between particles. Therefore the approximation breaks down as the structure further collapse and overdensity becomes much larger $\delta \gg 1$. The alternative for a complete treatment for later evolution would be to resort to numerical simulations that decompose the matter distribution into collisionless dark matter particles, namely the N-body problem. In section 1.4, we explain further the application of N-body particles simulations in cosmology.

### 1.1.3 Formation of the First Stars and Galaxies

From the virial theorem, we can derive the mass required by dark matter halos to collapse under the influence of gravity and form structures. As a consequence of the collapse, the gas within the halo heats up by either adiabatic compression or shock heating. By assuming that the absolute magnitude of the gravitational potential energy of the halo is twice its kinetic energy, we can then derive the characteristic temperature that the gas reaches during its virialisation at redshift $z$, namely the virial temperature, here (e.g. Tegmark et al., 1997)

$$T_{\mathrm{vir}} \simeq 2\times 10^{3}\,\mathrm{K}\left(\frac{\mu}{1.22}\right)\left(\frac{M_{\mathrm{halo}}}{10^{6}\,\mathrm{M}_{\odot}}\right)^{\frac{2}{3}}\left(\frac{1+z}{20}\right) \tag{1.1.12}$$

Where $M_{\mathrm{halo}}$ is the halo mass, $\mu$ is the mean molecular weight of the gas in the halo normalised by the value for the primordial neutral gas. A cloud of baryonic gas further collapses into the halo gravitational potential under the condition that it can dissipate thermal energy. The thermodynamic cooling is expressed in terms of two timescales that quantify how fast the cloud loses thermal pressure $t_{\mathrm{cool}}$ and how fast the gas is collapsing $t_{\mathrm{ff}}$, defined as (e.g. Benson, 2010)

$$t_{\mathrm{cool}} = \frac{3\,k_{\mathrm{B}}\,T}{2\,n_{\mathrm{gas}}\,\Lambda(T,Z)} \qquad\qquad t_{\mathrm{ff}} = \sqrt{\frac{3\,\pi}{32\,G\rho_{\mathrm{gas}}}} \tag{1.1.13}$$

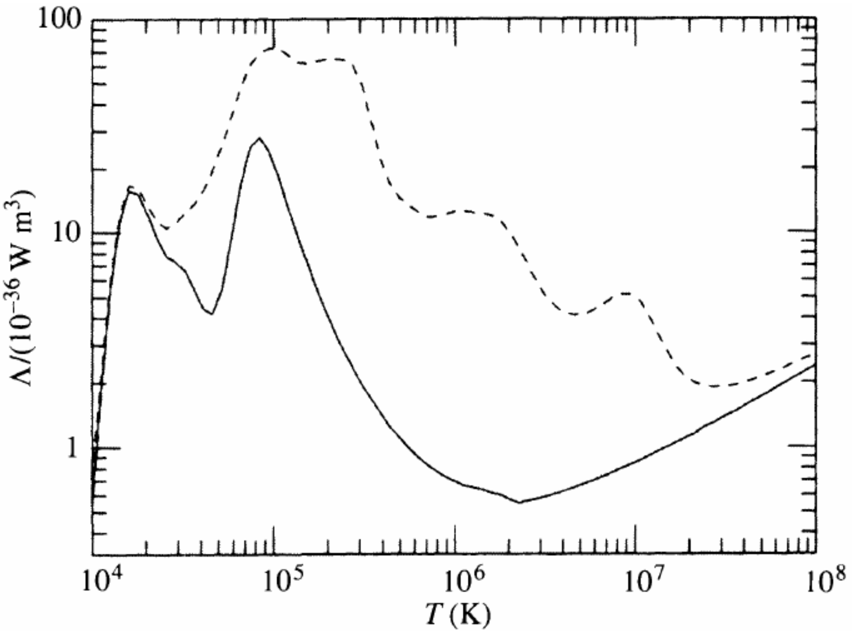

Figure 1.2: The cooling function $\Lambda$ for gas with temperature $T$ and in collisional ionisation equilibrium. The metallicity $Z$ also play an essential role in the cooling mechanisms. The solid line indicates the cooling function of the primordial gas $Z \approx 0$ and with the dashed line for gas of solar composition $Z = Z_{\odot}$. This figure is taken from Binney & Tremaine (1987).

Where $\rho_{\rm gas}$, $T$, $n_{\rm gas}$ and $\Lambda(T, Z)$ are the density, temperature, number density and cooling function of the collapsing gas, respectively. The latter variable quantifies the energy lost over time and volume in a gas cloud of temperature $T$. Moreover, its efficiency strongly depends on the metallicity $Z$ of the cloud. For instance, in figure 1.2 we show the cooling function computed with the `Cloudy` code (Ferland et al., 1998) for the primordial gas (solid line) and gas of solar composition, $Z = Z_{\odot}$ (dashed line). Depending on the temperature of the gas, the main mechanism for cooling can vary between radiative recombination, collisional ionisations, bound-bound transitions and Bremsstrallung emission or a combination of these processes. Therefore, depending on the metallicity, temperature, and density, a cloud of primordial gas can cool faster than its free-fall timescale, $t_{\rm cool} < t_{\rm ff}$, and further collapse to form the first stars and galaxies. From the computed cooling function for the primordial gas (solid line in figure 1.2), we can notice that the atomic cooling mechanism becomes efficient only for temperature $T \geq 10^4\,$K. Therefore, only halos with a certain mass range can host star-forming galaxies. For instance, from equation (1.1.12) we understand that only halos with mass $M_{\rm halo} \geq 10^8 {\rm M}_{\odot}$ have virial temperature of $T_{\rm vir} \geq 10^4\,$K and therefore possibly host star-forming galaxies.

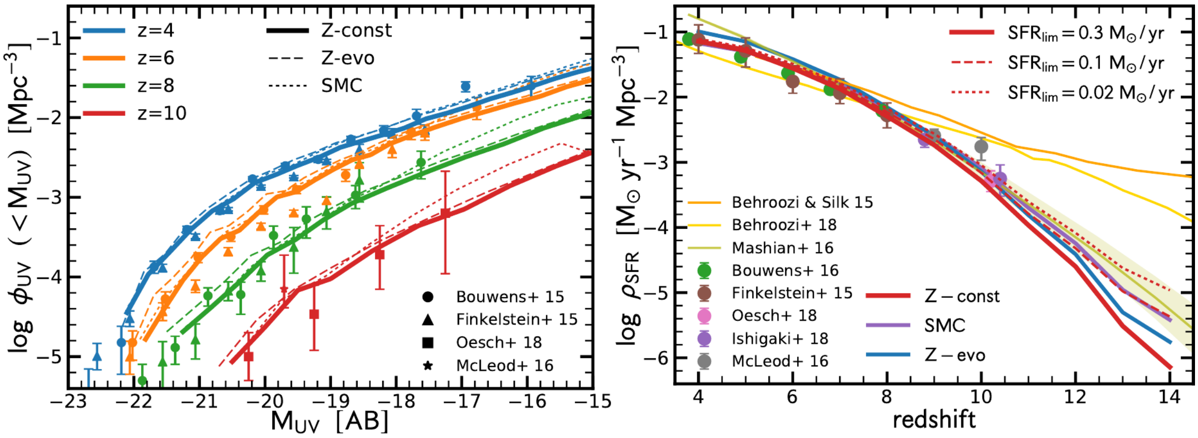

Figure 1.3: *Left panel*: The predicted evolution, from observational data, of the UV luminosity functions at four different redshift, $z$ = 4, 6, 8 and 10. *Right panel*: The star formation rate density history derived from the LFs. The SFR model fit observational data and it is limited to three mass-rate (solid, dashed and dotted line). In both panels, three different approach on metallicity are considered: constant ($Z$-const), time evolving ($Z$-evo) and a mass related metallicity model (SMC). Both figures are taken from Tacchella et al. (2018).

These star-forming galaxies supplied enough ionising photons to be considered the driving sources of reionisation (Robertson et al., 2010; Madau & Dickinson, 2014). Therefore, meaningful information on the process and duration of reionisation can be derived from studying the galaxies star formation rate (SFR) history. $\rho_{\rm SFR}$ during reionisation can be provided from the observed IR and rest-frame UV luminosity functions $\phi$ (LFs) of high-redshift galaxies. For this reason, in the past decade, several surveys have been focusing on measuring the IR and rest-frame UV spectra, such as the *Hubble Space Telescope*[3], the Cluster Lensing And Supernova survey with Hubble[4] (CLASH) and the Hubble Frontier Fields Survey[5]. Moreover, the upcoming James Webb Space Telescope[6] (JWST) will observe objects at even higher redshift as their visible emission shift to the infrared wavelength. By integrating the observed LFs for a limiting magnitude (or luminosity), one can infer the rest-UV luminosity density $\rho_{\rm UV}$ (e.g. Tacchella et al., 2018; Finkelstein et al., 2019), generally this refer to 1′500 Å emission (FUV specific luminosity) but can be defined explicitly for other UV wavelengths (Bouwens et al., 2014). In figure 1.3, left panel, we show an example of the predicted evolution of the UV luminosity function over magnitude

[3] https://hubblesite.org/
[4] https://archive.stsci.edu/prepds/clash/
[5] https://frontierfields.org/
[6] https://www.jwst.nasa.gov/

$M_{\rm UV}$ by Tacchella et al. (2018). The observational data are from Bouwens et al. (2016), Finkelstein et al. (2015), Oesch et al. (2018), Ishigaki et al. (2018) , McLeod et al. (2016) and extend between redshift $z = 4$ and 10. Three models that consider different metallicity models: solid line for a constant value of $Z = 0.02\,Z_\odot$, dashed line for evolving metallicity and a mass related metallicity model by Meurer et al. (1999) (SMC). Its is customary to associate the specific UV luminosity $L_{\rm UV}$ to the SFR $\rho_{\rm SFR}$ by a conversion factor $\kappa_{\rm UV}$ (the same can be assumed for IR) of the order of magnitude $\sim 10^{28}\,{\rm erg\,s^{-1}\,Hz^{-1}/(M_\odot\,yr^{-1})}$ (e.g. Madau & Dickinson, 2014; Robertson et al., 2015a) that is sensitive to the stellar population metallicity, binary stars and the Initial Mass Function (IMF). In figure 1.3, right panel, it shows the SFR history estimated from the UV LF for three lower-luminosity limit (e.g. $M_{\rm UV} = -17$ equivalent to $SFR_{\rm lim} = 0.3\,{\rm M_\odot}/yr$) and the same assumption on the metallicity. The model fits the UV luminosity functions derived from multi-wavelength imaging and spectroscopic surveys data, as mentioned above. The rest-UV luminosity density $\rho_{\rm UV}$ is fundamental to understand the evolution of the IGM during reionisation. When combined with the ionising photon production efficiency $\xi$, we quantify the total ionising photons produced by the sources. Meanwhile, when these two quantities are multiplied with the escape fraction $f_{\rm esc}$, we can determine the portion of these photons that effectively escape into the IGM Finkelstein et al. (2019). The result defines the ionising emissivity and the number of ionising photons produced by stars that actively contribute to the reionisation of the Universe (see § 1.3.1 for discussion).

The first generation of stars formed under particular conditions when compared to their modern equivalent. An important physical ingredient for these stars is molecular hydrogen $H_2$ in the early Universe. The thermodynamic propriety of $H_2$ (rotation and vibrational lines) allows gas to collapse to temperature of $T \sim 300\,{\rm K}$ (Saslaw & Zipoy, 1967; Matsuda et al., 1969; Haiman et al., 1996; Abel et al., 2000; Bromm et al., 2002; Yoshida et al., 2003). Currently, we have no direct observational constraints; however, a series of models propose that these stars should have formed at $z > 10$ inside dark matter collapsed structure of relative small halos, named *mini-halos*, with a mass between $10^5\,{\rm M_\odot}$ and $10^8\,{\rm M_\odot}$ (Bromm et al., 1999; Abel et al., 2001; Yoshida et al., 2003; O'Shea & Norman, 2007). With the current cosmological model, these progenitors, named Population III (PopIII) stars, would have formed from primordial gas composed primarily of hydrogen and helium. Therefore, stellar models show that the first generation of stars should be incredibly massive, with a mass range between $M \sim 20 - 130\,{\rm M_\odot}$ (Umeda & Nomoto, 2003), short-lived due to their lack of metals and produce most of the UV ionising photons, thus kick-starting the reionisation process. One main uncertainty is related to the initial stellar mass function (IMF) of PopIII stars, which is expected to be substantially different to present-day stars due to the absence of metals in the primordial gas. However, UV and Lyman-Werner ($E \sim 11.2 - 13.6\,\mathrm{eV}$) radiation can easily dissociate molecular hydrogen. As a consequence, the radiative feedback of the first PopIII stars either stopped or delayed the further formation of new sources (Machacek et al., 2001; Sazonov & Sunyaev, 2015).

The supernovae explosion of PopIII stars enriched for the first time the interstellar medium with metals. Moreover, the expanding SN shock waves can trigger the formation of the next generation of stars, Population II stars (PopII) (Greif et al., 2010), generally located in globular clusters (van Albada & Baker, 1973), and with a chemical composition that has very few elements heavier than the helium (Schlaufman et al., 2018). Likewise, the SN explosion of PopII stars eventually causes the formation of more recent stars rich with metals, like our Sun, that is named Population I (PopI) stars. The increase of metallicity in the IGM due to PopIII (then PopII) SN explosion has the effect of increasing the efficiency of atomically cooling lines (see figure 1.2, dashed line). This facilitates the formation of PopII stars that are expected to grow in number substantially more than their progenitors. Moreover, the injection of metals into the surrounding IGM is considered to alter the IMF of PopII stars. For lack of observational evidence, this process is still poorly determined (Maio et al., 2010). Even though PopIII stars have a shorter lifetime and are relatively fewer than their PopII counterparts, they are able to produce UV photons more efficiently, mainly due to their extreme mass and their lack of metals which leads to a higher production of ionising photons per unit star formation (Wilkins et al., 2016; Mebane et al., 2018).

## 1.2 Epoch of Reionization

In its initial stage, the Universe was composed of energetic photons that promptly ionise every atom that recombined. However, because of the cosmic expansion, this primordial plasma gradually starts to cool down. When the Universe was approximately a thousand times smaller in size than nowadays, at $z \sim 1100$ and temperatures of approximately $3000\,\mathrm{K}$, these energetic photons have lost enough energy and can no more prevent the formation of neutral atoms, primarily hydrogen and helium. At this point, the primordial plasma becomes optically thin, and photons are free to travel throughout the Universe as well as a relatively small fraction of free electrons, less then 10%. In the beginning, this remanence radiation is rather energetic, with photons in the near-infrared side of the electromagnetic spectrum. However, because of the cosmological Doppler redshift, today their wavelengths move to microwaves range. Thereby this residual radiation is referred to as the Cosmic Microwave Background (CMB).

Under the influence of gravity, dark matter starts to collapse into high-density regions. On a large scale, the Universe is believed to be homogeneous. However, on a small scale, the baryonic matter needs to thermally cool, following the criterion discussed in section 1.1.2, in order to collapse even further and eventually form the first luminous cosmic objects. One hundred million years after the Big Bang, approximately at redshift $z \sim 25$, these sources, which may have been star-forming galaxies and quasi-stellar objects (QSOs), start to produce ionising photons, which over a period of approximately 500 million years completed the re-ionisation of the Universe—transitioning our Universe from an initially cold and neutral state to a final hot and ionised state.

This period is also known as the *Epoch of Reionization* (EoR). It combines a wide range of physical processes, from cosmology and galaxy formation to radiative transfer (RT) and atomic physics. It is a process of importance in structure formation since it has been a direct consequence of creating first structures and ionizing sources, affecting the formation of the late universe inhabitant.

The commonly agreed picture of EoR considers that the first sources start to independently ionise their surrounding neutral gas, creating their so-called *ionised bubble* or *HII regions* in a pre-overlap phase (Choudhury & Ferrara, 2006). Continuing to expand the mean free path of ionising radiation eventually overlap with nearby companions, such that over time these initially isolated bubbles form a vast interconnected ionised region that stretches until ultimately the entire Universe is fully ionised.

### 1.2.1 Observational Constraint

The Epoch of Reionization is one of the least understood periods in the history of the Universe due to the lack of direct observations. For example, very little is known about the history of neutral hydrogen, such as the physics of the ionising sources at high redshifts $z \in [6; 12]$, the stellar initial mass function (IMF), the luminosity function of star-forming galaxies, the escaping fraction of the ionisation radiation from high-redshift galaxies, and more. Nowadays, two types of sources are considered the dominant driver: star-forming galaxies and QSOs, but also more exotic theorised sources should be, at least partially, included, e.g., decaying dark matter, evaporating primordial black holes (Hansen & Haiman, 2004; Avelino & Barbosa, 2004; Mapelli et al., 2006; Chen & Miralda-Escude, 2008).

Instead, several indirect constraints can probe the reionisation process.

### Scattering with CMB photons

Free electron produced during reionisation can interact with cosmic microwave background (CMB) photons. This constitute one of the major indirect probe of reionisation and is often referred as secondary anisotropies of the CMB, meanwhile the primary anisotropies are correlated by the density fluctuations form the last scattered surface. Here we overview three of the main secondary interactions and their physical processes. In the first case, free electrons interact via Thomson-scattering with CMB photons. This elastic interaction does not affect the kinetic energy of both its components. Instead, it dumps the CMB temperature angular power spectra at the small scales, below the horizon scale (e.g. Komatsu et al., 2011a; Planck Collaboration et al., 2016). The Thomson-scattering optical depth can quantify this process depending on the presence of free electrons along the line of sight, defined as

$$\tau_{\rm e}(z) = c\,\sigma_{\rm T} \int_0^z \frac{n_{\rm e}(z')}{(1+z')\,H(z')}\,dz' \tag{1.2.1}$$

where $\sigma_{\rm T} = 6.65 \times 10^{-25}\,{\rm cm}^2$ is the Thomson cross section and $n_{\rm e}$ is the electron density at a given redshift. The CMB temperature fluctuation power spectrum are damped by an exponentially proportional to the electron optical depth $e^{-2\tau_{\rm e}}$. Recent results (e.g. Aghanim et al., 2020) measured this quantity and set the upper limit to $\tau_{\rm e} = 0.0544^{+0.0070}_{-0.0081}$, which implies that reionisation reaches its mid-point at redshift $z_{\rm mid} = 7.68 \pm 0.79$ and therefore a fully ionised Universe for $z \approx 6$. The example in figure 1.4 shows the influence of Thomson scattering optical depth on the CMB angular power spectra for the case of a sudden and fully ionised Universe at redshift $z_{\rm i}$. The models show that scales $l \geq 100$ are affected mainly by high optical depth values (Hu, 1995). Another essential constraint can be obtained from the CMB polarisation power spectrum. Temperature anisotropies result from primordial fluctuations, and these eventually polarise the CMB anisotropies up to the horizon scale, $l \sim 100$, at the last scattering surface. Instead, the signal for larger-scale collapse for larger scale due to the absence of coherent contribution. From this argument, the detections of a polarised signal at a large scale $l < 100$ signify CMB photons Thomson scattering with free-electron produced during reionisation (Rees, 1968). These new features will peak in the polarisation power spectra at scales corresponding to the redshift at which reionisation ends $l \propto \sqrt{z_{\rm re}}$ and its amplitude would be proportional to the electron optical depth, derived with equation (1.2.1). A third process is the so-called Sunyaev & Zeldovich (1972) effect (SZ). In the pre-overlap phase of reionisation,

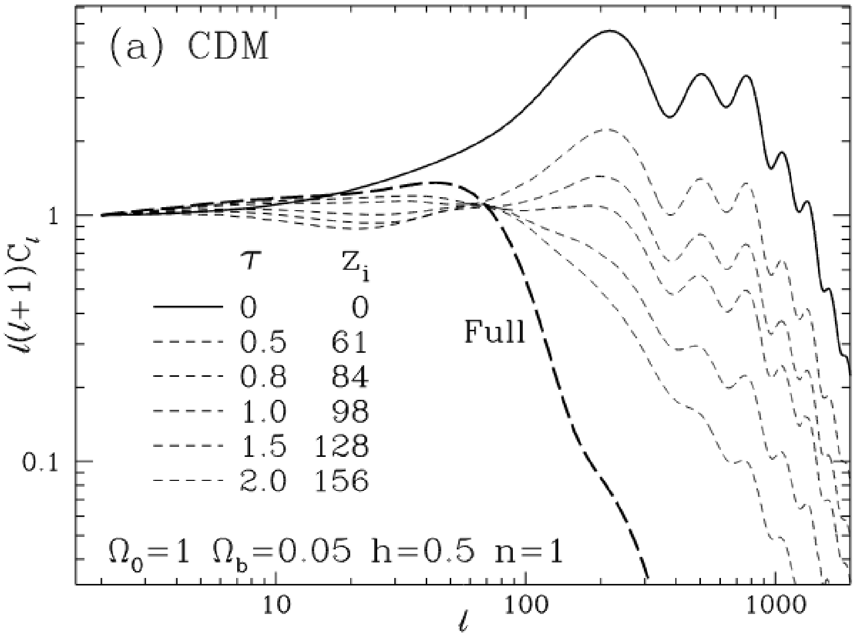

Figure 1.4: Effect of the optical depth on the CMB temperature (TT) angular power spectra for standard ΛCDM. Here the models consider uniform and sudden fully reionisation at redshift $z_\mathrm{i}$. This figure is taken from Hu (1995).

the high energy free electrons in the hot plasma surrounding the source can interact via inverse Compton-scattering with the CMB photons. Here, CMB thermal anisotropies rise at small scale $l > 2000$ (Mesinger et al., 2012) due to the peculiar motion of hot ionised gas. This velocity component is a combination of the thermal velocity of the singular electron, referred as the thermal SZ effect (tSZ) and the collective motion of the entire cluster, the kinetic SZ effect (kSZ). In the first case, tSZ shifts low-frequency CMB photons to higher energy distorting the CMB black body spectrum. The latter case gives a proportional shift in temperatures depending on the cluster velocity while maintaining its characteristic spectral shape. We can quantify the influence of the kSZ effect on CMB temperature anisotropies along the line of sight $\mathbf{n}$ with the following formula (McQuinn et al., 2005, e.g.)

$$\frac{\Delta T}{T}(\mathbf{n}) = c\,\sigma_\mathrm{T} \int_0^z \frac{n_\mathrm{e}(z')\cdot(\mathbf{n}\cdot\mathbf{v}(z'))}{(1+z')\,H(z')}\, e^{-\tau_\mathrm{e}}\, dz' \tag{1.2.2}$$

where $\mathbf{v}$ indicates the peculiar motion of the cluster hosting free electron and the exponential factor gives the probability of scattering.

### Lyman-$\alpha$ Radiation and Absorption lines

Observation of absorption lines in the spectra of distant galaxies or QSOs can tell us of the state and proprieties of the intergalactic medium (IGM). Neutral hydrogen in the IGM along the line of sight, at redshift $z$, can absorb UV radiation of these distant sources, with a wavelength of $\lambda_{\mathrm{Ly}\alpha} = 1216\,\text{Å}\,(1+z)$. The absorption features in QSO spectra can be divided into two main categories, when the column density of neutral hydrogen is below $N_{\mathrm{HII}} \lesssim 10^{17}\,\mathrm{cm}^2$ this observed feature is known as Lyman-$\alpha$ forest. Meanwhile, when the scattering with the resonant line becomes so intense that almost all radiation with wavelengths $\lambda < \lambda_{\mathrm{Ly}\alpha}$ is absorbed, this is known as the Gunn-Peterson (GP) effect. The latter absorption features were predicted in 1965 by Gunn & Peterson (1965). The magnitude of the absorption depends on the column density of the neutral hydrogen, and it is quantified by the GP optical depth, defined by (e.g. Choudhury & Ferrara, 2006; Ferrara & Pandolfi, 2014)

$$\tau_{\mathrm{GP}}(z) = \frac{1.8 \times 10^5}{h\,\Omega_{\mathrm{m}}^{1/2}} \left(\frac{\Omega_{\mathrm{b}}\,h^2}{0.02}\right) \left(\frac{1+z}{7}\right)^{3/2} x_{\mathrm{HI}}(z) \tag{1.2.3}$$

here, approximated for high redshift $z \gg 1$, where $x_{\mathrm{HI}}$ is the volume fraction of neutral hydrogen at redshift $z$. The GP optical depth becomes tremendously large even for extremely small neutral fraction $x_{\mathrm{HI}} > 10^{-3}$. In figure 1.5 we can see an example of the GP effect and Lyman-$\alpha$ forest in the spectra of a QSO. The broad peak at $\lambda_{\mathrm{Ly}\alpha} \simeq 8670\,\text{Å}$ correspond to the Ly$\alpha$ emission from the quasar. This feature is used as an indicator of the quasar rest frame, and in this case, it corresponds to redshift $z \simeq 6.13$ (Becker et al., 2014). The bluer side of the QSO Ly$\alpha$ emission restframe, $\lambda_{\mathrm{obs}} < 8500\,\text{Å}$, show the Lyman-$\alpha$ forest (and other Lyman transitions) as the result of UV photons absorption in the presence of residual neutral gas along the line of sight. The intense absorption between wavelength $8400\,\text{Å} < \lambda_{\mathrm{obs}} < 8600\,\text{Å}$ create a gap in the spectra with virtually no flux and is known as the GP trough. The GP optical depth provides the primary evidence for a highly ionised IGM at an intermediate redshift between $z = 2$ and 5 (Fan et al., 2006a; McGreer et al., 2011, 2014). Nevertheless, this method is useful only for highly ionised IGM as it is necessary a small fraction of neutral hydrogen $x_{\mathrm{HI}} \sim 10^{-3}$ to be optically tick $\tau_{\mathrm{GP}} \sim 100$ and absorb all photons at wavelength $\lambda_{\mathrm{Ly}\alpha}$. The absorption features decreases in QSOs spectra at redshift $z \sim 6$ (Fan et al., 2006b; Greig et al., 2016; Davies et al., 2018), suggesting that cosmic reionisation is completed at that time. Other indirect constraints considers the Lyman-$\alpha$ damping wing in high-z quasar spectra (e.g. Schroeder et al., 2013; Totani et al., 2016; Davies et al., 2018; Greig et al., 2019) and the number density of galaxies emitting Lyman-$\alpha$ radiation (e.g. Ota et al., 2008; Ouchi et al., 2010; Konno et al., 2014; Robertson et al., 2015b), since the latter provides an efficient way to track early star-forming galaxies at high redshift and systems with virtually no dust. The two main method by which Lyman-$\alpha$ emitting galaxies are detected is either

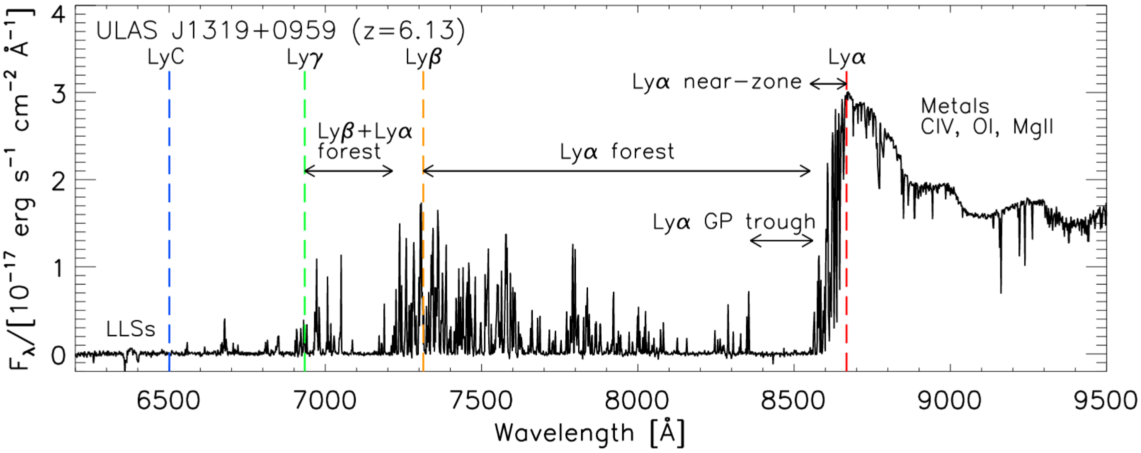

Figure 1.5: High-resolution spectrum versus the observed wavelength of quasar `ULAS J1319+0959` obtained with the X-Shooter spectrograph on the Very Large Telescope (VLT). The broad peak at $\lambda_{\rm Ly\alpha} \simeq 8670\,$Å is the Ly$\alpha$ emission and it places the QSO at redshift $z \simeq 6.13$. The bluer side of the spectra, $\lambda_{\rm obs} < 8500\,$Å, shows absorption lines due to the presence of residual neutral hydrogen along the line of sight, while the gap in the spectra at $8400\,$Å $< \lambda_{\rm obs} < 8600\,$Å is the GP trough. This figure is taken from Becker et al. (2015).

by narrow-band wavelength imaging centred on the redshifted $\lambda_{\rm Ly\alpha}$ or spectroscopy on a wider redshift range but smaller field of view. Both methods can make use of gravitational lensing to amplify the faint brightness of galaxies generally at redshift $z > 5$.

### Hyperfine Transition of Neutral Hydrogen: the 21-cm line

The neutral hydrogen ground state is split into two energy levels, the triplet and singlet state. The electron and nucleus spin interaction changes the atom's energy state, emitting a photon with a rest-frame wavelength of $\lambda_0 = 21.16\,$cm and frequency of $\nu_0 = 1.42\,$GHz. Transition between these two hyperfine levels is known as the 21-cm signal. Predicted by van de Hulst (1945) in collaboration with Jan Oort, it was then theoretically computed to study the structure of the Milky Way. This spin-flip transition has an extremely small Einstein coefficient $A_{10} = 2.85 \times 10^{-15}\,{\rm s}^{-1}$, which corresponds to a spontaneous emission lifetime of approximately $10^7$ years. Nevertheless, because of the widespread and abundance of neutral hydrogen in the early universe IGM, this signal will be a unique signature of EoR (e.g. Madau et al., 1997; Furlanetto et al., 2006).

In radio astronomy is convenient to express the intensity of a signal, observed at a particular frequency $\nu$, in its equivalent brightness temperature $T_{\rm b}(\nu)$. When a radiation background (in our case, the CMB) passes through a gas cloud of neutral hydrogen with temperature $T_\mathrm{S}$, the temperature of the emergent radiation is then

$$T_\mathrm{b}(\nu) = T_\mathrm{S}\,(1 - e^{-\tau_\nu}) + T_\mathrm{CMB}(\nu)\,e^{-\tau_\nu} \tag{1.2.4}$$

the first term on the right-hand side quantifies the emission probability of 21-cm photons from within the cloud. On the other hand, the second term gives the proportion of incoming radiation transmitted to the cloud. Here $\tau_\nu$ is the 21-cm optical depth for diffuse IGM, and by integrating along the entire line of sight, we define the total 21-cm absorption that has an exact solution expressed as

$$\tau_{\nu_0}(z) = \frac{3}{32\pi}\frac{h\,c^3\,A_{10}}{k_\mathrm{B}\,T_\mathrm{S}\,\nu_0^2}\frac{n_\mathrm{HI}(z)}{H(z)} \tag{1.2.5}$$

where $\nu_0 = 1.42\,\mathrm{GHz}$ is the frequency of the 21-cm line and $n_\mathrm{HI}$ the density distribution of neutral hydrogen at redshift $z$.

The quantity measured by radio telescope is the differential brightness temperature $\delta T_\mathrm{b} \equiv T_\mathrm{b} - T_\mathrm{CMB}$, such that when we assume a small 21-cm optical depth (e.g. Furlanetto et al., 2006; Mellema et al., 2006b) we have

$$\delta T_\mathrm{b}(z) = \frac{T_\mathrm{S} - T_\mathrm{CMB}}{1+z}\,(1 - e^{-\tau_\nu}) \approx \frac{3}{32\pi}\frac{h\,c^3\,A_{10}}{k_\mathrm{B}\,T_\mathrm{S}\,\nu_0^2}\left(1 - \frac{T_\mathrm{CMB}}{T_\mathrm{S}}\right)\frac{n_\mathrm{HI}(z)}{(1+z)H(z)} \tag{1.2.6}$$

Here the $(1+z)$ term at the denominator, after the first equivalence, accounts for the $T_\mathrm{CMB}$ decrease due to the Doppler redshift. It is important to notice that $\delta T_\mathrm{b}$ can be arbitrarily positive or negative. The spin temperature is associated with the ratio between the hydrogen atom number density in the singlet $n_0$ and triplet state $n_1$, and it is given as (Field, 1958)

$$\frac{n_1}{n_0} = 3\,exp\left(-\frac{T_*}{T_\mathrm{S}}\right) \tag{1.2.7}$$

A standard approach is to define the spin temperature as a weighed fraction, based on the different process that can contribute to the evolution of $T_\mathrm{S}$, defined as (Field, 1959)

$$T_\mathrm{S} = \frac{T_\mathrm{CMB} + y_\mathrm{k}\,T_\mathrm{k} + y_\alpha\,T_\alpha}{1 + y_\mathrm{k} + y_\alpha} \tag{1.2.8}$$

For instance, the factor $y_\mathrm{k}$ accounts for the collisional excitation of the 21-cm signal with free-electron or other hydrogen atom and its contribution to the increase/decrease of the gas kinetics temperature $T_\mathrm{k}$. The decoupling of Ly-$\alpha$ radiation by the Wouthyusen-Field effect (Wouthuysen, 1952) is quantified by the factor $y_\alpha$ and the colour temperature $T_\alpha$. This process is also called the Lyman-$\alpha$ pumping mechanism due to the fact that photons transition the atom via photo-excitation to the Ly-$\alpha$ spectral line series, and a 21-cm photon is induced once the atom decay and return to the ground state. Furthermore,

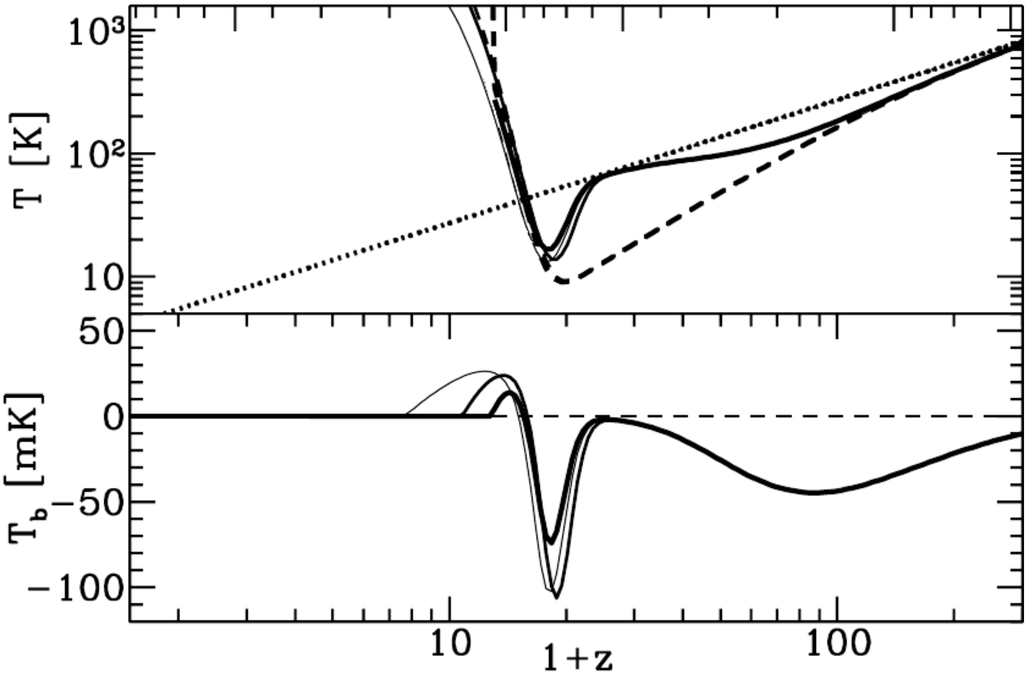

Figure 1.6: *Top panel*: evolution of the spin temperature $T_\mathrm{S}$ (solid black line) and its components over redshift. Different thickness of the line indicates the effect of the components, from equation (1.2.8), on the spin temperature. *Bottom panel*: corresponding evolution of the differential brightness $\delta T_\mathrm{b}$. This figure is taken from Pritchard & Loeb (2012a)

absorption of CMB photons can alter the spin temperature, and from equation (1.2.6) we understand that the 21-cm signal is observable only in the case when the spin temperature differs from the CMB temperature $T_\mathrm{S} \neq T_\mathrm{CMB}$, otherwise the two are indiscernible.

Therefore, the expected 21-cm signal is divided into epochs, represented by cardinal heating or cooling mechanisms that characterised its evolution. The sign of the differential brightness $\delta T_\mathrm{b}$ depends mainly on the relation between the spin and CMB temperature. In figure 1.6, we show an example of the evolution of the spin temperature (top panel) and the consecutive variation on the brightness temperature (bottom panel) from the work of Pritchard & Loeb (2012a). The former shows the variation of $T_\mathrm{S}$ (boldface solid line) for three models, compared to the IGM $T_\mathrm{k}$ (boldface dashed line) and the CMB $T_\mathrm{CMB}$ temperatures (dotted line). The bottom panel show the corresponding evolution of the differential brightness for three different source models (solid lines of different thickness) that regulate the amplitude and position of the peak/trough.

Ab initio, $z > 200$, Campton scattering between free electron and the CMB photons keeps the spin temperature coupled to the CMB. At the same time, the Universe is dense enough such that the collisional coupling between atoms is highly efficient. Therefore after recombination, both the gas and spin temperatures are strongly coupled to the CMB, $T_{\rm S} = T_{\rm k} = T_{\rm CMB}$ and the differential brightness is close to zero $\delta T_{\rm b} \approx 0$. For redshift $z \leq 200$, the Campton scattering becomes inefficient, such that the gas in IGM decouple from the CMB and start to adiabatically cool $T_{\rm k} \propto (1+z)^2$. Meanwhile, the CMB temperature decreases proportionally to redshift, $T_{\rm CMB} \propto (1+z)$. Therefore, the differential brightness becomes for the first time negative $\delta T_{\rm b} \leq 0$ around redshift $z \sim 100$ in what is defined as the *Dark Ages* of reionisation.

As a consequence of the cosmic expansion, the collisional coupling efficiency begins to decline. At this point, just before star formation becomes significant, the 21-cm signal is expected to fade again as $T_{\rm S} \approx T_{\rm CMB}$. With the formation of the first luminous sources between redshift $z = 20 - 30$, Lyman-$\alpha$ continuum radiation is emitted by the sources that heat the surrounding gas in IGM. As a result, the spin temperature couple back with the gas kinetic temperature $T_{\rm S} \rightarrow T_{\rm k}$, upon which is still adiabatically cooling. The differential brightness becomes again predominately negative in what is known as the *Cosmic Dawn*. The Experiment to Detect the Global EoR Signature (EDGES, Bowman et al., 2018) detected an absorption feature in the sky-averaged 21-cm radio signal at redshift $z = 17.2$ ($\nu_{\rm obs} = 78\,{\rm MHz}$) that is much deeper than expected. This detection put in discussion the physical processes that have been employed until now. Several works have proposed alternative models to explain this exceedingly deep absorption, by either including non-standard physics cooling mechanisms (e.g. Barkana, 2018; Fraser et al., 2018; Pospelov et al., 2018). Other efforts consider varying the proprieties of dark matter particle and astrophysical parameters or by including an excess of radio background radiation in addition to the CMB (e.g. Fialkov et al., 2018; Ewall-Wice et al., 2018; Feng & Holder, 2018).

The appearance of the first luminous object is a direct consequence of the first X-ray sources, such as SN explosions, neutron stars and black holes (e.g. Nath & Biermann, 1993; Zaroubi et al., 2007; Mirocha, 2014; Sazonov & Sunyaev, 2015). The radiation produced by these sources can travel long distances (of a few Mpc in size) before being absorbed by the neutral gas in the IGM. At this stage, the spin temperature decouples for the last time from the CMB radiation $T_{\rm S} \gg T_{\rm CMB}$ as it gets heated and eventually partially ionised by the X-ray background radiation. Hence, for redshift $z < 20$, the 21-cm signal is sensitive to the presence of neutral hydrogen and thus the density fluctuations. Finally, with the beginning of the *Reionisation Era*, the UV radiation starts to propagate and ionise the vast neutral IGM, the 21-cm globally averaged signal start to decline until it becomes completely transparent $\delta T_\mathrm{b} = 0$ as a consequence of the decline of neutral hydrogen in the Universe.

### 1.2.2 Current & Future Direct Observations

The neutral hydrogen present in the IGM at redshift $z$ emits 21-cm photons with frequency $\nu_0 = 1.42\,\mathrm{GHz}$. When observed, the 21-cm signal would have redshifted to the radio band of the electromagnetic spectrum, such that while the observed wavelength increases to scale of a few metres $\lambda_\mathrm{obs} \sim 1 - 5\,\mathrm{m}$, the observed frequency shifts to lower values, $\nu_\mathrm{obs} \sim 50 - 200\,\mathrm{MHz}$, such that

$$\lambda_\mathrm{obs}(z) = \lambda_0\,(1+z) \qquad\qquad \nu_\mathrm{obs}(z) = \frac{\nu_0}{1+z} \tag{1.2.9}$$

In principle, we can probe the reionisation process by observing the redshifted signal produced during the cosmic reionisation, following equation (1.2.6). Various radio experiments, such as Low Frequency Array[7] (LOFAR; e.g. van Haarlem et al., 2013), Murchison Widefield Array[8] (MWA; e.g. Tingay et al., 2013) and the Hydrogen Epoch of Reionization Array[9] (HERA; e.g. DeBoer et al., 2017), have been trying to detect this signal. Recently, these facilities have provided useful upper limits on the 21-cm power spectrum (e.g. Mertens et al., 2020; Trott et al., 2020) that have been used to learn the properties of reionisation (e.g. Ghara et al., 2020; Mondal et al., 2020; Greig et al., 2020a,b). However, the 21-cm signal during EoR will be highly non-Gaussian and therefore the power spectrum will not give the full characterisation (e.g. Mellema et al., 2006b; Ichikawa et al., 2010a; Watkinson & Pritchard, 2015; Majumdar et al., 2018; Giri et al., 2019c). In the coming years, the Square Kilometre Array[10] (SKA) will be built. The low-frequency component of the SKA will be sensitive enough to detect the 21-cm signal produced during EoR and create images of its distribution on the sky (Mellema et al., 2013; Wyithe et al., 2015; Koopmans et al., 2015). A sequence of multiple 21-cm images from different redshifts (or observed frequency) will constitute a three-dimensional set of data known as the tomographic dataset. This observation will enable direct studies of the sizes and shapes of ionised/neutral regions during the EoR. For instance, the statistical proprieties of the signal is often described using the 21-cm power spectrum $P_{21}(k)$ (see section 2.3.4 for results discussion). The power spectra quantify the amplitude fluctuations in the 21-cm signal

[7] https://www.astron.nl/telescopes/lofar/

[8] https://www.mwatelescope.org/

[9] http://reionisation.org/

[10] https://skatelescope.org

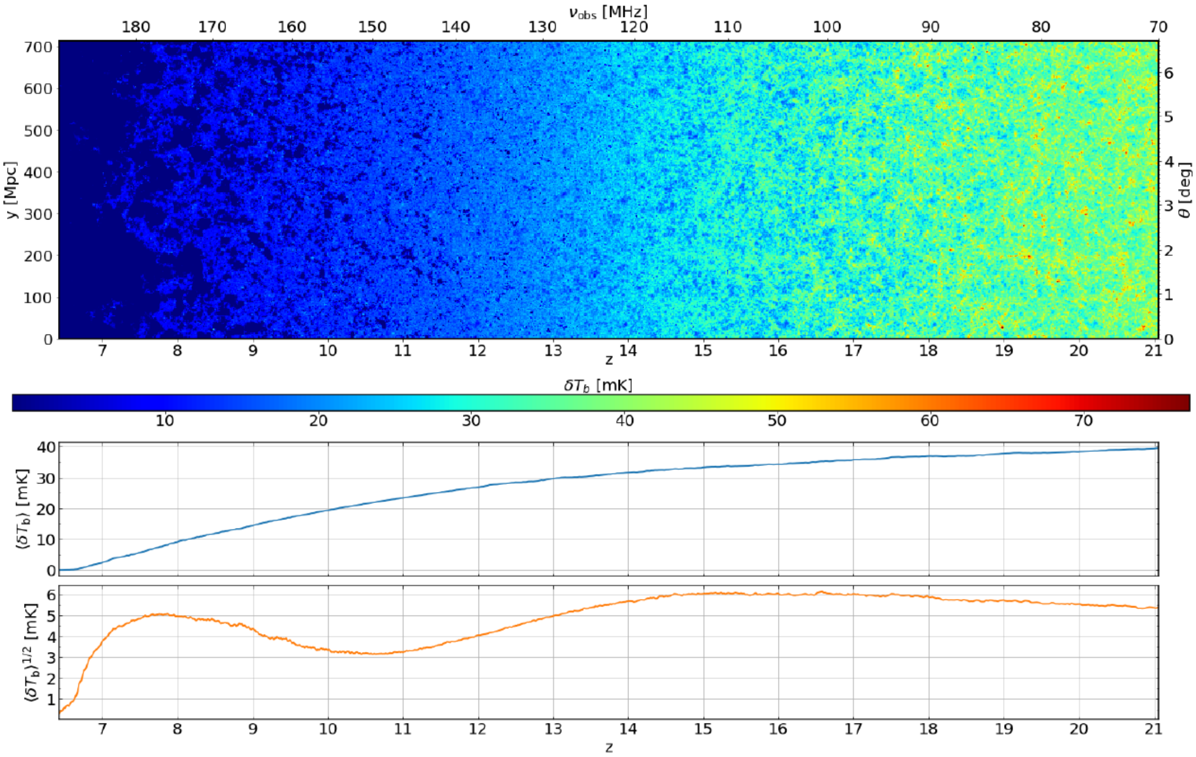

Figure 1.7: Slice of a simulated tomographic dataset from our largest fully-numerical simulation. The redshift evolution of the 21-cm, which is quantified by the differential brightness $\delta T_b$ observed by radio telescopes. Here, we show the redshift range relevant for the period of the *Reionisation Era*. On the y-axis the box size in comoving Megaparsec (cMpc) for given redshift $z$ and corresponding cosmic time on the x-axis.

as a function of the amplitudes scale $2\pi k^{-1}$. The Fourier transform of the fractional perturbation to the brightness temperature $\delta_{21}(\mathbf{x}) \equiv \left[\delta T_{\rm b}(\mathbf{x}) - \delta\bar{T}_{\rm b}\right]/\delta\bar{T}_{\rm b}$ define the power spectrum such that

$$\left\langle \tilde{\delta}_{21}(\mathbf{k}_1)\,\tilde{\delta}_{21}(\mathbf{k}_2)\right\rangle = (2\pi)^3\,\delta_{\rm D}(\mathbf{k}_1+\mathbf{k}_2)P_{21}(\mathbf{k}_1)$$

However the fluctuation in the 21-cm signal are highly non-Gaussian (e.g. Mellema et al., 2006b; Giri et al., 2019c) in such a way that the power spectra does not provide a full statistical description of the signal. Therefore, modern studies proposed topological descriptor such as Euler characteristics and Betti numbers (Friedrich et al., 2011; Elbers & van de Weygaert, 2019; Giri & Mellema, 2021; Kapahtia et al., 2021, e.g.) as an alternative. The latter method in particular, it provides a more complete statistical description of the size and shape of the 21-cm regions and the subsequent vast interconnected ionised regions. In the case of a 3D object the Euler characteristic is defined as a sum of the first three Betti numbers, defined as $\chi = N_o - N_t + N_c$, where $N_o$ is the zeroth term and indicates the number of isolated objects, meanwhile the first and second term are the number of tunnels $N_t$ and cavities $N_c$, respectively. Giri & Mellema (2021) demonstrated that this values have a distinctive evolution during the different stages of reionisation, making them an useful tools to describe the state of the IGM from future SKA tomographic dataset (see § 3.4.6 for results discussion).

In figure 1.7 top panel, we show an example of a simulated tomographic dataset. The slice is approximately 700 cMpc wide and, to our knowledge, it is currently the largest fully-numerical reionisation simulation available. For reference the comoving spatial resolution is $\Delta x = 2.381\,\mathrm{cMpc}$, the corresponding angular $\Delta\theta$ and frequency resolution $\Delta\nu$ at redshift $z$ is evaluated with

$$\Delta\theta = \frac{\Delta x}{D_\mathrm{C}(z)} \qquad\qquad \Delta\nu = \frac{\nu_0\, H(z)\, \Delta x}{c\,(1+z)^2} \tag{1.2.10}$$

Here $H$ is the Hubble equation and $D_\mathrm{C}$ is the comoving distance, equation (1.1.7). In our case for $z = 9$ we have $\Delta\theta = 1.22\,\mathrm{arcsec}$ and $\Delta\nu = 185.85\,\mathrm{kHz}$. On the redshift/frequency axis, we can see the evolution of the 21-cm signal. From the appearance of the first sources, at the early stage of reionisation $z \sim 20$, the neutral IGM is heated by X-ray radiation (e.g. Ross et al., 2019) in order that the kinetic component in the equation (1.2.8) becomes the dominant term. Thus the parenthesis term in the right-hand side of the equation (1.2.6) is approximated to unity as $\frac{T_\mathrm{CMB}}{T_\mathrm{S}} \to 0$. This approach is known as the *Spin Saturated Approximation* (see § 1.2.1 for discussion) and is relevant for the *Reionisation Era* (Furlanetto et al., 2004). Hence, with this approximation the differential brightness can be observed for the majority in emission, $\delta T_\mathrm{b} > 0\,\mathrm{mK}$. Therefore, at high redshift $z > 18$, the signal correlates with the high-density regions and is well above $\delta T_\mathrm{b} \geqslant 40\,\mathrm{mK}$. The ionising fronts from sources then start to ionise their surrounding IGM and the global signal $\langle\delta T_\mathrm{b}\rangle$ gradually decreases (the middle panel in figure 1.7). The first regions with a lack of signal, $\delta T_b = 0\,\mathrm{mK}$, start to appear at redshift $z \approx 15$, indicating the presence of ionised hydrogen. At this stage, the fluctuation in the signal reaches the first peak with $\langle\delta T_\mathrm{b}\rangle^{1/2} \sim 6\,\mathrm{mK}$ (bottom panel), signifying that the ionised regions have reached sizes of a few Mpc and have surpassed the simulation resolution size $\Delta x$. Iliev et al. (2014) demonstrated how this features becomes typical for large enough simulations (of volumes size $L > 100\,\mathrm{Mpc}$) and its amplitude and redshift-position dependent on the sources model employed. Over time, these ionised regions grow, at first of a few teens Mpc size, and eventually for redshift $z \approx 8$, they merge with neighbours bubbles into vast regions. This transition is signed by a first decline in the signal fluctuations, $\langle\delta T_\mathrm{b}\rangle^{1/2} \sim 3\,\mathrm{mK}$, and subsequently a second peak up to $\langle\delta T_\mathrm{b}\rangle^{1/2} \sim 5\,\mathrm{mK}$. Ultimately, for redshift $z < 7$, the Universe is entirely ionised and therefore, it appears transparent with virtually no neutral hydrogen left $\langle \delta T_{\rm b} \rangle \sim 0\,{\rm mK}$ except inside high density regions on small scales, below $\sim 2\,{\rm Mpc}$ in size.

Tomographic images from the SKA telescope will be prone to instrumental restrictions such as limited resolution (Braun et al., 2019). However, a more concerning problem is the foreground radio-loud emission as it can contaminate and overcome the feeble 21-cm reionisation signature. The foreground contamination can be of galactic or extragalactic origin, it requires substantial knowledge of its physics, and can be up to a few million times brighter than the 21-cm signal. Moreover, the sensitivity of the SKA telescope depends on the contrast between the collected 21-cm signal and the antennas noise, as the instrumental systematics can be from three to up to six orders of magnitude larger and outshine the reionisation signal. The noise fluctuations $\sigma_{\rm noise}$ for interferometric radio telescope is widely studied (e.g. McQuinn et al., 2006; Ghara et al., 2016, 2017) and can be modelled by a Gaussian field with root-mean-squared (RMS) of

$$\sigma_{\rm noise} = \frac{\sqrt{2}\, k_{\rm B}\, T_{\rm sys}}{A_{\rm eff}\, \sqrt{\Delta\nu\, t_{\rm int}}} \tag{1.2.11}$$

where $k_{\rm B}$ is the Boltzmann constant, $A_{\rm eff}$ the effective collecting area, $t_{\rm int}$ the correlator integration time and $\Delta\nu$ the frequency band (expression on the right in equation (1.2.10)). The system temperature is defined in function of the observed frequency, such that $T_{\rm sys} \approx 60\,{\rm K}(\frac{\nu}{300\,{\rm MHz}})^{-2.55}$. In figure 1.8 left panel, we show the distribution of a simulated Gaussian noise field after 1000 hours observation for redshifts $z = 7$, 8, 9, 10, 11 and 12. The frequency resolution has a redshift dependency $\Delta\nu \propto (1+z)^{-1/2}$ such that the noise RMS increases for observations at higher redshift. On the right panel, we show a 714 cMpc slice of the realisation at redshift $z = 9$ or frequency $\nu_{\rm obs} = 142\,{\rm MHz}$, calculated with equation (1.2.9). As we can see, at this observed frequency the noise reaches temperatures of $\pm 2200\,{\rm K}$. In our case, we select the variables of equation (1.2.11) based on the baseline design of the SKA1-Low core station[11]. Hitherto, for the case of SKA the effective collecting area is considered to be the area of the core station (35 m in diameter) such that $A_{\rm eff} = 962\,{\rm m}^2$. Other parameters are derived from the same setup, such that $t_{\rm int} = 10\,{\rm s}$, $T_{\rm sys} = 404.1\,{\rm K}$ and frequency resolution is $\Delta\nu = 185.85\,{\rm kHz}$, calculated with equation (1.2.10) for $z = 9$ and $\Delta x = 2.381\,{\rm cMpc}$.

[11] SKA1 system baseline design revision 1 (2013):`https://www.skatelescope.org/wp-content/uploads/2013/05/SKA-TEL-SKO-DD-001-1_BaselineDesign1.pdf`

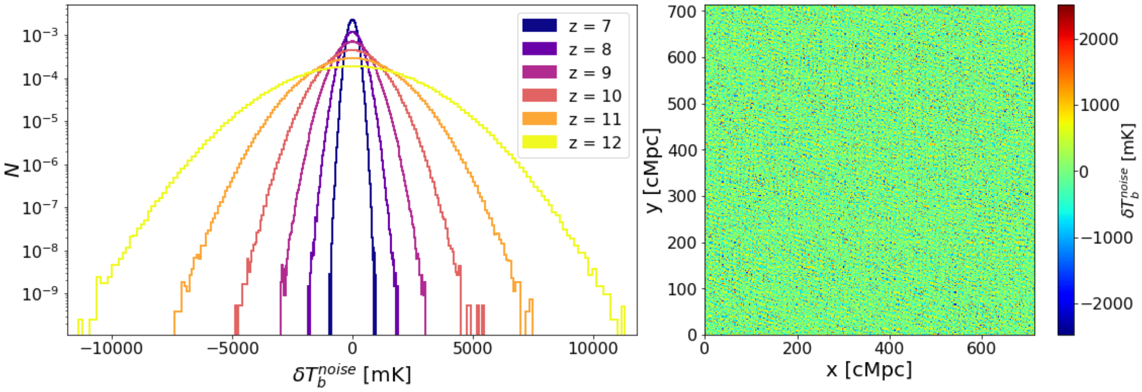

Figure 1.8: Simulated noise for SKA telescope. *Left panel*: noise distribution at different redshift $z$ = 7, 8, 9, 10, 11 and 12. *Right panel*: a slice of 714 Mpc per size of the cube noise at redshift $z = 9$.

## 1.3 Theoretical Background

In the presence of radiation, the ionization evolution of the primordial gas is given by a differential equation, one for every component (e.g. Ferrara & Pandolfi, 2014; Park et al., 2016). Therefore the evolution of the neutral hydrogen fraction $x_{\rm HI} = n_{\rm HI}/n_{\rm H}$ and ionised fraction $x_{\rm HII} = 1 - x_{\rm HI}$, where $n_{\rm H} = n_{\rm HI} + n_{\rm HII}$, is given by balance between the number of ionising photons and number of ion-electron that recombine into atoms, so

$$\begin{aligned} \frac{dx_{\rm HII}}{dt} &= (1 - x_{\rm HII})(\Gamma_{\rm ion} + C_{\rm ion}\, n_{\rm e}) - \alpha_{\rm B}(T)\, x_{\rm HII}\, n_{\rm e} \\ \frac{dx_{\rm HI}}{dt} &= -\frac{dx_{\rm HII}}{dt} \end{aligned} \tag{1.3.1}$$

For simplicity, we consider only the presence of hydrogen, so only one differential equation is required to be solved. In the first equation, on the right-hand side, the positive term indicates the contributors to ionizing the gas. $C_{\rm ion}$ is the term for collisional ionization with free electron or ions, relevant only for high density regions and important temperature $T_{\rm gas} \gg 10^4$ K (Hui & Gnedin, 1997; Shull et al., 2004; Ciardi & Ferrara, 2005). $\Gamma_{\rm ion}$ is the photo-ionization rate, and it represents the total number of ionizing photons per second per hydrogen atom emitted by the sources.

$$\Gamma_{\rm ion} \equiv 4\pi \int_{\nu_{\rm HI}}^{\infty} \sigma_{\rm HI}(\nu)\, \frac{J_\nu}{h\,\nu}\, d\nu \tag{1.3.2}$$

where $\nu_{\rm HI}$ is the frequency threshold required to ionize hydrogen, $J_\nu \equiv J(\mathrm{t}, \mathbf{r}, \hat{\mathbf{n}}, \nu)$ is the angular-averaged specific ionising intensity and $\sigma_{\rm HI}$ the photo-ionization cross section for hydrogen.

The negative term in equation (1.3.1) expresses the recombination rate of ionised hydrogen with free electron. Here, hydrogen is the only element that contributes to the

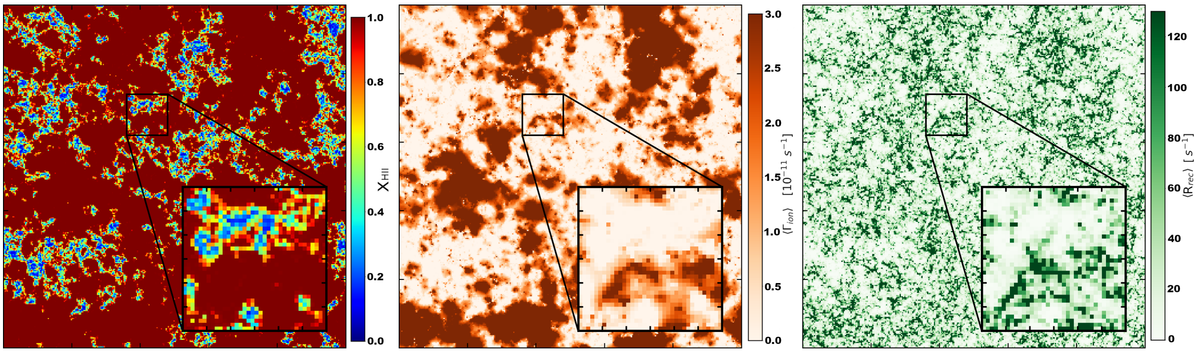

Figure 1.9: Visual representation of equation (1.3.1) performed with the C$^2$RAY radiative transfer code. Each panel show a slice of a simulated box with size 714 Mpc per side at redshift $z \approx 7$, when the volume is 90% ionised. *Left panel*: the hydrogen ionisation field. *Central panel*: the volume averaged photo-ionisation rate is the positive term in the same equation *Right panel*: the volume averaged recombination rate for a primordial gas composed of only atomic hydrogen. In each panel we show a slice zoom-in on a region of size approximately 140 Mpc per side. The simulation volume resolution is of 13.5 Mpc$^3$.

electron number density, hence the electron number density is simply $n_{\rm e} = n_{\rm HII} = x_{\rm HII}\,\bar{n}_{\rm H}$. Such that,

$$\mathcal{R} \equiv \alpha_{\rm B}(T)\,x^2_{\rm HII}\,\bar{n}_{\rm H} \tag{1.3.3}$$

where $\alpha_{\rm B} \propto T^{-0.7}$ is the (temperature-dependent) Case B recombination coefficient. It indicates the number of electron-proton recoupling per second for hydrogen at temperature $T$ and ignores recombination to the ground state. The proper mean number density of neutral hydrogen $\bar{n}_{\rm H} \equiv n_{\rm H}(z)$ at redshift $z$ is defined by the cosmology such as

$$n_{\rm H}(z) = \frac{\Omega_{\rm b}\,\rho_{c,0}}{m_{\rm p}}(1+z)^3 \tag{1.3.4}$$

The recombination rate directly depends on the fraction of the ionised gas $x_{\rm HII}$. This quadratic dependency increases recombination in high-density regions of the IGM and can significantly alter the duration of reionisation and the morphology of residual neutral gas (Mao et al., 2019; Bianco et al., 2021b). For every ionised atom that recombines, an additional photon must be reinvested to ionising it again, resulting in a substantial loss in the photon budget to keeping the IGM highly ionised.

In figure 1.9 we show three 714 Mpc per side volume slices of a reionisation outputs simulation computed with a radiative transfer code. This image gives a visual representation of the three component in equation (1.3.1). On the left panel we have the ionised field, corresponding to the left hand side. Red and crimson colours indicate highly ionised hydrogen $x_{\rm HII} \geq 0.8$ and blue colour indicates neutral $x_{\rm HII} \leq 0.2$, while green/aquamarine denotes the transition phase $x_{\text{HII}} \sim 0.5$. The central panel shows the photo-ionisation rate, equation (1.3.2), corresponding to the positive term on the differential equation right hand side. It quantifies the number of ionising photons emitted by the sources and follows their distribution. As we can see regions with the largest contributions, $\langle\Gamma_{\text{ion}}\rangle > 2.5 \times 10^{-11}\, s^{-1}$, resides in highly ionised regions. Finally, the right-most panel show the number of recombination per second, equation (1.3.3). The factor corresponds to the negative on the right hand side of equation (1.3.1), it correlates with the density distributions and achieves high recombination in the denser regions, $\langle\mathcal{R}\rangle > 100\, s^{-1}$. Both right hand side quantities shown in figure 1.9 are averaged on the simulation volume resolution, $V_{\text{cell}} = 13.5\,\text{Mpc}^3$.

### 1.3.1 Source Term

Emission of atomic line radiation in IGM can drastically cool the baryonic gas that has elevated viral temperature down to $T_{\text{IGM}} \simeq 10^4$K, allowing the gas to lose thermal energy and fall into the gravitational potential well of dark matter halos. Therefore, a common assumption supported by observational evidence is that each dark matter halo hosts a star-forming galaxy. Hence, the cumulative number of ionizing photon $\dot{N}_\gamma$ produced by the sources per hydrogen atom can be estimated based on the hosting halo mass and is given by (e.g. Furlanetto, 2006a; Choudhury, 2009; Pritchard & Loeb, 2012b)

$$\dot{N}_\gamma = \xi\, \frac{\mathrm{d}f_{\text{coll}}}{\mathrm{d}t}(M > M_{\text{min}}) \tag{1.3.5}$$

where $\xi$ is the efficiency factor and $f_{\text{coll}}$ is the fraction of baryonic matter that collapsed into halos with mass above a certain minimum $M_{\text{min}}$.The efficiency factor is defined as $\xi = f_*\, f_{\text{esc}}\, N_{\text{i}}$. Here $f_*$ is the mass fraction of baryonic gas converted into stars. $f_{\text{esc}}$ is the fraction of ionizing photons that escape into the IGM and $N_{\text{i}}$ quantify the number of ionizing photons produced per atom converted into stars, which depends on the initial mass function of (IMF) of the stellar population. Here the collapsed fraction $f_{\text{coll}} \equiv f_{\text{coll}}(M > M_{\text{min}})$ is defined by (e.g. Monaco, 1997; Haiman & Holder, 2003)

$$f_{\text{coll}} = \frac{1}{\rho_{\text{m}}} \int_{M_{\text{min}}}^{\infty} M\, n(M, z)\, dM \tag{1.3.6}$$

where $\rho_{\text{m}} \equiv \rho_{\text{c},0} \cdot \Omega_{\text{m}}$ is the comoving matter density, $M$ is the halo mass and $n(M, z)dM$ is the comoving number density of halos with masses within the range $(M,\, M + dM)$ at redshift $z$. The minimum mass for halos that host star-forming galaxies is $M_{\text{min}}$, in general this corresponds to the mass for the viral temperature $10^4$ K at which hydrogen atomic line emission becomes the main cooling mechanism (see discussion in § 1.1.2). The halo number density distribution can be acquired from analytical prescriptions derived from numerical N-body simulations (e.g. Jenkins et al., 2001; Tinker et al., 2008; Watson et al., 2013).

for escaping fraction of ionising radiation (Ferrara & Loeb, 2013; Gnedin et al., 2008)

### 1.3.2 Sub-grid inhomogeneity and Clumping Factor

Depending on how gas density fluctuations vary in space and over time (local degree of *"clumpiness"*), the recombinations in the IGM can significantly affect the progress and nature of the reionization process. For every ionised atom that recombines with a free electron, an additional ionizing photon should be produced in order to ionize it again and keep the IGM highly ionised. In this way potentially a substantial portion of the sources photon budget could be depleted. In equation (1.3.3) we defined the recombination rate for the case of a pure hydrogen gas. For a region of volume $V$ we can calculate the volume-weighted averaged recombination rate as

$$\langle\mathcal{R}\rangle_V = \frac{1}{V}\int_V \alpha_{\rm B}(T)\, n_{\rm HII}^2(\mathbf{r})\, d^3r = \alpha_{\rm B}(T)\, \left\langle n_{\rm HII}^2\right\rangle_V \tag{1.3.7}$$

This indicates the number of electron-proton recombination per second in the volume. Here for simplicity, we assumed that the temperature is constant throughout the volume, and so the atomic recombination coefficient $\alpha_{\rm B}$ can be considered position-independent. However, it is important to note that this assumption is not justified when the volume $V$ is large enough to include a considerable amount of radiative sources and these start to ionise and heat the surrounding IGM.

The volume-weighted averaged recombination rate has a quadratic dependency with the enclosed ionised gas fraction. This relation can be factorised out by the so-called Clumping Factor approach, where the volume-averaged of the density squared $\left\langle n^2\right\rangle$ is approximated by the average density squared $\langle n\rangle^2$ and a proportional term, namely the clumping factor, defined as (e.g. Gnedin & Ostriker, 1997; Kaurov & Gnedin, 2015)

$$C_{\rm HII} = \frac{\langle n_{\rm HII}^2\rangle}{\langle n_{\rm HII}\rangle^2} = \frac{\langle x_{\rm HII}^2\rangle}{\langle x_{\rm HII}\rangle^2} \tag{1.3.8}$$

this term is an indicator that gauges how much the gas is agglomerated in structures under gravitational collapse, hence, it can be considered as an indicator for unresolved (sub-grid) structures within the volume. Such that equation (1.3.7) becomes

$$\langle\mathcal{R}\rangle_V = \alpha_{\rm B}(T)\, C_{\rm HII}\, \langle n_{\rm HII}\rangle_V^2 \tag{1.3.9}$$

We illustrate with an example the importance of the clumping factor. In figure 1.10, we consider two volumes with the same amount of enclosed averaged gas density but

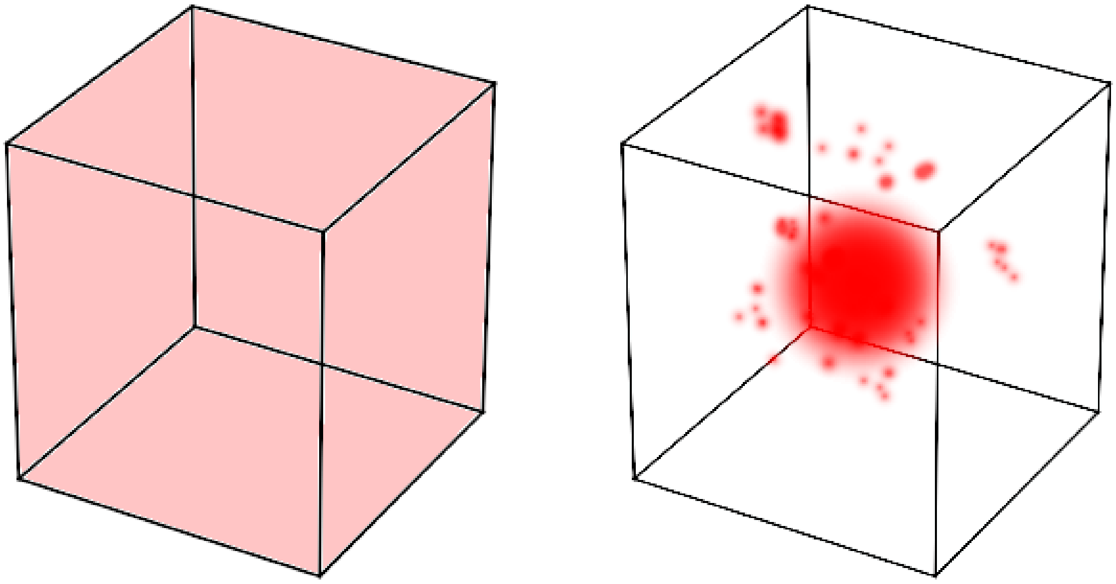

Figure 1.10: Two visual representations of gas in IGM within a few teens Mpc scale cubic volume. We consider the same enclosed average density (red shadow) in both cases but with different spatial distributions. *Left panel*: the gas is uniformly scattered onto space (generally assumption in simulations). *Right panel*: the gas is clustered around the high-density peak. Although these two volumes enfold the same amount of gas, they differ in the recombination rate due to higher clustered regions in the cubic volume on the right panel.

different spatial distribution. On the left panel, for uniformly distributed gas the two quantities $\langle n_{\mathrm{HII}}^2 \rangle$ and $\langle n_{\mathrm{HII}} \rangle^2$ are equivalent and therefore the correction term is close to unity $C_{\mathrm{HII}} \simeq 1$. This condition is true only at high redshift when density fluctuations are relatively small or under-dense regions in the late Universe. However, at lower redshift, as structures formation take part, the gas starts to cluster in higher density regions, as in figure 1.10 right panel. Consequently, the two terms grows apart $\langle n_{\mathrm{HII}}^2 \rangle \gg \langle n_{\mathrm{HII}} \rangle^2$ resulting in a larger clumping factor up to a few hundreds of magnitude. Although the two volumes have the same gas density, the latter case will experience a higher recombination rate since the free electron and proton are locally closer to each other.

Consequently, if not correctly treated, the clumping factor approach can under- overestimate the importance of sub-grid inhomogeneities on radiation absorption and recombination in the IGM, altering the reionisation process and the sources ionising photon budget. Therefore, the clumping factor can be applied as a correction term in low-resolution simulations. In chapter 2 we provide a detailed comparison between clumping factor models, and we present a thorough treatment of the clumping factor approach for the calculation of sub-grid inhomogeneities recombination in large reionisation simulations.

## 1.4 Simulating the Epoch of Reionization

In the absence of direct observation, numerical simulations play a crucial role in understanding the reionisation process's underlying physics. Moreover, reionisation is a process that occurs on a wide range of scales. For example, the physics of the radiation sinks and ionising sources occur on a scale between a few parsecs up to a few hundred kpc sizes. On the other hand, we require large comoving scales $\geq 100\,\mathrm{Mpc}$ to account for the abundance of sources (Iliev et al., 2014) and the long mean free path of soft X-ray photons with energy $E_\gamma \sim 0.1\,\mathrm{keV} - 12\,\mathrm{keV}$ (Trümper & Hasinger, 2008). Over the last two decades, proprieties of IGM at low redshift has been broadly studied through the GP effect (Rauch, 1998; Fan et al., 2006a). Meanwhile, significant progress has been made in modelling and solving the radiative transfer equation of ionising fronts from point sources (e.g. Ciardi & Ferrara, 1997; Iliev et al., 2006, 2009; Zahn et al., 2011).

In this chapter, we present the two main numerical approaches employed to simulate the EoR. We distinguish two main methods, in § 1.4.1 we present the fully-numerical simulation, and in § 1.4.4 we discuss the semi-numerical alternative.

### 1.4.1 Fully numerical Approach

Fully numerical simulations provide the most correct and physical approach to study the reionisation process by precisely solving the equation (1.3.1) for several elements of the primordial gas (i.e. H I , H II , He I, He II and He III). These simulations can run simultaneously N-body and hydrodynamic algorithm to solve the radiative transfer equation for the collision dynamic of cosmic gas for a better representation of galaxy formation (e.g. Springel et al., 2005; Rosdahl et al., 2013; Vogelsberger et al., 2014; Weinberger et al., 2020). In figure 1.11 right panel, and example of a 6 Mpc/h per side slice of the neutral gas density distribution for a radiation hydrodynamics with adaptive mesh refinement simulation (`RAMSES-RT` code: Rosdahl et al., 2013). On the left panel, a zoom-in on a region of 500 kpc/h show the same density distribution and the position of the dark matter halos (orange circles).

The wide range of scales combined with the minimum number of particles required to constrain relevant mass resolution make this method computationally expensive and time-consuming. Over the years, better performing algorithm and increasing computational power in high-performance computers exponentially increased the number of particles

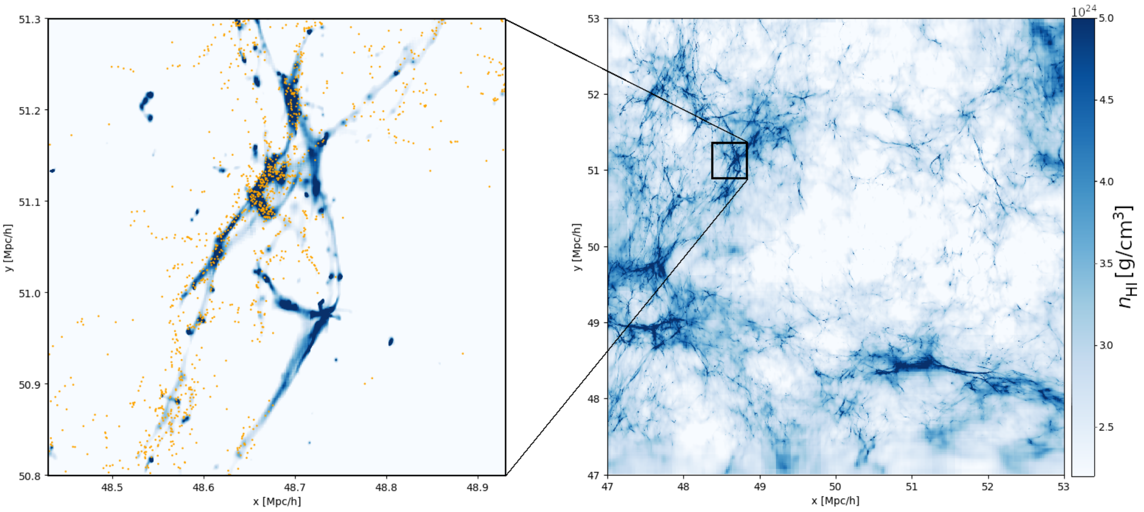

Figure 1.11: Slice example for the RAMSES-RT code at redshift $z = 8.892$. *Left*: a zoom into a region of size 500 kpc/h per side. In blue the neutral hydrogen density distribution and with orange circles the position of the halos identified with ROCKSTAR halo finder (the circles radius are not in proportion to the actual size of the halo). *Right*: the same density distribution but on a larger region of size 6 Mpc/h per side.

employed in high-resolution N-body simulations. As a reference, in figure 1.12 we show the increases of particle number employed in high-resolution N-body simulations over the years of publication.

On the other hand, a computationally cheaper approach consists of separate reionisation simulations into three steps. At first, we employ N-body simulations to keep track of the evolving dark matter density ignoring the gas dynamics. Subsequently, we identify the collapsed structures that host star-forming galaxies with a halo-finder algorithm and apply radiative transfer (RT) codes to simulate the ionising photon propagation into the neutral IGM. However, this approach requires an analytical model that assigns the distribution and hydrodynamic of gas, especially at small scales.

### 1.4.2 N-body Simulations

Cosmological observations provide constraints on the parameter for the ΛCDM universe (e.g. Hinshaw et al., 2013; Planck Collaboration et al., 2020; Valentino et al., 2020; Abbott et al., 2021) and for a prescribed transfer function of the primordial power spectrum, numerical simulations can generate a random Gaussian filed for density matter distribution (e.g. Lewis et al., 2000; Hahn & Abel, 2011). The growth of cosmological structure can then be derived from the first-order Lagrangian perturbation theory (LPT) (Zel'Dovich,

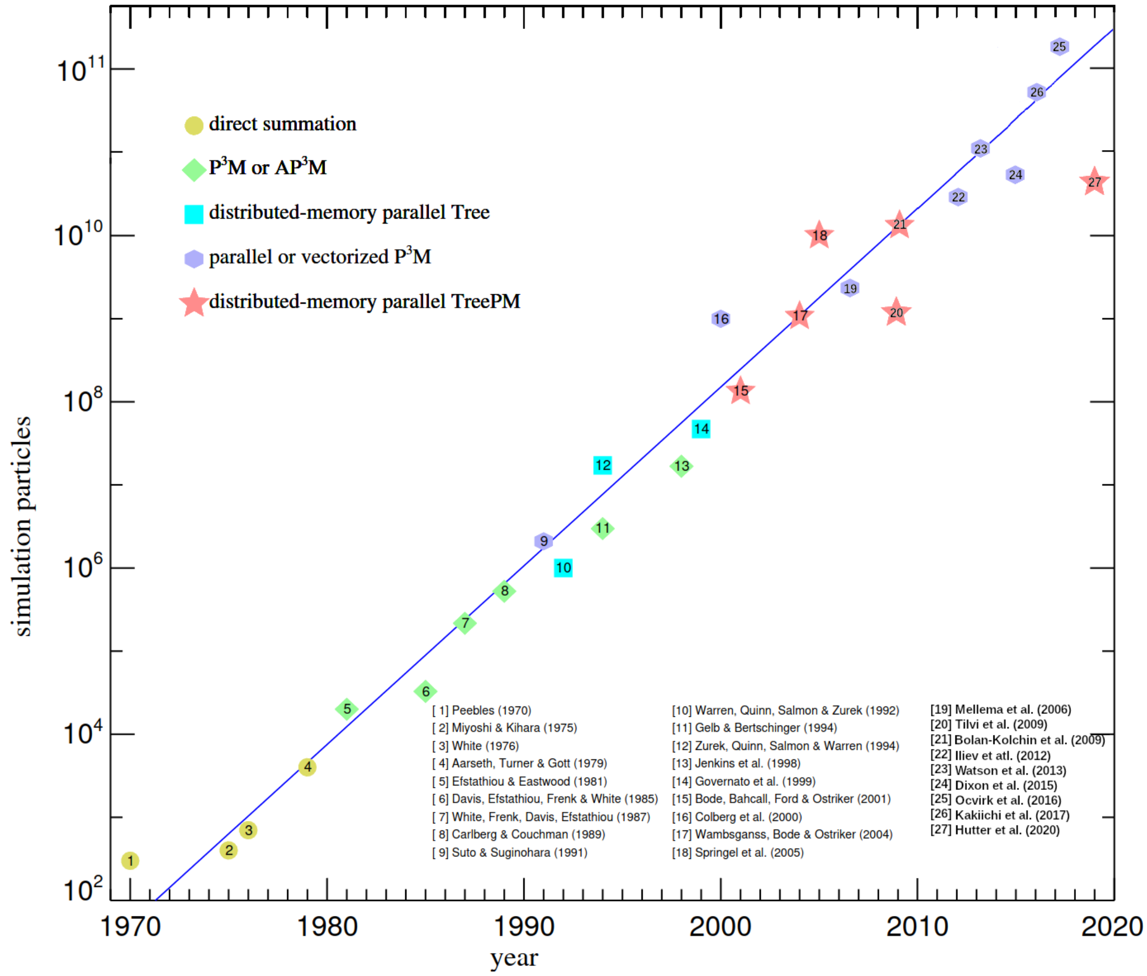

Figure 1.12: The number of particles in N-body simulations over the year of publication showing exponential growth, the different symbol indicates the computational algorithm that implements the particle interactions. This diagram is taken from Springel et al. (2005). We extended the chart by adding the most recent N-body simulation results that increased the simulated particle of a few orders of magnitude after 2005.

1970). However, gravitational instability during cosmic expansion increases the amplitude of density perturbations, and the theory breaks down at small scales and relatively early redshift (Scoccimarro et al., 1998; White, 2014, 2015). For this reason, the cosmological initial condition for particles position and velocity are computed with LPT until redshift of a few hundred, such that N-body simulations can start computing the growth and evolution of the matter fluctuations from redshifts well after the recombination era ($z \sim 1100$).

N-body simulations compute the matter density distributions and the growth of structure in an expanding universe. To simulate the reionisation history, we require larger volumes of a few hundred Mpc per side (Iliev et al., 2014) that accounts for the abundance of sources (Trac & Gnedin, 2011) and rare objects like for high-redshift QSOs. It also provides a broader range of density, with larger and deeper voids as well as higher overdense regions. If we consider $N_{\text{part}}$ the total number of N-body particles and $V_{\text{box}} = L^3_{\text{box}}$ the simulated volume, with side length $L_{\text{box}}$ in comoving units, the mass resolution depends on $N_{\text{part}}$, $V_{\text{box}}$ and cosmological parameters, defined by

$$M_{\text{part}} = \frac{M_{\text{box}}}{N_{\text{part}}} = \rho_{c,0}\,\Omega_m\,\frac{L^3_{\text{box}}}{N_{\text{part}}} \tag{1.4.1}$$

Here, $\rho_{\text{c},0} \equiv \rho_{\text{c}}(z = 0)$ is today critical density, calculated with equation (1.1.5). If we consider a simulation able to compute $N_{\text{part}} \sim 10^{11}$ particles, in a volume of $L_{\text{box}} =$ 500 Mpc per side, the mass particle would then be $M_{\text{part}} \sim 1.6 \times 10^8\,\text{M}_\odot$. Halo finder algorithms are then employed to identify density peaks and the N-body particles that are gravitationally bounded to it, these high density regions are collapsed structures that could host star-forming galaxies (e.g. Knollmann & Knebe, 2009; Sutter & Ricker, 2010; Rasera et al., 2010; Behroozi et al., 2012; Harnois-Déraps et al., 2013; Watson et al., 2013). A minimum of 50 N-body particles are then group together to constitute a dark matter halo (Knebe et al., 2011). These halos are then employed in Radiative Transfer simulations as the source term of ionising radiation.

In numerical simulations halos are considered hosting the potential sources of ionizing radiation (see discussion in § 1.1.2). Therefore, the cumulative number of ionising photon per hydrogen atom $\dot{N}_\gamma$, equation (1.3.5), produced by these sources is related to the halo mass $M_{\text{halo}}$ and is defined as

$$\dot{N}_\gamma = f_*\,f_{\text{esc}}\,N_{\text{i}}\frac{M_{\text{halo}}\,\Omega_{\text{b}}}{t_{\text{s}}\,\mu\,m_{\text{p}}\,\Omega_{\text{m}}} \tag{1.4.2}$$

where $m_{\text{p}}$ the proton mass, $\mu$ is the mean molecular weight of the gas and $t_{\text{s}}$ is the sources lifetime. The first three term are the components of the efficiency factor, as presented in § 1.3.1. The fraction in the right-hand-side gives the number of baryons in the typical volume per halo particle divided by sources lifetime, derived from equation (1.4.1). Haloes hosting ionizing sources are named atomically-cooling halos (hereafter ACHs) and can be further divided into two mass ranges. Haloes with mass above $M_{\text{halo}} \geq 10^9\text{M}_\odot$ are considered high mass halos (HMACHs). These halos are massive enough to continue to accrete gas from the surrounding IGM, even when this is photo-heated and -ionised by sources, at temperature several order of magnitude above $10^4$ K. Therefore, their efficiency factor remains substantially unchanged throughout reionisation. Haloes with mass above $M_{\text{halo}} \geq 10^8\text{M}_\odot$ are considered low mass halos (LMACHs). They are affected by the radiative feedback, resulting in complete or partial suppression of their star-forming rate depending on their mass (Dixon et al., 2016).

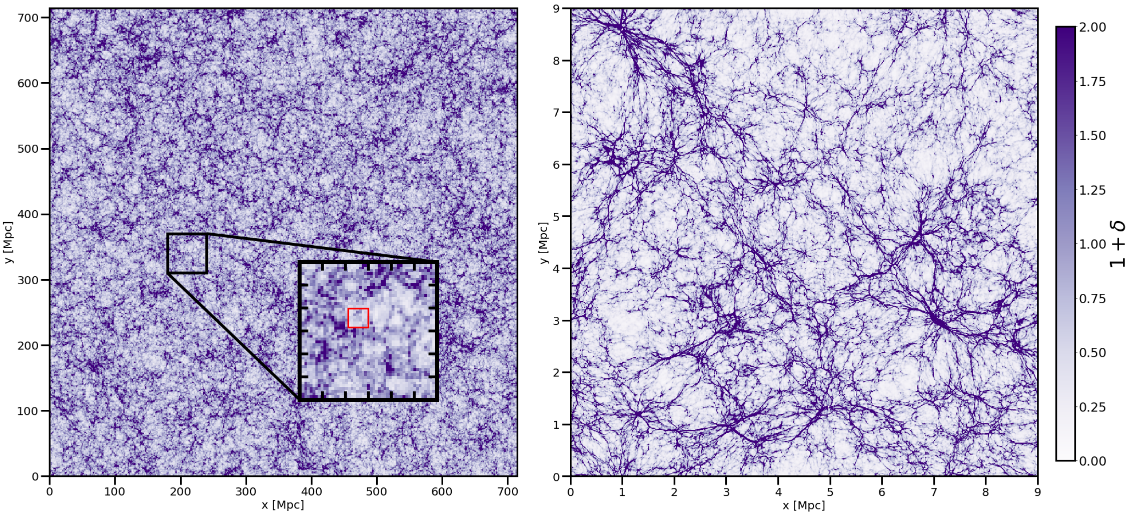

Figure 1.13: Example of the density distribution for the CUBEP$^3$M code. *Left*: for a large box, with volume of $(714\,\mathrm{Mpc})^3$ and coarse-resolution of 1.2 Mpc per side. An inset panel shows the overdensity field on a zoom-in region of size 80 Mpc per side. *Right*: overdensity field for a small, high-resolution simulation, with respectively volume of $(9\,\mathrm{Mpc})^3$ and resolution 7.5 kpc per side.

The computational limitation inevitably restrains the minimum mass of dark matter halos resolved in our simulations. In the example mentioned above, for $N_{\mathrm{part}} \sim 10^{11}$ and $L_{\mathrm{box}} = 500\,\mathrm{Mpc}$ the halos we could find in our simulation would then have mass above $M_{\mathrm{halo}} \geq 8 \times 10^9\,\mathrm{M}_\odot$. This limitation becomes problematic when simulating the larger scale or cosmic reionisation ($> 100\,\mathrm{Mpc}$) since we do not resolve single sources (e.g. galaxies, stars, QSOs, AGN, etc.), instead halos with masses of several orders of magnitude larger. Moreover, computing the ionising photon propagation is time-consuming. Therefore, it is required to further smooth the N-body and halo finder results into coarse grid-mesh (Shapiro et al., 1996). This approach requires analytical prescriptions to model the unresolved (sub-grid) quantities. Several efforts have been made to perform numerical simulations that combine the astrophysical and physical processes at small scales ($\leq 1\,\mathrm{Mpc}$), which resolve the physics of all known radiation absorbent, with the large scale simulations, of the typical size of a few hundred comoving Mpc. Some of these prescriptions account for the sub-grid inhomogeneous recombination in IGM (e.g. Mao et al., 2019; Bianco et al., 2021b), the number and distribution of low mass halos (Ahn et al., 2015a), the radiative contribution by QSOs (Ross et al., 2019) to heat the IGM and the effect of radiation suppression on ionising sources (Dixon et al., 2016). In figure 1.13 we show two slices of the N-body density distribution of the IGM computed with CUBEP$^3$M code. On the left panel, a large volume of $(714\,\mathrm{Mpc})^3$ with a coarse spatial resolution of 2.381 Mpc. As a visual example of the dynamic range in reionisation simulations, the inset panel zooms on a region of 60 Mpc per side. The red square corresponds to the simulated volume of the density slice on the right panel, of volume $(9\,\mathrm{Mpc})^3$ and with a resolution of 7.5 kpc.

### 1.4.3 Radiative Transfer Simulation

Radiative Transfer (RT) simulations solve the evolution of the ionizing radiation field. When a ray passing through a medium of absorbers along the direction $\hat{\mathbf{n}}$, the variation in the specific intensity $I_\nu \equiv I(\mathrm{t}, \mathbf{x}, \mathbf{n}, \nu)$ changes accordingly to the following differential equation (e.g. Rybicki & Lightman, 1986; Abel et al., 1999; Gnedin & Abel, 2001)

$$\frac{1}{c}\frac{\partial I_\nu}{\partial t} + \frac{\hat{\mathbf{n}}\cdot\nabla I_\nu}{\bar{a}} - \frac{H(t)}{c}\left(\nu\frac{\partial I_\nu}{\partial \nu} - 3\,I_\nu\right) = j_\nu - \alpha_\nu\,I_\nu \tag{1.4.3}$$

where $\bar{a} \equiv \bar{a}(t) = a(t)/a_{\mathrm{ems}}$ is the redshift divided with the redshift of the photon emitted at $z_{\mathrm{ems}}$ with frequency $\nu$ and it takes into account the cosmic expansion. Here $j_\nu$ is the emission coefficient, defined as the energy emitted per unit time and $\alpha_\nu$ is the mass absorption coefficient of a gas cloud, also known as opacity coefficient.

This equation describes a seven-dimensional space problem, with two angular coordinates, three-position and one time and frequency. Given the absorption and emission coefficient, we could, in principle, solve exactly equation (1.4.3). However, if we consider $N_{\mathrm{RT}}$ to be the number of radiating sources in our simulation, the direct solution scale with the number of sources and requires $\mathcal{O}(N_{\mathrm{RT}}^{5/3})$ operations per frequency $\Delta\nu$ and time step $\Delta\mathrm{t}$. Therefore, RT codes are often run on post-processed N-body simulations density field (e.g. Iliev et al., 2006; Trac & Cen, 2007; Mcquinn et al., 2007; Ciardi et al., 2012; Dixon et al., 2016) and valid approximations can reduce the number of operations to scale close to linearly and become computationally feasible.

Several application are proposed and studied (Iliev et al., 2009), we can distinguish ray-tracing methods (e.g. Abel et al., 1999; Mellema et al., 2006a), moment based (e.g. Aubert & Teyssier, 2008), Monte Carlo approximation (e.g. Ciardi et al., 2001; Ghara et al., 2018), local depth approximation (e.g. Gnedin & Ostriker, 1997). In this work, we focus on the approach adopted by the $\mathtt{C^2RAY}$ simulation (Mellema et al., 2006a; Friedrich et al., 2012). A photo-conserving radiative transfer code that casts rays for each sources onto a coarsened density grid. By assuming that the ionizing photon mean free path $R_{\mathrm{mfp}}$ is much smaller than $c\,\Delta t$ (e.g. Gnedin & Abel, 2001), and that all ionizing radiation from recombination is absorbed locally $j_\nu = 0$, we can simplify equation (1.4.3) such that

$$\hat{\mathbf{n}}\cdot\nabla I_\nu = -\alpha_\nu\,I_\nu \tag{1.4.4}$$

where the solution is

$$I_\nu = I_{\nu,0}\, e^{-\int \alpha_\nu(s)\, ds} = I_{\nu,0}\, e^{-\tau_\nu} \tag{1.4.5}$$

here the ionizing intensity decreases exponentially by the absorption coefficient $\alpha_\nu$ along the direction of radiation propagation. Is common to express the solution, equation (1.4.5), by the optical depth defined as $d\tau_\nu = \alpha_\nu\, ds$, measured along the line of sight.

### 1.4.4 Semi-Numerical Approach

As we mentioned in the previous chapter, § 1.4.1, fully numerical simulations provide a self-consistent model of the astrophysical and physical process in reionisation. However, they are prone to computational limitations. With a series of nuanced approximations, a semi-numerical approach can increase the computational efficiency and enable the simulation of large cosmological volume, otherwise unattainable by fully numerical approaches. The main approach to derive the growth of H II regions is the excursion set formalism. This method was first applied to reionisation by Furlanetto et al. (2004), based on an analogous method in the context of galaxy formation (e.g. Bond et al., 1991; Lacey & Cole, 1994) that employs the Press-Schechter formalism. The basic requirement of ionising sources is that the total number of photon from sources must leastwise match the amount of hydrogen atom in their close surrounding IGM (Trac & Gnedin, 2011). Therefore, a spherical region of the IGM with overdensity $\delta$, radius size $R$ and corresponding mass $M = 4\pi/3\, \rho_{c,0}\, R^3$ is considered fully ionised under the following condition:

$$f_{\rm coll}(M,\, \delta) > \xi^{-1} \tag{1.4.6}$$

Here $f_{\rm coll}$ is the fraction of gas collapsed to form ionising sources and $\xi$ is the efficiency factor. In the case of semi-numerical simulations, analytical prescriptions of the halo number density $n(M)$ are often favoured to halo finder algorithm because of their swiftness (e.g. Mesinger et al., 2010). In the case of the Press & Schechter (1974) mass function the collapsed fraction, equation (1.3.6), takes the form

$$f_{\rm coll} = erf\left[\frac{\delta_{\rm crit}(z)}{\sqrt{2}\,\sigma_{\rm min}}\right] \tag{1.4.7}$$

where $\delta_{\rm crit}(z) = 1.686/D(z)$ is the collapse critical density and $\sigma_{\rm min} \equiv \sigma(M_{\rm min})$. is the density fluctuations variance at the halo minimum mass $M_{\rm min}$ to host star-forming galaxies.

Intuitively, radiative sources must produce enough ionising photons to keep the surrounding IGM plasma warm and ionised to avoid gas recombination. Therefore, the smoothing scale $R$ is established by iterating the radius from a maximum scale $R_{\rm max}$

down to the mesh-grid size $R_{\mathrm{cell}}$, until the condition in equation (1.4.6) is satisfied. This maxima $R_{\mathrm{max}}$ can be set to the mean free path of the ionising photon, which is the largest distance that ionising UV photons can travel before being almost entirely absorbed by the neutral IGM. This quantity can be extrapolated by low redshift observations of Lyman limit absorption systems (e.g. Storrie-Lombardi et al., 1994; Miralda-Escude, 2003; Ribaudo et al., 2011; Prochaska et al., 2015; Shull et al., 2017; Chen et al., 2020a; Becker et al., 2021; Cain et al., 2021).

## 1.5 Machine Learning and Neural Networks

Machine learning (ML) is a field of data science that aims to model and understand physical processes by determining feature patterns from large amounts of data with iterating updating algorithms. The idea is to mimic the learning process in biological organisms. The first concept of machine learning started in the field of computational neuroscience. Several studies formulated different mathematical expressions that explained and imitated, to some extent, the information processing in biological systems (e.g. McCulloch & Pitts, 1943; Widrow & Hoff, 1960; Rosenblatt, 1962). From these early studies, it was possible to develop a statistical approach for recognising patterns in data analysis, formulating the first *multilayer perceptron*, or nowadays commonly known as neural networks.

A first practical application of machine learning algorithm was performed by the pioneering work of Samuel (1959), where he developed a program that progressively increased its ability to play the board games of draughts. Several other pioneering works further enhanced the domain of Artificial Neural Network (ANN). As in the case of the self-optimising algorithm, started by Linnainmaa (1976), which put the basis for the error back-propagation formalism (Rumelhart & Zipser, 1985) (discussion in § 1.5.2). The work of Fukushima (1980), later inspired the development of image input neural network (e.g. Le Cun et al., 1997), we present its conceptual idea in § 1.5.3. At the end of the nineteenth century, several other works further improved and modernised ANN with the help of more complex and performant algorithms (e.g. Hopfield, 1982; Tesauro, 1995; Hochreiter & Schmidhuber, 1997), based on the nature of the problem. On the other hand, the availability of vast and more complete datasets standardised the neural network benchmark (e.g. Deng et al., 2009; LeCun & Cortes, 2010), allowing a better and unbiased comparison between models.

The scope of this chapter is to provide an overview of the central concept of the machine learning application. First, we present the basic framework of a neural network

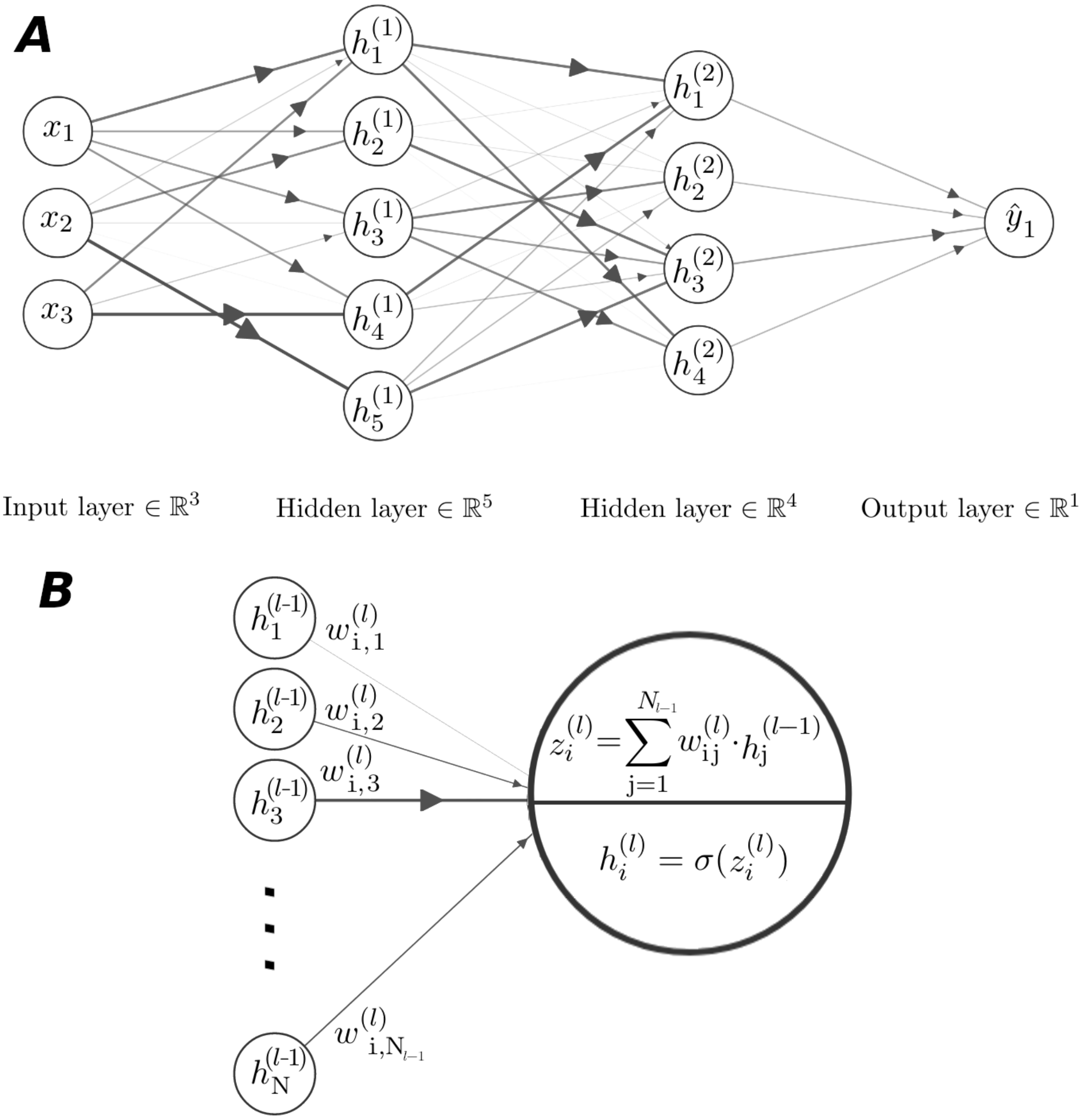
A
$x_1$
$x_2$
$x_3$
$h_1^{(1)}$
$h_2^{(1)}$
$h_3^{(1)}$
$h_4^{(1)}$
$h_5^{(1)}$
$h_1^{(2)}$
$h_2^{(2)}$
$h_3^{(2)}$
$h_4^{(2)}$
$\hat{y}_1$
Input layer $\in \mathbb{R}^3$
Hidden layer $\in \mathbb{R}^5$
Hidden layer $\in \mathbb{R}^4$
Output layer $\in \mathbb{R}^1$
B
$h_1^{(l\text{-}1)}$
$w_{\mathrm{i},1}^{(l)}$
$h_2^{(l\text{-}1)}$
$w_{\mathrm{i},2}^{(l)}$
$h_3^{(l\text{-}1)}$
$w_{\mathrm{i},3}^{(l)}$
$h_{\mathrm{N}}^{(l\text{-}1)}$
$w_{\mathrm{i},\mathrm{N}_{l-1}}^{(l)}$
$z_i^{(l)} = \sum_{\mathrm{j}=1}^{N_{l-1}} w_{\mathrm{ij}}^{(l)} \cdot h_{\mathrm{j}}^{(l-1)}$
$h_i^{(l)} = \sigma(z_i^{(l)})$

Figure 1.14: *Panel A*: Example of the architecture of a fully connected neural network with two hidden layers. Each circle correspond to a neuron, and the arrow opacity is proportional to the importance of the weight. *Panel B*: An illustration that show a specific neuron inputs $(l)$ based on the previous layer $(l-1)$ outputs. Here we use a `ReLU` activation function $\sigma$ to illustrate the non-linear step.

in § 1.5.1. Then, in § 1.5.2, we introduce the concept of network optimisation with the back-propagation algorithm. Here, we consider the supervised learning method, an iterating training algorithm based on comparison examples between network predictions and inputs associated with known labelled outputs. Finally, in § 1.5.3 we present the Convolutional Neural Network (CNN) model for the recognition of topological features in image inputs, relevant for the results discussed in chapter 3.

### 1.5.1 Feed-Forward Neural Network

The general framework of ANN is divided into three parts, an input layer that consists of a vector $\mathbf{x} = (x_1, \cdots, x_{N_0})$, one or more hidden layers that extrapolate features from the input data, and finally an output layer $\hat{\mathbf{y}} = (y_1, \cdots, y_{N_M})$. The hidden layer is formed with a group of neurons that elaborates the inputs with a succession of linear transformation followed by a non-linear operation called the activation function. ANN can have one or more interconnected hidden layers to form a series of adjoint layers. The linear transformation is generally a dot product with a matrix of trainable weights $W^{(l)} = \left(w_{\mathrm{ij}}^{(l)}\right) \in \mathbb{R}^{N_l \times N_{l-1}}$ where $l = 1, \cdots, M$ is the index number of hidden layers associated with each neuron. This operation returns a vector $\mathbf{z}^{(l)} = (z_1, \cdots, x_{N_l})$ employed as input for the non-linear transformation.

$$h_{\mathrm{i}}^{(l)} = \sigma\left(z_i^{(l)}\right) = \sigma\left(\sum_{\mathrm{j}=1}^{N_{l-1}} w_{\mathrm{ij}}^{(l)} \cdot h_{\mathrm{j}}^{(l-1)}\right) \tag{1.5.1}$$

This equation is referred to as the basic model of a *feed-forward network function*. Here, $\sigma$ is a non-linear function and $\mathbf{h}^{(l-1)} = (h_1, \cdots, h_{N_{l-1}})$ is the output of the prior hidden layer. By definition this would be the input vector in the case of the first step $l = 1$, such that $\mathbf{x} \equiv \mathbf{h}^{(0)}$. Whereas, in the case of the last step $l = M$ this would be the output layer, $\hat{\mathbf{y}} \equiv \mathbf{h}^{(M)}$. In figure 1.14 *Panel A* we show a schematic example of a neural network, with an input $\mathbf{x} \in \mathbb{R}^3$, two hidden layers, of dimension $\mathbf{h}^{(1)} \in \mathbb{R}^5$ and $\mathbf{h}^{(2)} \in \mathbb{R}^4$, with a final scalar output $\hat{\mathbf{y}} \in \mathbb{R}^1$. The arrow thickness indicates the significance of the weight for the following hidden states or output. *Panel B* focus on one neuron and gives a visual representation of equation (1.5.1).

The non-linear operation $\sigma$ can vary depending on the nature of the problem to solve (Aurélien, 2019). Historically this was a step or sigmoid function, but a series of alternative has been proposed in the last decades (e.g. Jarrett et al., 2009; Glorot et al., 2011), that increased the performance of neural networks (Maas et al., 2013). In figure 1.15 we show some of the most commons activation functions used today for neural networks. The top two are a sigmoid and `Softmax` functions respectively, which take any values and return a value between 0 and 1, whereas at the bottom, two examples of Rectified Linear Unit activators (Maas et al., 2013). This last category of activators is largely used today due to its faster and effective training of complex networks.

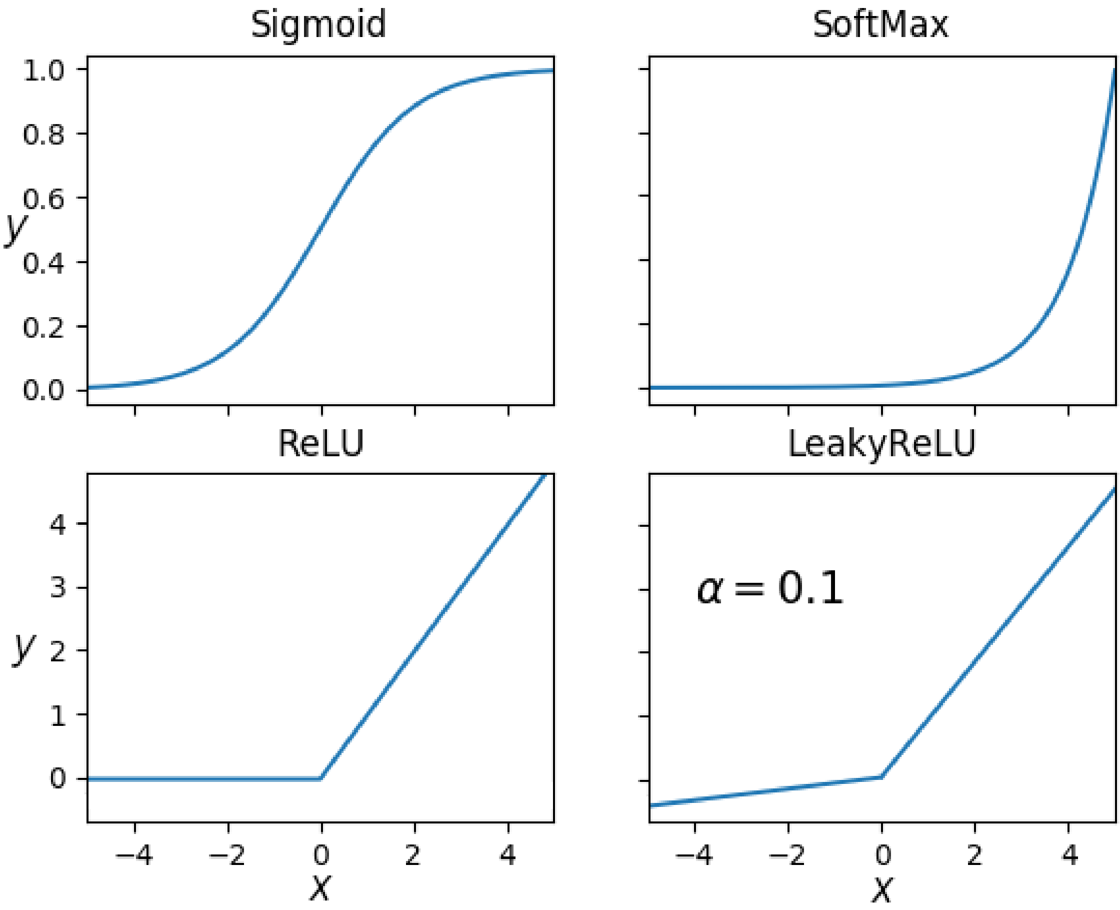

Figure 1.15: Examples of activations functions largely employed in neural networks. Top left, sigmoid function and top right `Softmax`. In these two cases, the result $y$ is a value between 0 and 1. The bottom left, the Rectified Linear Unit function (`ReLU`), return its value for $x \geqslant 0$, while 0 for negative terms. Bottom right the Leaky version of a Rectified Linear Unit (`LeakyReLU`) with gradient $\alpha = 0.1$ for negative terms.

### 1.5.2 Network Training by Error Back-Propagation

The previous section presented the basic structure of the feed-forward network function and how machine learning models approximate any problem as a sequence of linear and non-linear transformations, from an input vector $\mathbf{x}$ to an output $\hat{\mathbf{y}}$, characterised by a series of weight parameters $W^{(l)}$ with index $l = 1, \cdots, M$, also known as the network hyper-parameters. In order to train a network in supervised learning, the weights must be optimised to minimise a certain loss function or error between the network output prediction $\hat{\mathbf{y}}$ and the actual data $\mathbf{y}$. For illustration purposes, here we consider this to be the mean squared error (MSE), and we represent all the network parameters with the

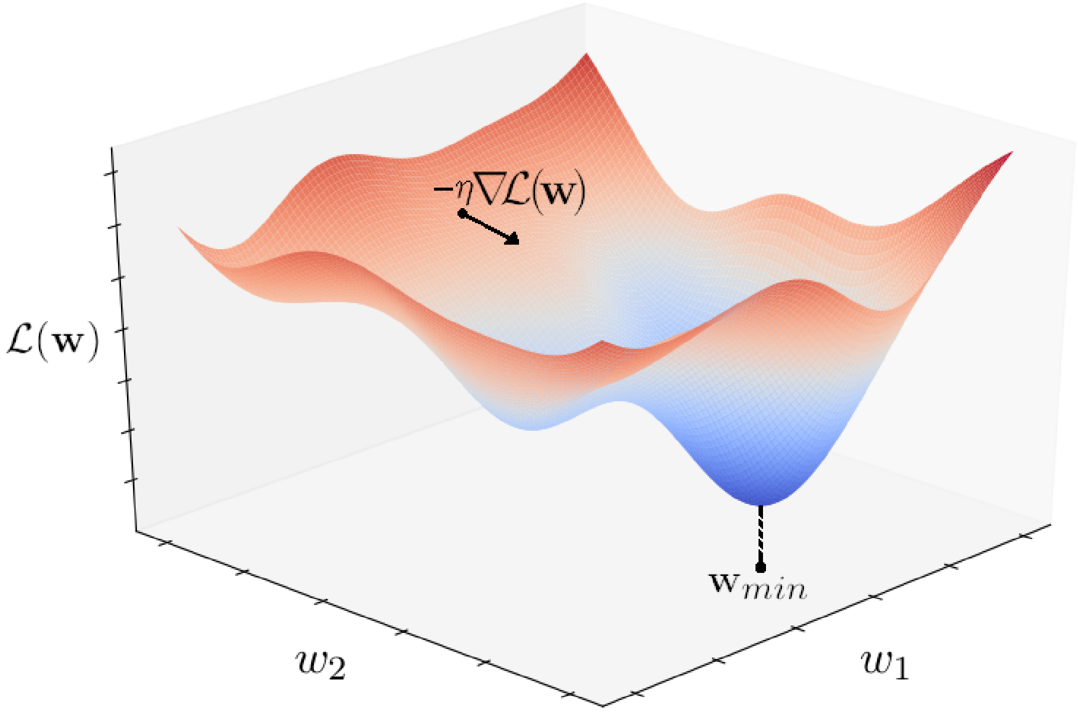
$-\eta\nabla\mathcal{L}(\mathbf{w})$
$\mathcal{L}(\mathbf{w})$
$\mathbf{w}_{min}$
$w_2$
$w_1$

Figure 1.16: An illustration that visualise the loss function $\mathcal{L}$ as a hyper-surface in the multi-dimensional space of the network parameters, in this case, represented by $w_1$ and $w_2$. The algorithm takes a step $\eta$ in the direction of the descending gradient to minimise the network error. This process is repeated until the algorithm finds a global minimum, for instance, a set of parameters $\mathbf{w}_{min}$ that optimise the network.

variable $\mathbf{w}$, such that

$$\mathcal{L}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(y_{\mathrm{i}} - \hat{y}_{\mathrm{i}}(\mathbf{w}))^2 \tag{1.5.2}$$

Here $N$ is the number size of a subsample of the data. The loss function can be visualised as a hyper-surface in the multi-dimensional space of the network hyper-parameters. The idea is to find a set of weights that correspond to the global minima $\mathbf{w}_{min}$ and therefore minimise the error. In figure 1.16 we give a visual representation of this parameter space.

An efficient technique for optimising a network model equation (1.5.1) is the so called *error back-propagation* (Rumelhart & Zipser, 1985). This method is an iterating updating process, and it consists of dividing the set of data into subsamples of size $N$, often referred to as the batch sample. This method iterates through the data, and for each batch, it estimates a set of initials weights. Subsequently, the gradient is evaluated, and a step of length $\eta > 0$ in the parameter space is taken toward the descending gradient $-\eta\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w})$. The quantity $\eta$ is known as the learning rate and it quantifies the rate of convergence. Recent studies propose different methods to optimise the learning speed and avoid the convergence being stuck in local minima (Delyon, 2000; Smith, 2017; Patterson & Gibson, 2017). Finally, the weights are updated accordingly for the next step. This process is repeated until convergence $\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}_{min}) \approx 0$. At that point, the network is considered trained and able to represent the data with a certain level of accuracy. Recent works have investigated alternatives for the gradient descendent algorithm (e.g. Ruder, 2017; Zhang, 2019; Chandra et al., 2019; Gong et al., 2020; Yadav, 2021), and several improvements have been proposed to increased the efficacy and speed of the convergence.

### 1.5.3 Convolutional Neural Network

The linear transformation in the feed-forward network model ensures that each neuron operates its inputs independently and no information is shared with nearby neurons while this elaborates its output. Therefore, this approach is not optimal for multi-dimensional data, as in the case of images and audio input signal, where the topological features in data contain essential proprieties inherent to the underlying physical process.

Convolutional neural networks (CNN) are mainly designed to work on image inputs. The first work employing CNN was performed by Le Cun et al. (1997). However, their popularisation and widespread application were possible only after the work of Krizhevsky et al. (2012), where it presented a network at the *ImageNet 2012 Challenge*, able to correctly identify more than a thousand classes of objects in a large set of high-resolution images and decrease radically the error rate achieved by its predecessors. In the context of CNN, the data are generally three-dimensional tensors with sizes $\mathbb{R}^{H\times L\times W}$, where $H$ and $L$ are the spatial image length in pixels, and $W$ corresponds to the number of feature maps, stacked together to form groups also known as channels. A few examples can be RGB colour images (e.g. Long et al., 2014), multi-waveband data or hyperspectral images (e.g. Wei et al., 2020). Their framework is the same as for the feed-forward network. However, fully-connected hidden layers are replaced by a series of convolutional and or pooling layers, with a characteristic features detector kernel filter. Hence, in equation (1.5.1) the vector input is replaced by an image input. *Panle A* in figure 1.17 provides an example of convolution on a $4 \times 4$ input image, represented as a matrix, with just one feature map for simplicity. We consider a $3 \times 3$ kernel filter $F_c = (f_{\mathrm{ij}}) \in \mathbb{R}^{3\times 3}$ of values.

$$F_c = \begin{pmatrix} f_{00} & f_{01} & f_{02} \\ f_{10} & f_{11} & f_{12} \\ f_{20} & f_{21} & f_{22} \end{pmatrix} = \begin{pmatrix} 0 & 2 & 1 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \tag{1.5.3}$$

The kernel iterates through the feature maps, the element-wise product between the

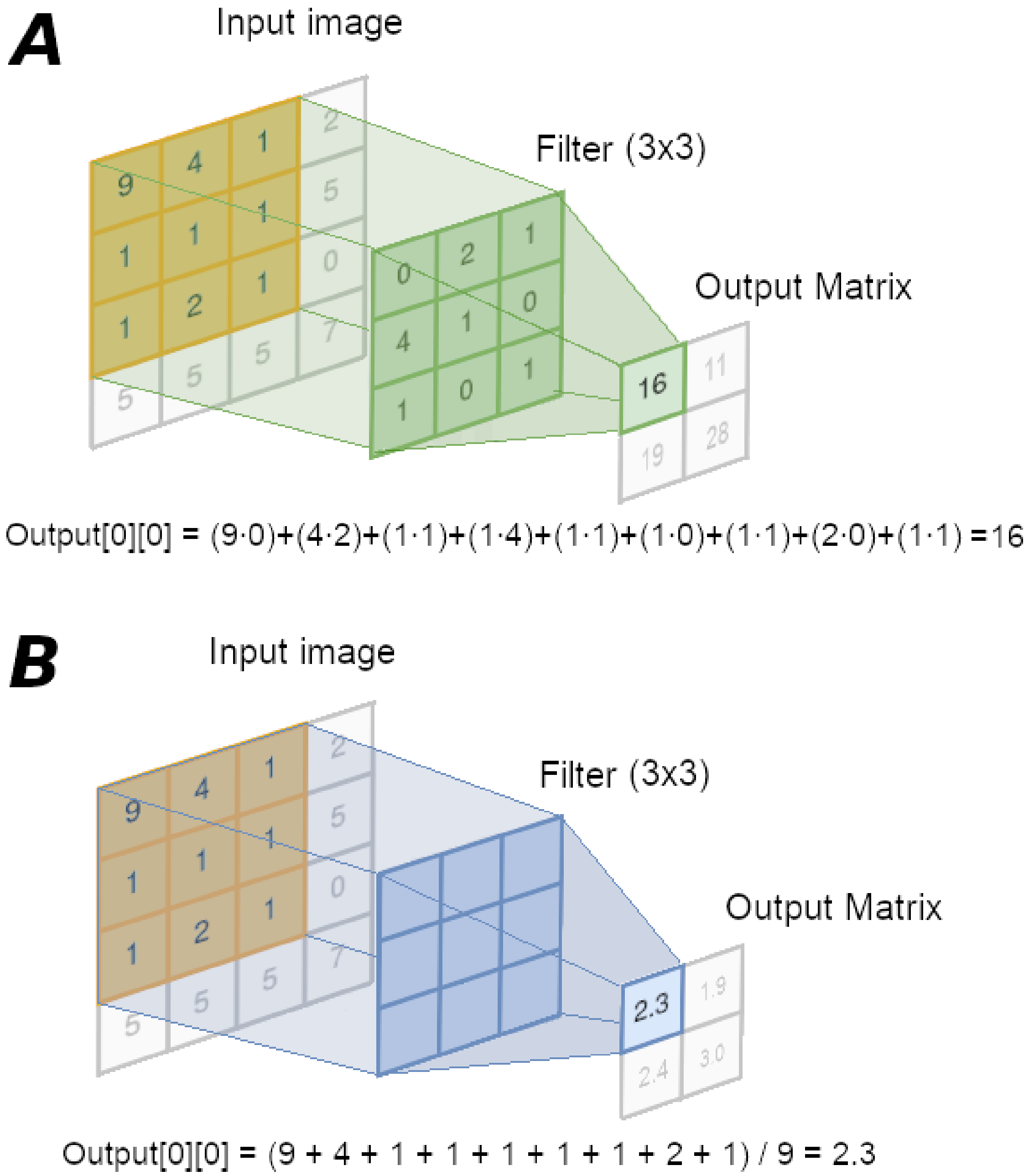

Figure 1.17: Example of two possible operations on an input image, represented as a matrix, in convolutional networks. In *Panel A* a discrete convolutional and *Panel B* an average pooling operation. In both cases, we show in orange the area in the input image where a 3 × 3 kernel filter (in green and blue, respectively) operates. At the bottom of each panel, the mathematical expression represented by the operation. We only show the result of the first index in the output matrix.

filtered area (in orange) and the kernel (in green) is summed up. This procedure is repeated for each feature map present in an input image. In figure 1.18, we show a few examples of kernel filters applied on a 512 × 384 pixels RGB image, with three features

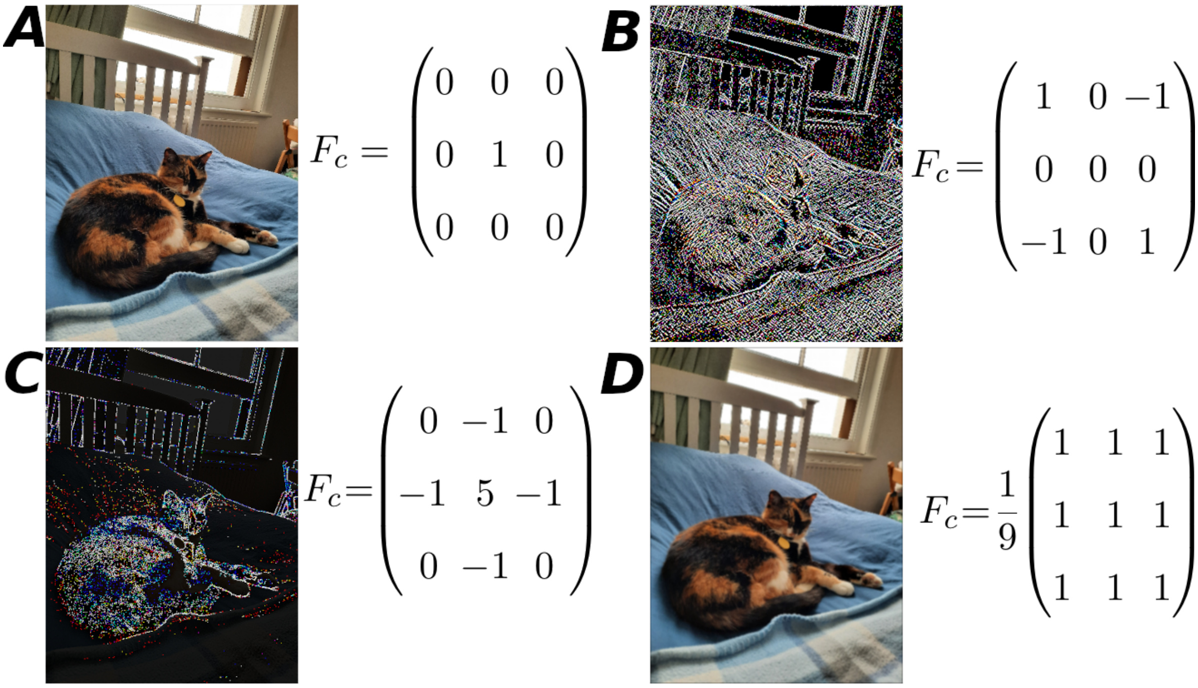
A
$F_c = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
B
$F_c = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}$
C
$F_c = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{pmatrix}$
D
$F_c = \frac{1}{9}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$

Figure 1.18: An example of four filters applied to a 512 × 384 pixels photo of Lichen, the author's neighbour's cat. *Panel A*: the original image. *Panel B*: edge detector filters that emphasises contrast between coloured pixels at the borders of the filtered region. *Panel C*: edge detector filters that focuses on the contrast produced by the light reflectance. *Panel D*: the image is blurred by a normalised box-linear filter.

maps, respectively red, blue and green channels. In particular, *Panel B* and *C* depict two edge detector filters. In the first case, the filtered image emphasises the contrast between coloured pixels at the borders of the filter, while in the second case focus on the contrast produced by the light reflectance in the image. The main difference between convolution and pooling operations, is that the latter apply a function to the filtered area. The standard approach considers max-pooling (Sudholt & Fink, 2017; Huang et al., 2018) or average pooling (He et al., 2015; Oquab et al., 2015) but there are more generalised approaches depending on the nature of the problem (Christlein et al., 2019; Gholamalinezhad & Khosravi, 2020). In figure 1.17 *Panel B*, we show an arithmetical average pooling operation, with a 3 × 3 kernel (in blue) on the same input matrix as in the previous example. In figure 1.18 *Panel D*, we show the filter associated with the averaged pooling operation, also known as a normalised box-linear filter. We apply the pooling operation on the same 512 × 384 pixels RGB image, and the resulting image appears blurred.

In the case of CNN, the linear transformation in equation (1.5.1) is substituted by a discrete convolution or pooling operation. Therefore, based on the example in figure 1.17, the weight matrix for the $l$-th hidden layer with filter $F_c$ takes the form of a Toeplitz matrix with elements.

$$W^{(l)} = \begin{pmatrix} f_{00} & f_{01} & f_{02} & 0 & f_{10} & f_{11} & f_{12} & 0 & f_{20} & f_{21} & f_{22} & 0 & 0 & 0 & 0 & 0 \\ 0 & f_{00} & f_{01} & f_{02} & 0 & f_{10} & f_{11} & f_{12} & 0 & f_{20} & f_{21} & f_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & f_{00} & f_{01} & f_{02} & 0 & f_{10} & f_{11} & f_{12} & 0 & f_{20} & f_{21} & f_{22} & 0 \\ 0 & 0 & 0 & 0 & 0 & f_{00} & f_{01} & f_{02} & 0 & f_{10} & f_{11} & f_{12} & 0 & f_{20} & f_{21} & f_{22} \end{pmatrix} \tag{1.5.4}$$

Here the $\mathbb{R}^{4\times4}$ input matrix is flattened to form a $\mathbb{R}^{1\times16}$ vector, and the output is then reshaped to form a $\mathbb{R}^{2\times2}$ matrix. The weight matrix can be generalised for input images with dimensions $H \times L \times W$, and in that case equation (1.5.4) corresponds to a three-dimensional tensor. Even so, the weight parameters are updated via error back-propagation, with the same method explained in § 1.5.2.

### 1.5.4 Examples of Deep Structured Learning Networks

Modern computer vision requires very large networks, with multiple hidden layers and more than a million trainable weight parameters $W^{(l)}$, to be able to learn patterns and extrapolate features from complex visual inputs. This type of networks are part of the broad category of deep structured learning networks and are employed in various fields in science. In our case, we discuss an application of deep learning networks in the context of cosmic reionisation, illustrated with figure 1.19. This work was carried out by Gillet et al. (2019). They developed a CNN intended to extrapolate the astrophysical proprieties (output astrophysics) of the first sources from the 21-cm tomographic dataset (input layer) produced with a semi-numerical reionisation simulation. In § 1.1.2 and § 1.4.4, we introduced some of the astrophysical parameters that their network recover, such as $\xi$ and $T_{\mathrm{vir}}$, but here, in addition, they consider the soft X-ray emissivity $L_{\mathrm{X}}$ and the energy threshold $E_0$ for X-ray self-absorption in galaxies interstellar medium (ISM).

The input layer consists in a 21-cm tomographic dataset, top panel in figure 1.19, meanwhile at the bottom the four recovered astrophysical parameters, $\xi$, $T_{\mathrm{vir}}$, $L_{\mathrm{X}}$ and $E_0$. We can see that their network is structured in two main parts. The first constitutes a succession of convolutional and pooling layers. The cyan $3 \times 3$ grid represents the iterating kernel filters. The pooling and convolutional operations condensate the input into an array layer (flattening) that contains a compressed representation of the input image. This array is known as *low-resolution latent space* and portrays only the most relevant features encoded in the image. For this purpose, the contracting path is often referred to as the encoder. The discussion about the latent space is relevant later for our result discussion in chapter 3. The second part consists of a fully connected neural

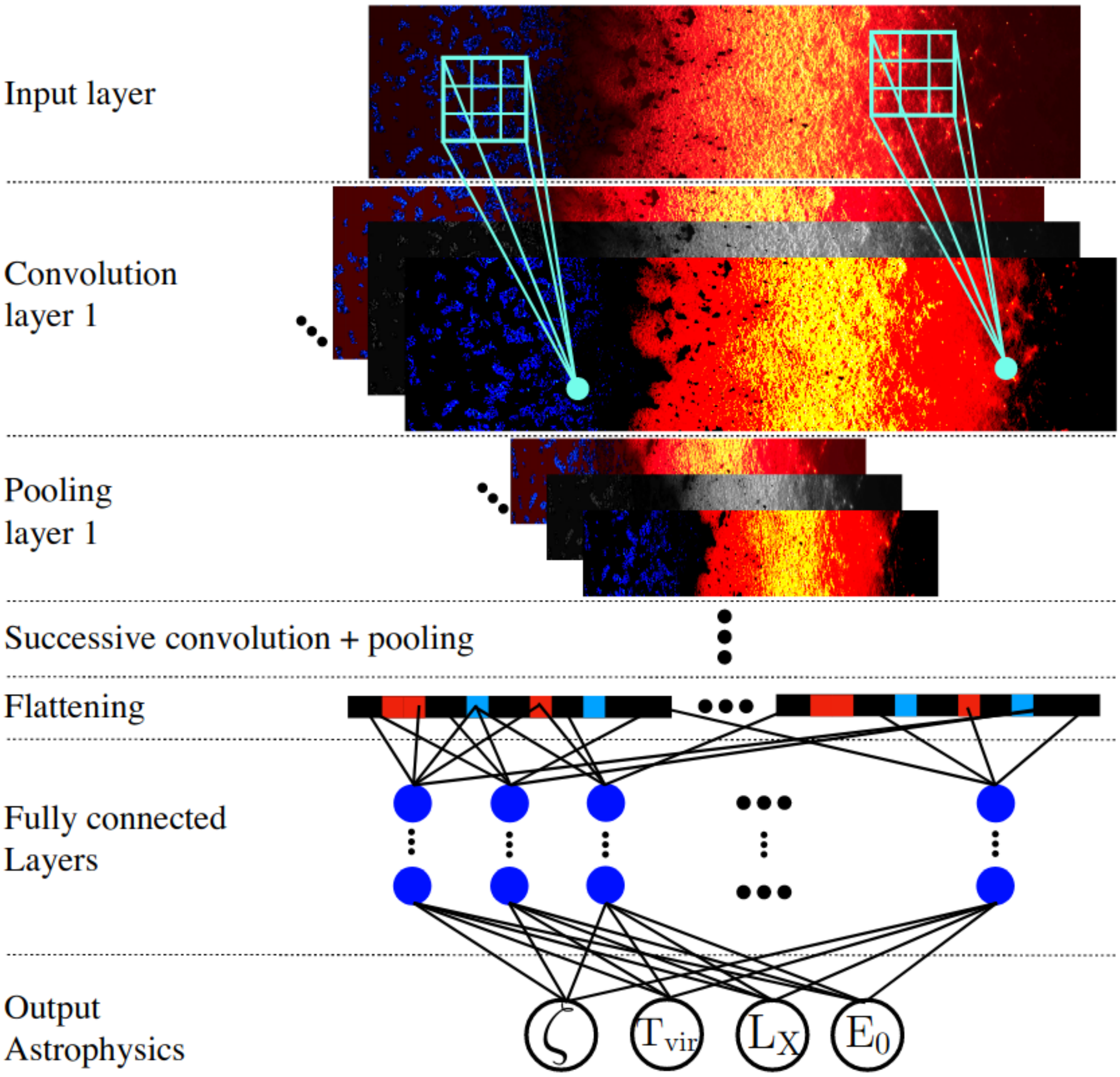

Figure 1.19: Example convolutional neural network from the work of Gillet et al. (2019). This network is particularly designed to extrapolate astrophysical parameters from simulated 21-cm tomographic dataset and it is a perfect example of how different models can be combined together to form a deep neural network.

network that takes the latent space as input and returns the astrophysical output following the same structure of the feed-forward model and back-propagation method presented in § 1.5.1 and § 1.5.2 respectively.

The pioneering work of Long et al. (2014) extended the framework of CNNs. It introduced a deep learning network that is able to associate each image pixels with the objects present in visual inputs. This pixel-wise discernment process is a common problem in the field of data science and is known as image segmentation. In their work, they defined a class of networks named auto-encoders. As shown in the previous example, the primary strategy is to divide the network into two parts. The first constitutes an encoder that

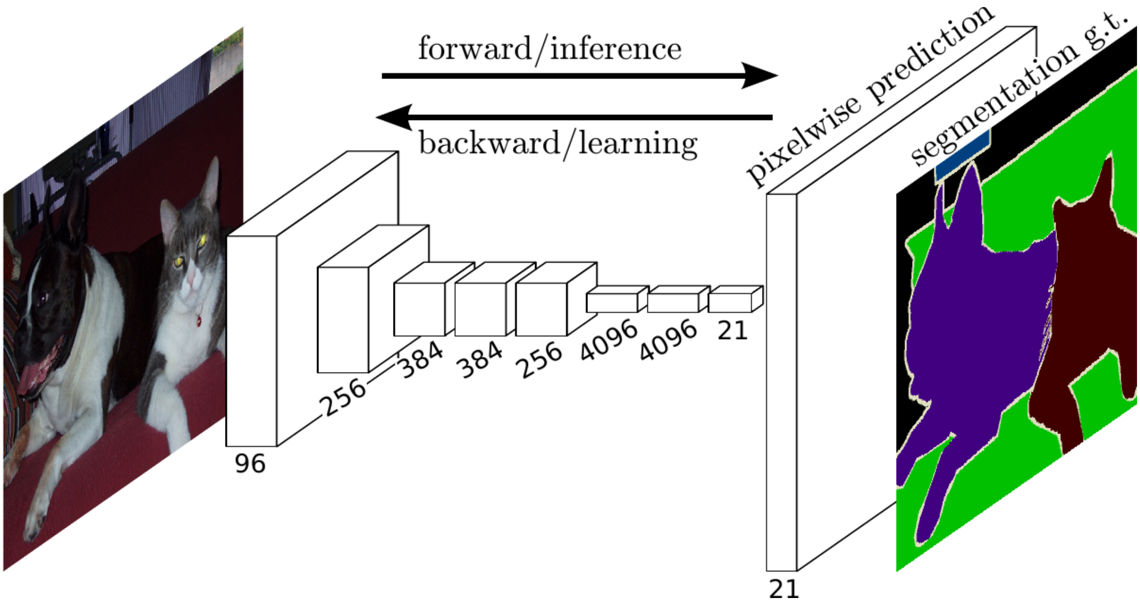

Figure 1.20: Schematic representation of an auto-encoder employed for image segmentation. The visual input is first compressed with a series of convolution and pooling operations that reduce the input to a low dimensional latent space of sizes $10 \times 10$ and 21 feature maps. Then, the result is used to produce a pixel-wise map that distinguishes different objects in the image. This diagram is taken from Long et al. (2014).

contracts the visual input, resulting in a low dimensional latent space.

Contrary to what we have seen in the previous example, the second part consists of an expanding path. This process is also named decoder. Its goal is to use the information compressed in the low dimensional latent space, extrapolate the most relevant image features to reconstruct the input image, and organise pixels in groups. Each pixel is then associated with the different subjects in the picture as faithfully as possible. Analogously to what we discussed in § 1.5.3 this procedure can be expressed mathematically by vector-matrix multiplication. However, in this case, a transpose version of the matrix $W^{(l)}$, equation (1.5.4) would be employed to expand the low dimensional latent space instead of reducing it. For this reason, this operation is often mentioned as transposed convolution, sometimes called deconvolution, and up-sampling operation, respectively.

In figure 1.20 we show the architecture of their network and an example applied to everyday life image. We can see that the $500 \times 500 \times 3$ input image is progressively compressed to form the latent space with dimension $10 \times 10$ and 21 feature maps. The latent space is then up-sampled into a pixel-wise prediction with sizes $500 \times 500 \times 21$. On the right side of the same figure, the resulting output shows a colour base map that distinguishes the different subjects presented in the input image.

## 1.6 Thesis Outline

The primary project of this thesis is to improve the numerical implementations of the sub-grid IGM inhomogeneities in vast cosmological simulations with the $\mathtt{C}^2\mathtt{RAY}$ code. Moreover, we employ an artificial intelligence algorithm to develop a new technique for identifying H II regions from image observation of the 21-cm signal employed in the upcoming Square Kilometre Array telescope.

In section 1.4 we provide an overview of the two primary methods employed today for simulating the epoch of reionisation. Meanwhile, in section 1.5 we introduce the concept of machine learning and artificial neural networks, the notion of error back-propagation for self-updating algorithms and the convolutional neural network model for image segmentation.

In chapter 2 we present a new approach that correctly quantifies the effect of local recombinations on the scale below the large numerical simulation resolution. We then demonstrated how sub-grid inhomogeneity distribution in IGM alters the reionisation history and the derived topological summary statistics in simulations.

In chapter 3 we employ the notions on deep neural networks introduced in section 1.5 and apply them to reionisation in order to develop a new technique for the extrapolation of H II  regions from noisy image observations of the 21-cm signal and compare it to previous methods.

Finally, in chapter 4 we provide concluding remarks and a summary of the work presented in the thesis.

# Chapter 2

# The impact of inhomogeneous sub-grid clumping on cosmic reionisation II: modelling stochasticity

This chapter addresses the problem introduced in section 1.3.2. Here, we study the impact of unresolved sub-grid inhomogeneity in vast EoR cosmological simulation and compare various modelling approaches for the gas clumping factor and analyse the effect on the reionisation process.

The content in this chapter can be found in Bianco et al. (2021b), published on the MNRAS *as is*. Ilian T. Iliev defined the theme of the project following the results by Mao et al. (2019). The author wrote and developed the code to calculate the clumping factor field from the high-resolution numerical simulation while Iliev ran the RT simulations. The author then implemented the clumping factor models to $\texttt{C}^2\texttt{-Ray}$ code. Sambit K. Giri provided the Python package tools for the statistical analysis of the RT simulation outputs, and the author adapted them for our simulations. The author then developed the code for the analysis of the main EoR quantities. The author made all the figures and wrote the text with help from Ilian Iliev. Kyungjin Ahn, Sambit K. Giri, Yi Mao, Hyunbae Park and Paul R. Shapiro provided valuable comments on the methods and results.

## 2.1 Introduction

In simulations, the recombination rate $\mathcal{R}$ is discrete, averaged on a mesh giving $\langle\mathcal{R}\rangle = \left\langle\alpha_{\rm B}({\rm T}){\rm x}_{\rm i}^2\,{\rm n}^2\right\rangle$, where $\alpha_B(T)$ is the (temperature-dependent) Case B recombination coefficient, $x_{\rm i}$ is the ionized fraction, $n$ is the number density and for simplicity we assumed pure hydrogen gas. This indicates the number of electron-proton recombination per second in a volume, for a given gas chemistry, within each grid cell. Early semi-analytical models have adopted a common methodology named *Clumping Factor Approach*, that defines the averaged recombination rate in terms of a *clumping factor* ${\rm C} = \left\langle{\rm n}^2\right\rangle / \langle{\rm n}\rangle^2$, which corrects for the difference between the cell-averaged $\langle n\rangle^2$ and the actual value, thereby accounting for unresolved small-scale (sub-grid) structure in simulations (Gnedin & Ostriker, 1997; Tegmark et al., 1996; Ciardi & Ferrara, 1997; Madau et al., 1999; Valageas & Silk, 2004). If not correctly treated, this approach can underestimate the impact of sub-grid inhomogeneities on absorption of radiation. In some cases this term is just completely ignored, i.e.: ${\rm C} = 1$ (Onken & Miralda-Escudé, 2004; Kohler et al., 2007), but the more common and simplistic approaches consist in either a constant global term (Cen, 2003; Zhang et al., 2007) or a time evolving global term (Iliev et al., 2005; Mellema et al., 2006b; Iliev et al., 2007; Pawlik et al., 2009), averaged on the entire box volume, also referred as the *biased homogeneous* or *globally averaged clumping model*. Recently we presented our first work (see Mao et al., 2019, for reference), hereafter Paper I, where we investigated the impact of a spatially varying, local density dependent sub-grid clumping factor on reionization observables. In the present paper we extend the discussion and propose a more realistic and accurate treatment of the *Clumping Factor Approach*, that takes into account also the scatter around the mean clumping-density relation observed in high-resolution simulations.

We use a high-resolution N-body simulation of a small volume of side length 9Mpc, with spatial and mass resolution of approximately 200 pc and 5000 $\mathrm{M}_\odot$, to statistically describe IGM density fluctuations down to the Jeans mass in the cold, pre-reonization gas and then to implement these sub-grid density fluctuations into two large volume (714Mpc and 349Mpc of side length) reionization simulations. By adapting the small-scale sub-grid to the resolution of larger boxes we then model the correlation between density and clumping factor, comparing three different models (details in §2.2.3), in order to infer the clumping factor from the coarse density grid of the large volume, see §2.2.4. Finally we perform a radiative transfer simulation to study the effect of this sub-grid inhomogeneity approach on observables of reionization.

This paper is organized as follows. In § 2.2 we present the N-body and radiative

Table 2.1: N-body simulation parameters. Minimum halo mass is $10^5\,M_\odot$, $10^9\,M_\odot$ and $10^9\,M_\odot$, corresponding to 20, 40 and 25 particles, respectively in SB, LB-1 and LB-2. In all cases the force smoothing length is fixed at 1/20 of the mean inter-particle spacing.

| Label | Box size | $N_{particle}$ | fine mesh | spatial resolution | $m_{particle}$ |
|---|---|---|---|---|---|
| SB | 9Mpc | $1728^3$ | $3456^3$ | 260 pc | $5.12 \times 10^3 \mathrm{M}_\odot$ |
| LB-1 | 714Mpc | $6912^3$ | $13824^3$ | 5.17 kpc | $4.05 \times 10^7 \mathrm{M}_\odot$ |
| LB-2 | 349Mpc | $4000^3$ | $8000^3$ | 4.36 kpc | $2.43 \times 10^7 \mathrm{M}_\odot$ |

| Label | Box size | RT coarse-grained mesh[a] | RT coarse-grained cell size[b] |
|---|---|---|---|
| SB | 9Mpc | $8^3\,(53\%), 13^3\,(50\%)$ | 2.381, 1.394Mpc |
| LB-1 | 714Mpc | $300^3$ | 2.381Mpc |
| LB-2 | 349Mpc | $250^3$ | 1.394Mpc |

[a] SB density grid is coarsened to the to the required resolution for the LBs. In the column for SB, the coarsened mesh size and respective percentage of the overlapping volume for windows mesh function, calculated with equation (2.2.3).

[b] Spatial resolution of the RT coarse-grained mesh for SB, for the calculation of equation (2.2.5) and 2.2.6.

transfer (RT) simulation used, the numerical methods, §2.2.2 and our models in §2.2.3. In §2.2.4 we discuss the realisation of the clumping factor for large volumes from sub-grid inhomogeneity correlation. In §2.3 we analyse our RT simulation results and look into how our models influence the basic features of EoR: the reionization history in §2.3.1, the volume-averaged ionization fraction evolution, the integrated Thompson optical depth and then the Bubble size distribution in §2.3.3. To better understand the change in ionization morphology we describe a side-by-side comparison of box slice shot with zoom §2.3.2. In §2.3.4 we analyse the 21cm signal power spectra and the brightness temperature distribution. Our conclusions are summarized in §2.4.

## 2.2 Methodology

### 2.2.1 Numerical Simulations

We use N-body simulations to follow the evolution of cosmic structures, performed with the `CUBEP`$^3$`M` code (Harnois-Déraps et al., 2013). The code uses particle-particle on short-range and particle-mesh on long-range to calculate gravitational forces. We use set of three N-body simulations, whose parameters are summarized in table 2.1.

Our clumping factor modeling is based on small, high resolution volume box ($6.3\,h^{-1}$ Mpc= 9 Mpc, $1728^3$ particles, labelled SB in table 2.1). This has sufficient spatial and mass resolution to resolve the smallest halos that can hold cold, neutral gas. Our main larger-volume N-body simulation is referred to as LB-1 ($500\,h^{-1}$Mpc= 714 Mpc, $6912^3 \approx 330$ billion particles). A smaller simulation, LB-2, ($244\,h^{-1}$Mpc= 349 Mpc, $4000^3 = 64$ billion particles) will be used as comparison to analyse possible influence of box size and resolution in the realisation of sub-grid clumping factor and prove the stability of our method. For both of the large-volume simulations the minimum halos mass resolved is $10^9\,M_\odot$, while halos with $10^8\,M_\odot < M_{halo} < 10^9\,M_\odot$ are implemented using a sub-grid model (Ahn et al., 2015a), thereby all atomically-cooling halos (ACHs) with minimum mass $\mathrm{M_{halo}} \gtrsim 5 \times 10^8 \mathrm{M_\odot}$ are included. We are using updated N-body simulations compared to Paper I, we illustrate this further in §2.5.1.

An on-the-fly spherical overdensity halo finder (Harnois-Déraps et al., 2013; Watson et al., 2013), with overdensity parameter $\Delta = 130$, creates an halo catalogues at given redshift, that is later used as inputs for the radiative transfer simulation. The remaining particles are categorized as part of the IGM. In this work we do not include any effects from minihaloes $\mathrm{M_{halo}} < 10^8\mathrm{M_\odot}$. Even though these sources could have driven ionization in the early phase of EoR, their effect on later stage is expected to be minor because of molecular dissociation by UV background radiation from primordial luminous sources, up to a point that their contribution is negligible compared to heavier ACHs (Ahn et al., 2009). Initial conditions are generated using the Zel'dovich approximation and the power spectrum of the linear fluctuations is given by the `CAMB` code (Lewis et al., 2000). The SB N-body simulation starts at redshift z = 300, while LB-1 and LB-2 at $z = 150$, which gives enough time to significantly reduced non-linear decaying modes (Crocce et al., 2006), while at the same time fluctuations are small enough to ensure linearity of density field at the respective resolutions. The cosmological parameter are based on WMAP 5 years data observation and consistent with final Planck results, for a flat, ΛCDM cosmology with the following parameters, $\Omega_\Lambda = 0.73$, $\Omega_\mathrm{m} = 0.27$, $\Omega_\mathrm{b} = 0.044$, $\mathrm{H_0} = 70\,\mathrm{km\,s^{-1}Mpc^{-1}}$, $\sigma_8 = 0.8$, $\mathrm{n_s} = 0.96$ and the cosmic helium abundance $\eta_\mathrm{He} = 0.074$ (Komatsu et al., 2011b). Our method is general and can be applied in any cosmological background, but the specific fitting parameters we provide are based on the above values.

We simulate the Epoch of Reionization using the `C`$^2$`-Ray` code (Mellema et al., 2006b), a photon-conserving radiative transfer (RT) code based on short characteristic ray-tracing.

The LB-1 and LB-2 N-body simulations provide the IGM density fields and halo catalogues with masses, velocities, position and other variables, for a total of 76 snapshots, equally spaced in time ($\Delta t = 11.54\,\mathrm{Myr}$) in the redshfit interval z $\in$ [6; 50]. For computational feasibility, the density grid is coarsened for the radiative transfer simulation to $300^3$ (LB-1), and $250^3$ (LB-2). The high-resolution N-body simulation (SB) data input is initially interpolated onto a $1200^3$ (SB) grid, which can then be coarsened to the required resolution as discussed in the next section. These grids correspond respectively to cell sizes of length 2.381Mpc, 1.394Mpc and 7.5kpc. For brevity we will refer to these grids as the *sub-grid volumes* for SB, and *coarse volumes* in LB-1 and LB-2, noted $\langle . \rangle_{crs}$. Just as in Paper I, the interpolation of the particles onto a grid is performed with a Smoothed-Particle-Hydrodynamic-like method (SPH-like), which then yields coarse-grid density, velocity and clumping fields (see sect. 2.2 in Paper I for details).

Ionization sources for the radiative transfer simulations are characterised by the ionizing photon production rate per unit time $\dot{N}_\gamma$, given by

$$\dot{N}_\gamma = f_\gamma \frac{M_{halo}\,\Omega_b}{\Delta t_s m_p \Omega_0} \tag{2.2.1}$$

where $m_p$ is the proton mass, $M_{halo}$ is the total halo mass within coarse-grid cell, $\Delta t_s = 11.53\,\mathrm{Myr}$, the lifetime of stars set equal to the time between N-body snapshots. $\mathrm{f}_\gamma$ is the efficiency factor, defined as

$$f_\gamma = f_\star\, f_{esc}\, N_i \tag{2.2.2}$$

where $f_\star$ is the star formation efficiency, $f_{\mathrm{esc}}$ is the photons escape fraction and $N_i$ is the stars ionizing photon production efficiency per stellar atom, it depends on the initial mass function (IMF) of the stellar population, e.g. for Pop II (Salpeter IMF) $\dot{N}_\gamma \sim 4000$, the value for $f_\star$ and $f_{esc}$ are still uncertain, therefore these parameters can be tuned in order to match the observational constrain that we will discuss in §2.3. Here we adopt the partial suppression model of (Dixon et al., 2016), whereby for LMACHs located in a neutral cell the efficiency factor is set to $f_\gamma = 8.2$, while in an ionized cell (above 10%) we set $f_\gamma = 5$ to account for feedback. For HMACHs the efficiency factor has a constant value of $f_\gamma = 5$, equivalent to e.g. $\mathrm{N_i} = 5000$, $f_\star = 0.05$ and $f_{\mathrm{esc}} = 0.02$.

### 2.2.2 Coarse-Grid Method

Our clumping factor calculations are based on N-body data and neglect any hydrodynamical effects on the clumping factor. Accounting for the gas pressure provides additional smoothing at small scales and therefore our clumping factor values should be considered as upper limits. Moreover, we are interested in the reionization of the IGM and therefore exclude the halos from our calculations. The contribution of recombinations inside haloes is already taken into account in equation (2.2.1) through the photon escape fraction and should not be counted again.

In order to represent the N-body particles into a regular grid, we adopt the SPH-like smoothing technique described in §2.2 of Paper I, we refer the readers to e.g. Shapiro et al. (1996) for more general details on SPH smoothing methods. In LB-1 and LB-2 simulations we use regular meshes directly produced by the SPH code at the required resolution (the specific values used here are listed in table 2.1). In the SB simulation we adopt a more flexible approach, whereby we first produce all quantities on a very fine mesh (here $1200^3$), which is later coarsened as required in order to approximately match the cell sizes used in the LB simulations.

A window mesh function smooths the SB mesh-grid on a coarser-grained mesh, with size defined by equation (2.2.3). The method allows the windows function to overlap. The percentage of overlap $\mathrm{N}_{\%}$ is chosen in order to achieve the required resolution size of the LBs and at the same time obtain a large enough set of coarsened SB data, since $\mathrm{Mesh}^3_{\mathrm{crs-gr}}$ gives the total number of data point then interpolated by the clumping models (see figure 2.1).

$$\mathrm{Mesh}_{\mathrm{crs-gr}} = \frac{\mathrm{BoxSize}_{\mathrm{SB}}}{(1 - \mathrm{N}_{\%}) \cdot \mathrm{Res}_{\mathrm{LB}}} \tag{2.2.3}$$

where $\mathrm{Res}_{\mathrm{LB}}$ is the coarse grained resolution of the large box and $\mathrm{BoxSize}_{\mathrm{SB}}$ the box size of the small box. We employ the SB cell-wise quantities expressed with equations (2.2.4) and (2.2.7) to compute the parametrization of the correlation models. Hereafter, we will refer to them as the *sub-coarse-grid* or *SB data*, whereas in the case of LBs we name them *RT-mesh grid.* In our case we have $\mathrm{Mesh}_{\mathrm{crs-gr}} = 8$ with percentage overlap $\mathrm{N}_{\%} = 53\%$ for 714Mpc (LB-1) and $\mathrm{Mesh}_{\mathrm{crs-gr}} = 13$ with $\mathrm{N}_{\%} = 50\%$ for 349Mpc (LB-2).

We define the gas clumping factor based on the cell-wise averaged quantities (e.g. Iliev et al., 2007; Jeeson-Daniel et al., 2014; Mao et al., 2019)

$$\mathrm{C}_{\mathrm{IGM}} = \frac{\langle \mathrm{n}^2_{\mathrm{IGM,\,cell}}\rangle_{\mathrm{cell}}}{\langle \mathrm{n}_{\mathrm{IGM,\,cell}}\rangle^2_{\mathrm{cell}}} \tag{2.2.4}$$

where

$$\langle \mathrm{n}_{\mathrm{IGM,\,cell}}\rangle_{\mathrm{cell}} \equiv \frac{1}{\mathrm{V}_{\mathrm{cell}}} \int_{\mathrm{cell}} \mathrm{n}_{\mathrm{IGM,\,cell}}(\mathbf{r})\, \mathrm{d}^3\mathrm{r} \tag{2.2.5}$$

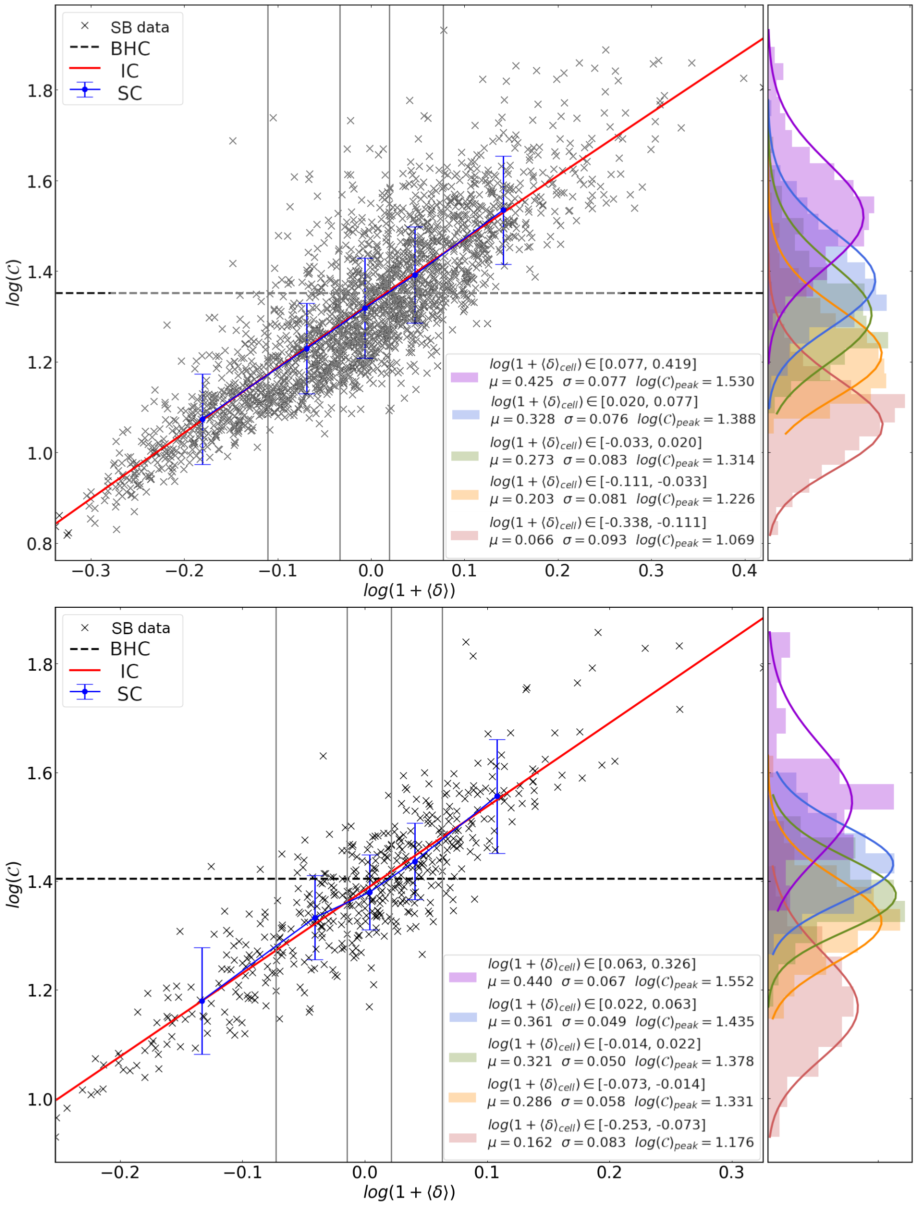

Figure 2.1: Sample correlation between local coarse IGM overdensity and coarse clumping factor at redshift z = 7.305 for LB-1 resolution (1.394 Mpc cells, top panels) and LB-2 resolution (2.391 Mpc, bottom). Shown are the coarsened SB N-body data at these resolutions (black crosses), the IC model (deterministic) fit (red line) and the globally-averaged clumping factor (horizontal dashed line). The (blue) points with error bars represent the expected value and standard deviation of the log-normal distribution (see text) in each overdensity bin. Vertical lines (solid grey) indicate the bin limits, whose sizes are adjusted so that each bin contains the same number of data points (approximately ∼ 400 (top) and 100 (bottom)). For each figure the right panel shows the log-normal distribution (solid line) of the clumping within each density bin vs. the actual data (shadow area), where we include in the legend a short description of the relevant parameters.

and

$$\langle \mathrm{n}^2_{\mathrm{IGM,\,cell}} \rangle_{\mathrm{cell}} \equiv \frac{1}{\mathrm{V}_{\mathrm{cell}}} \int_{\mathrm{cell}} \mathrm{n}^2_{\mathrm{IGM,\,cell}}(\mathbf{r})\, \mathrm{d}^3\mathrm{r}. \tag{2.2.6}$$

The mean cell over-density is defined

$$1 + \langle \delta \rangle_{\mathrm{cell}} = \frac{\langle \mathrm{n}_{\mathrm{IGM,\,cell}} \rangle_{\mathrm{cell}}}{\overline{\mathrm{n}}_{\mathrm{IGM,\,cell}}} \tag{2.2.7}$$

where $\overline{\mathrm{n}}_{\mathrm{IGM}}$ is the global average of the IGM number density over the entire box volume (in this paper, we always refer to quantities in comoving units).

### 2.2.3 Modeling the Overdensity-Clumping Correlation

In this work we consider several models for the parametrisation of the correlation between the local coarse overdensity $1 + \langle \delta \rangle_{\mathrm{cell}}$ and the coarse clumping factor $\mathrm{C}_{\mathrm{IGM}}$.

#### I) Biased Homogeneous Subgrid Clumping (BHC)

The simplest approach is to set a constant (redshift-dependent) clumping factor $C(z)$, for the entire simulation volume (e.g. Madau et al., 1999; Mellema et al., 2006b; Iliev et al., 2007; Kohler et al., 2007; Raičević & Theuns, 2011). In our case, we evaluate this globally averaged clumping for every SB simulation snapshot at the appropriate coarse resolution and then fit it with an exponential function of the form:

$$\mathrm{C}_{\mathrm{BHC}}(\mathrm{z}) \equiv \overline{\mathrm{C}}_{\mathrm{IGM}} = \mathrm{C}_0\, \mathrm{e}^{\mathrm{c}_1\,\mathrm{z} + \mathrm{c}_2\,\mathrm{z}^2} + 1 \tag{2.2.8}$$

where $\mathrm{C}_0$, $\mathrm{c}_1$ and $\mathrm{c}_2$ are the fitting parameters. We refer to this model as biased homogeneous clumping (Paper I) since that volume-averaged value is then multiplied by the local cell density to obtain the recombination rate, effectively biasing recombinations towards high-density regions.

#### II) Inhomogeneous Subgrid Clumping (IC) Model

This model, where the local gas clumping is set based on one-to-one, deterministic relation with the cell density, was first presented in Paper I. We include it here for comparison purposes. The relation of the clumping with the overdensity in equation (2.2.4) is fit by a quadratic function:

$$\log_{10}(\mathrm{C}_{\mathrm{IC}}(\mathrm{x}\,|\,\mathrm{z}_{\mathrm{i}})) \equiv \mathrm{y} = \mathrm{a}_{\mathrm{i}}\,\mathrm{x}^2 + \mathrm{b}_{\mathrm{i}}\,\mathrm{x} + \mathrm{c}_{\mathrm{i}} \tag{2.2.9}$$

where $\mathrm{x} = \log_{10}(1 + \langle \delta \rangle_{\mathrm{cell}})$ and $\mathrm{y} = \log_{10}(\mathrm{C}_{\mathrm{IGM}})$, the cellwise quantity from SB simulation. For each snapshot $z_i$ we evaluate the fitting parameters $\mathrm{a}_{\mathrm{i}}$, $\mathrm{b}_{\mathrm{i}}$ and $\mathrm{c}_{\mathrm{i}}$ using the coarse-grid field we derived in §2.2.2.

### III) Stochastic Subgrid Clumping (SC) Model

This model, first presented here, aims to account for the natural stochasticity of the relation between local clumping and overdensity, as observed in full numerical simulations. This stochasticity is due to various environmental effects beyond the dependence of the clumping on the local density, and results in a significant scatter around the mean relation used in the IC model (figure 2.1).

We model this scatter from the simple one-to-one relation by binning the SB coarse-grained clumping in several (here five) wide bins of overdensity $\Delta\delta_{\mathrm{j}}$. In each bin we fit the scatter using a log-normal distribution.

$$\mathcal{P}(\mathrm{x}\,|\,\mathrm{z_i}\,,\,\Delta\delta_{\mathrm{j}}) \equiv \frac{1}{\mathrm{x}\,\sigma_{\mathrm{ij}}\,\sqrt{2\pi}}\,\exp\left(-\frac{(\ln(\mathrm{x})-\mu_{\mathrm{ij}})^2}{2\,\sigma_{\mathrm{ij}}^2}\right) \tag{2.2.10}$$

where $\mathrm{x} = \mathrm{C_{IGM}}$. For each snapshot $\mathrm{z_i}$ and bin $\Delta\delta_{\mathrm{j}}$ we evaluate and record the parameters $\mu_{\mathrm{ij}}$ and $\sigma_{\mathrm{ij}}$.

A stochastic process is then applied to generate log-normal random values from two dimensional uniformly distributed variable $\mathrm{u_1}$, $\mathrm{u_2} \in [0,1]$, by using a modified[1] *Box-Müller transformation.*

$$\mathrm{C_{SC}}(\mathrm{z},\,\Delta\delta|\,\mu_{\mathrm{ij}},\,\sigma_{\mathrm{ij}}) = \mathrm{e}^{\mu_{\mathrm{ij}}+\sigma_{\mathrm{ij}}\cdot\sqrt{-2\ln(\mathrm{u_1})\cdot\cos(2\pi\,\mathrm{u_2})}} \tag{2.2.11}$$

where $\mu_{\mathrm{ij}}$ and $\sigma_{\mathrm{ij}}$ are the weighed log-normal parameters for LB-1 and LB-2. Finally, we note that the range of overdensities in the SB simulation is inevitably narrower due to the smaller volume compared to our target reionization volumes. For data beyond the SB limits, for high and low densities, we fix the mean value to the one given by the IC model, while standard deviation is fixed to the one obtained in the closest density bin. These distributions are then sampled randomly to create realisations of the clumping in large-volume simulations. A similar approach, but in a different context, has been used previously by Tomassetti et al. (2014) and Lupi et al. (2018), motivated by observation of density distribution in giant molecular clouds.

In figure 2.1 we show examples of the resulting parametrization obtained from the three models at redshift $\mathrm{z} = 7.305$, applied at the LB-1 and LB-2 RT resolutions. We show the coarse-grained N-body data, along with the BHC and IC models, as well as the mean, $\mathbb{E}[\mathrm{X}] = \mathrm{e}^{\mu+\frac{1}{2}\sigma^2}$, and the standard deviation, $\mathrm{SD}[\mathrm{X}] = \mathrm{e}^{\mu+\frac{1}{2}\sigma^2}\sqrt{\mathrm{e}^{\sigma^2}-1}$, of our proposed log-normal distribution of the stochasticity. On the side plot we show the coarse-data

[1] A random variable is defined log-normal distributed when the natural logarithm of the variable is normal distributed. Therefore our modification simply consist in taking the exponential of the transformation.

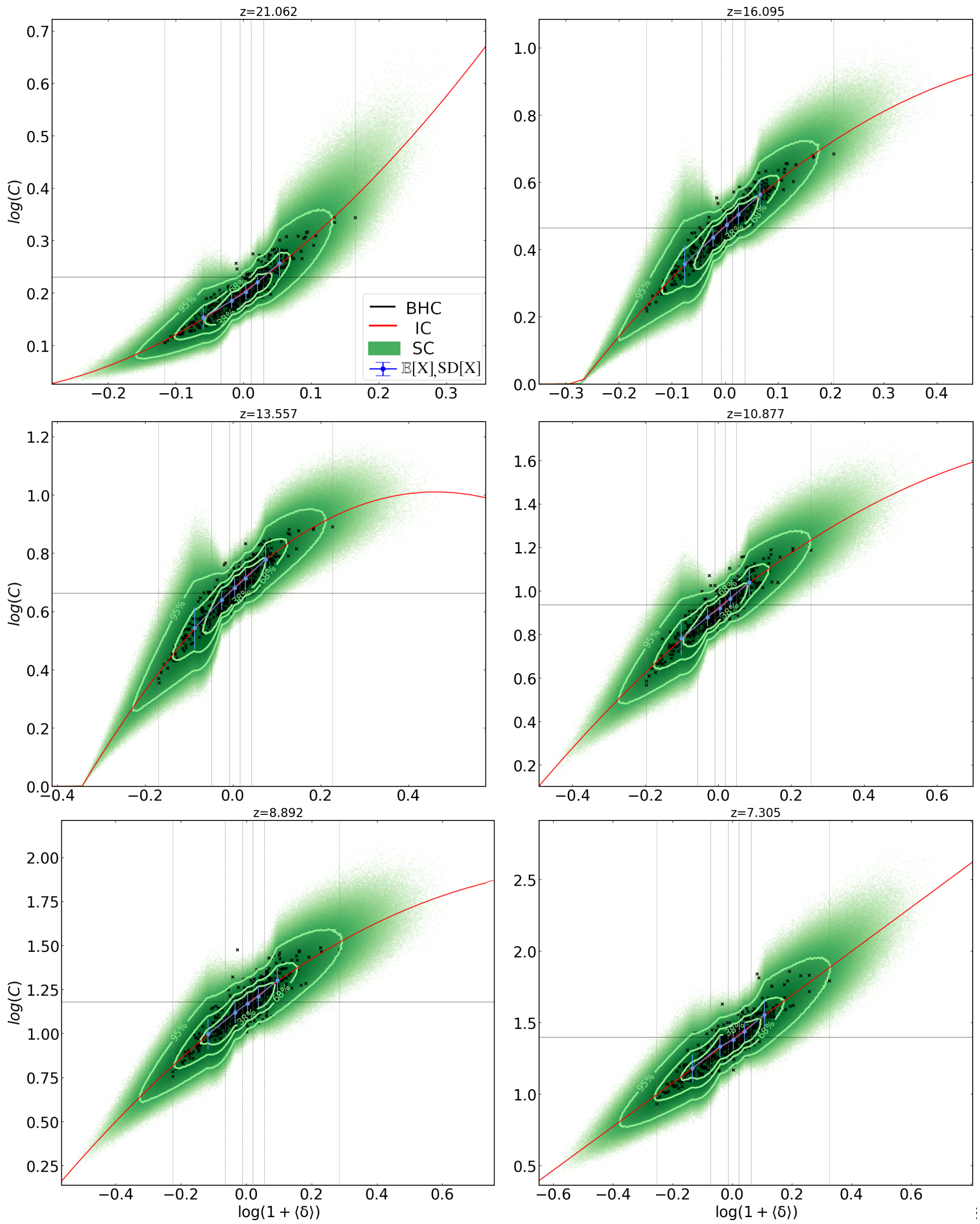

Figure 2.2: Realization of clumping factor for LB-1 at different redshift. The horizontal line (solid black) is the globally averaged clumping factor BHC. In red the one-to-one fit IC. Blue error barred points represent the expected value and standard deviation of the log-normal distribution. Vertical lines (grey dashed) indicate the bin limits. The green area indicates the SC realization estimated by equation (2.2.11). We plot the 38% ($0.5\,\sigma$), 68% ($1\,\sigma$) and 95% ($2\,\sigma$) confidence interval to highlight the realization distribution. Cross point are the *coarse* SB data used to calibrate the model parameters. In the case of $z = 7.305$ they correspond to the one of figure 2.1 (left column).

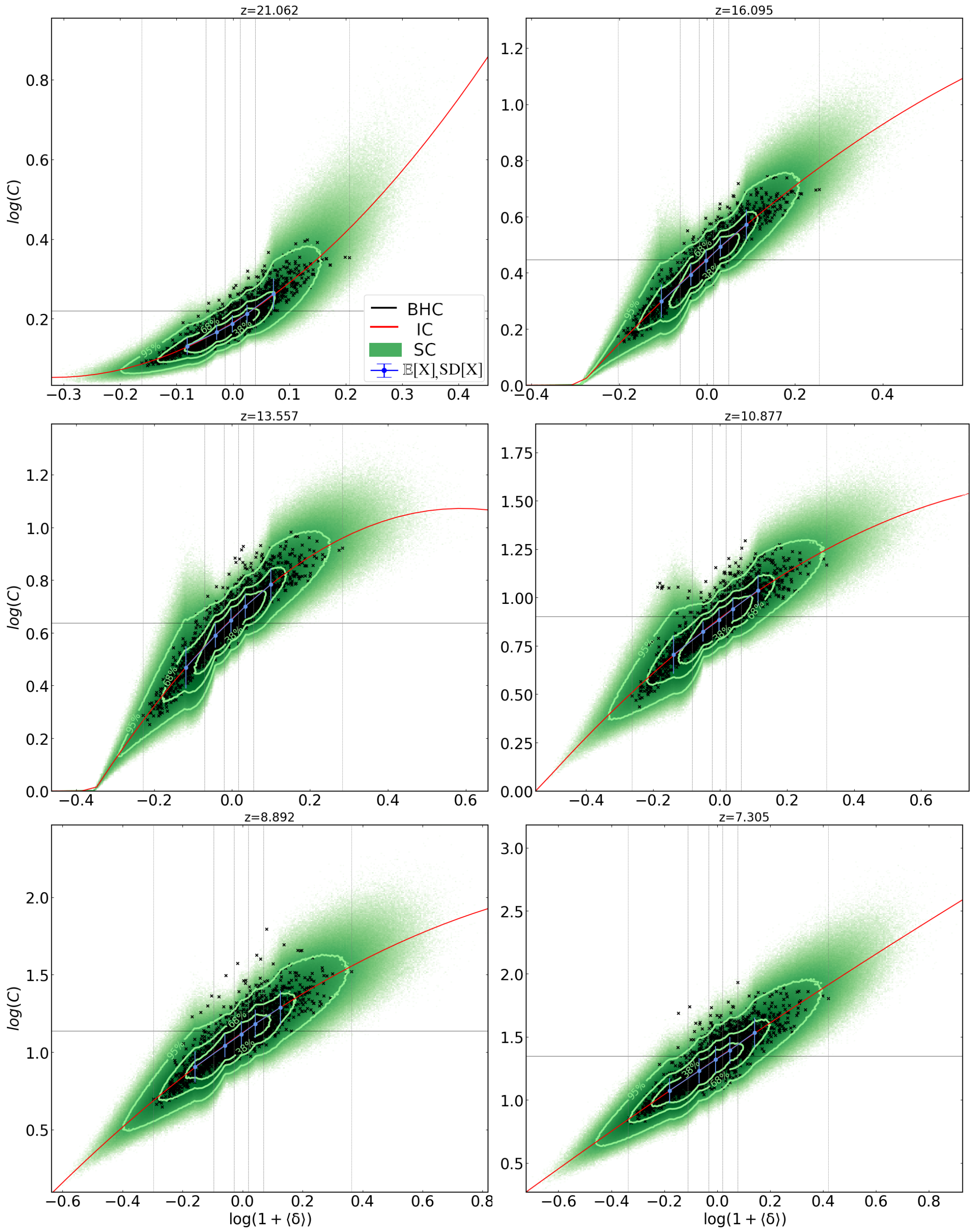

Figure 2.3: Equivalent explanation given for figure 2.2 but for LB-2. Cross point are the *coarse*-data used to calibrate the model parameters. In the case of z = 7.305 they correspond to the one of figure 2.1 (Bottom panel).

distribution (shadow histogram) and the resulting log-normal fit (solid line) with brief description of the density-bin limits and fitting parameters shown in the legend.

### 2.2.4 Clumping Implementation in Large-Scale Volumes

We used simulations LB-1 and LB-2 as examples of our method for creating large-volume simulation sub-grid clumping realisation. Results are shown in figure 2.2 and figure 2.3, respectively. In the figures we show the N-body data upon which the model is based (black crosses), the volume-averaged clumping factor BHC (black horizontal line), the one-to-one quadratic fit IC (red solid line), the expectation value $\mathbb{E}[X]$ and the standard deviation SD[X] of the log-normal distribution in each density bin (blue error-bar points) with the relative bins limits also shown (dashed vertical line). Finally, our SC model clumping realisation (green area) based on the density field of LB-2 is shown with contours corresponding to the 95% (outer), 68% (middle) and 38% (inner) confidence interval. Tables with parameters of the three models used in this paper can be found online[2].

The results illustrate the extend to which each subgrid clumping model reproduces the trends in the direct N-body data throughout the evolution. The BHC (mean-clumping) model roughly matches the peak of the contours and its evolution over time. The IC model (quadractic fit) captures well the general trend of the density-clumping relation and tracks well the highest density of data points. Finally, our new SC model realisation fully reproduces the data, including the scatter around the mean relation. The contours trace the majority of the simulation data quite closely, apart from a few outliers. However, a few things should be noted here.

First, as noted in § 2.2.3, the large volumes generally sample much wider range of environments than smaller ones used to produce the model, thus inevitably the large-volume realisation should extrapolate to over-densities outside the range sampled by the direct N-body data, for both larger and smaller over-densities. Second, again as discussed above, for statistical reasons we fixed the bin sizes so that they contain same number of data points, which inevitably results in quite uneven bin widths. These are very narrow near the peak density of points and are quite wide for extreme values of the over-density. The combination of these factors yields the 'flairing' of the realisaion (green) points at both large and small values of the over-density, and thus possible minor discrepancies with what we find in simulations. However, only small fraction of the points are in these regions, as demonstrated by the density of points, and therefore it is unlikely this will

[2] table for model parameters:
https://users.sussex.ac.uk/~mb756/fig/project1/tables.pdf

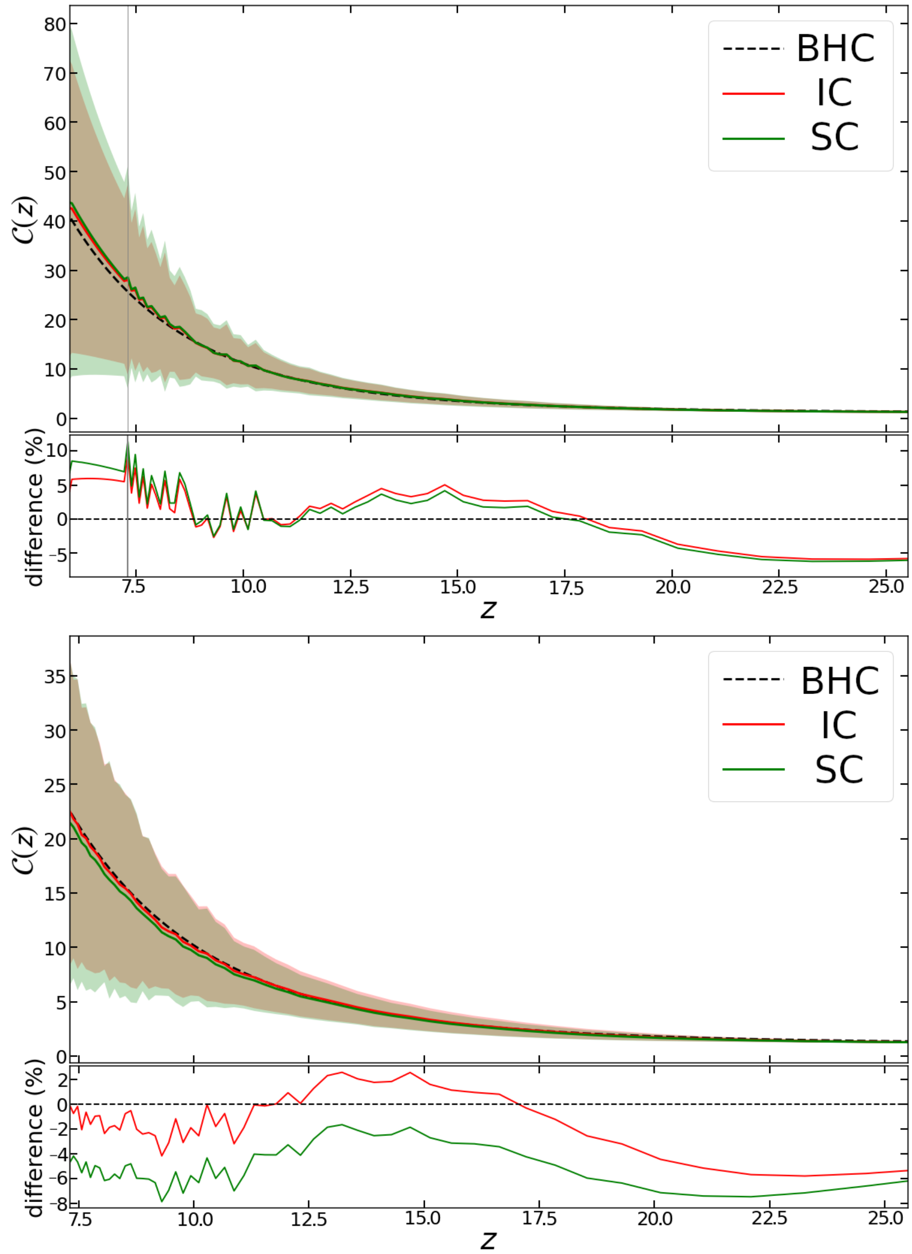

Figure 2.4: Comparison of mean clumping values for the three different models,the redshift evolution of the mean clumping factor for the three models, respectively, shows the range of the standard deviation. On the bottom plot we show the relative where the left image is for LB-1 and right for LB-2. On the upper plot we have the dashed black line is BHC, in red IC and in green SC model, the shadow error in percentage of the difference with BHC.

affect the results in any significant way. Obviously, the IC model is potentially affected in a similar way, since the quadratic fit is used beyond the range of the original data points.

As a consistency check, we compare the redshift evolution of the volume average of the clumping realizations based on the SC and IC models vs. the actual global mean $\mathcal{C}_{\rm glob}$ based on the simulated data (figure 2.4). The vertical line indicates the redshift at which the SB simulation was stopped, thus data beyond that is extrapolated. The relative errors of the mean values (bottom panels) are in agreement within the $6-7\%$ for LB-2 (right) and within 10% for LB-1 (left), throughout the relevant redshift range $6 \leq \mathrm{z} \leq 30$. At the highest redshifts ($z > 30$) the errors appear larger, however over that redshift range the density fluctuations are small and thus all clumping factors converge to 1 and do not contribute to the recombination rate.

Hence, the local density inhomogeneity does not significantly affect the global averages; however, we expect that the local clumping factor plays a greater role in the recombination and ionization at small scales (e.g. on the H II region size distribution, *ionized bubble* volume evolution, etc.). The proposed models are roughly consistent with results of previous papers Park et al. (2016), Iliev et al. (2007) and once our the RT-simulation are performed we expect to obtain similar confirmation from the work of Iliev et al. (2012).

## 2.3 Clumping Model Effect on Observational Signature

The sub-grid clumping model employed affects the local IGM recombination rates, which is then reflected in the derived observable signatures of reionization. In order to understand and try to quantify the importance of this choice, we perform three RT simulations where we fix the source production efficiencies of ionizing photons and vary solely the clumping model. At each time step the precomputed N-body density fields are used to create a realisation of the corresponding gridded clumping factor, as described in §2.2.3. These clumping grids are then stored and provided as additional inputs to full radiative transfer simulations with the C$^2$-RAY code (Mellema et al., 2006a). Specifically, the simulation used for this section is LB-1.

The simulation redshifts span the range $z = 40$ to 6, for a total of 125 snapshots. The corresponding aperture on the sky vary from 3.6 to 4.7 deg per side, and covers the redshifted 21-cm frequency range from 34 to 202 MHz. The resolution evolves from 43.5 arcmin to 56 arcmin in the spatial direction, and from 0.08 to 0.15 MHz in frequency.

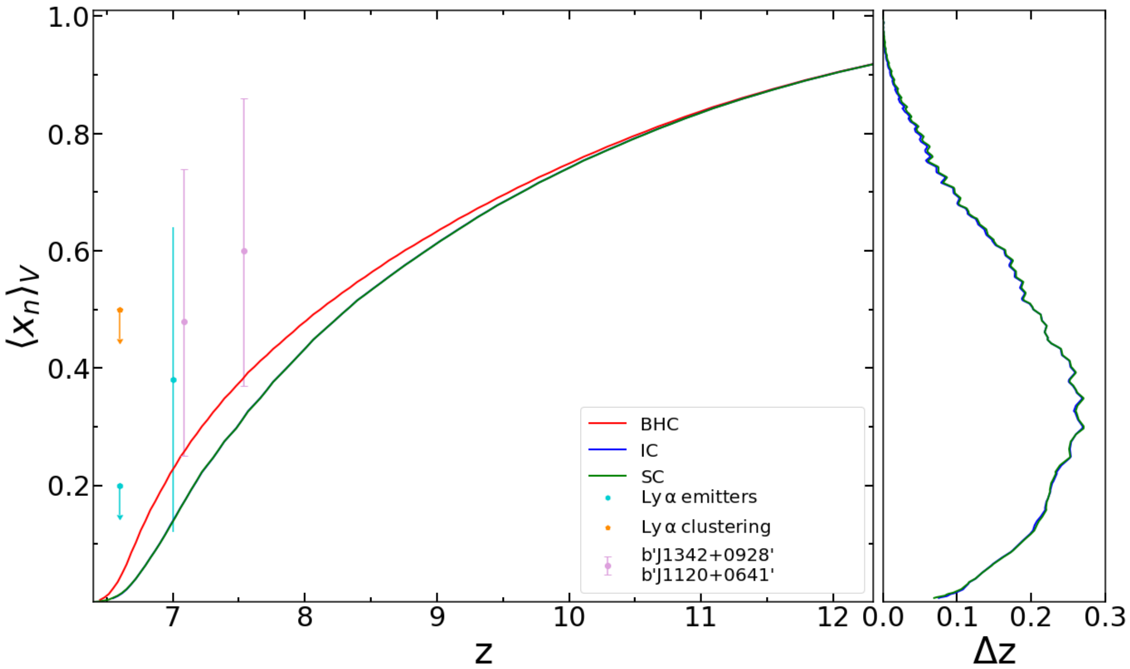

Figure 2.5: Left plot, the volume-averaged neutral fraction for BHC (solid red), IC (dashed blue) and SC (solid green) clumping models applied to simulation SB-2. On the right we show the redshift delay of IC and SC models compared to BHC. As a comparison we include observational constrains (see legend) from Ly$\alpha$ emitters (cyan circle) (Ota et al., 2008; Ouchi et al., 2010), Ly$\alpha$ clustering (orange circle) (Ouchi et al., 2010) and from high redshift quasars spectra (pink) (Davies et al., 2018).

### 2.3.1 Reionization History

Our results on the reionization histories are presented in figure 2.5 and table 2.2. Perhaps counter-intuitively, either of the more realistic, density-dependent clumping treatments (SC and IC) yield somewhat faster evolution and an earlier end of reionization compared to the BHC model. The former models diverge from BHC around $z \leqslant 12$, and thereafter the mean reionization is accelerated with a maximum difference at $\bar{x}_i = 70\%$, of $\Delta z \simeq 0.3$ at $z \simeq 7.5$, corresponding to a time difference of approximately 36 Myr. The end of reionization is delayed by $\Delta z = 0.1$, or 17 Myr. Here there is very little difference between the SC and IC models. Compared to the observational constraints, all three models reionize somewhat early, however these constraints are largely upper limits, and with significant uncertainties. Moreover, our main interest is the relative effect of different sub-grid clumping models, rather than a faithful reproduction of the constraints.

During reionization, free electrons scatter CMB photons via inverse Compton scattering, suppressing CMB anisotropies on all scales and introducing polarization on large

Table 2.2: Mean volume-averaged ionized fractions, $\bar{x}_i$, at a reionization milestones: 10%, 30%, 50%, 70% and 90% volume of the gas ionized. The last column $z_{reion}$ lists the end of reionization, defined as $\bar{x}_i = 99\%$. The second section lists the redshift and time differences with respect to the BSC model.

| Model | $z_{10\%}$ | $z_{30\%}$ | $z_{50\%}$ | $z_{70\%}$ | $z_{90\%}$ | $z_{reion}$ |
|---|---|---|---|---|---|---|
| BHC | 11.918 | 9.533 | 8.118 | 7.221 | 6.721 | 6.483 |
| IC | 11.918 | 9.611 | 8.340 | 7.480 | 6.905 | 6.583 |
| SC | 11.918 | 9.611 | 8.340 | 7.480 | 6.905 | 6.549 |
| $\Delta z$ | 0 | 0.078 | 0.222 | 0.259 | 0.184 | 0.1 |
| $\Delta t\,[Myr]$ | 0 | 5.8 | 22.9 | 34.4 | 28.9 | 17.2 |

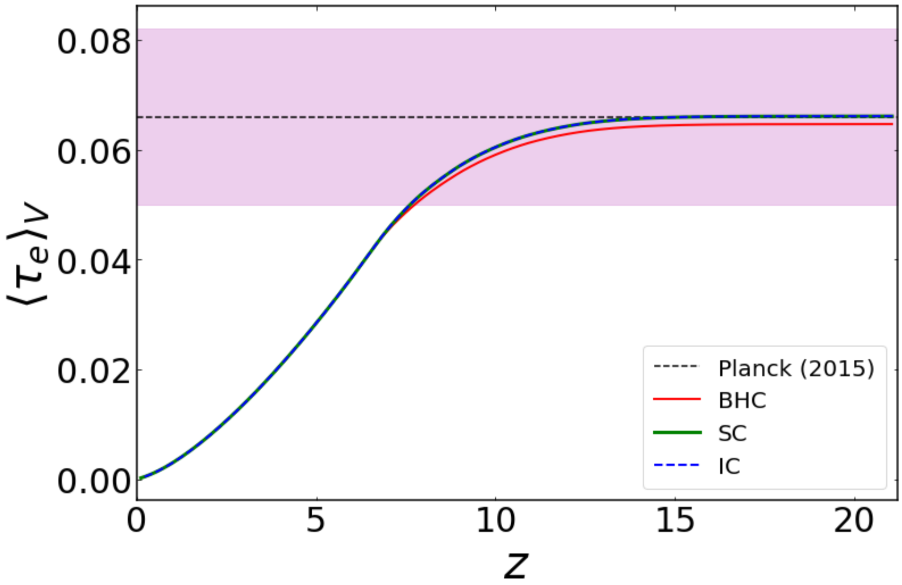

Figure 2.6: Thomson scattering optical depth to CMB photons integrated through our simulations, as labelled. Hown are also the *Planck* observational constraint (black dashed line) along with its relative 1-$\sigma$ confidence interval (violet shaded)

angular sizes. The contribution from free electron can be quantified by the integrated Thomson scattering optical depth along the line of sight, given by

$$\tau_e(z) = c\,\sigma_T \int_0^z \frac{n_e(z')}{(1+z')\,H(z')} dz' \tag{2.3.1}$$

where $\sigma_T = 6.65 \times 10^{-25} cm^2$ is the Thomson cross section, $c$ the speed of light and $n_e$ is the electron density at a given redshift.

In figure 2.6 we plot the volume mean of equation (2.3.1), integrated back in redshift. In agreement with the global reionization histories, the inhomogeneity-dependent models are very similar to each other and are slightly optically thicker than the BHC case, due to the more advanced reionization in the latter. Regardless of this small difference, all three cases are in close agreement with the *Planck-LFI* 2015 results Ade et al. (2014), which found $\tau_e = 0.066 \pm 0.016$ corresponding to an instantaneous reionization for redshift $z_{reion} = 8.8^{+1.7}_{-1.4}$.

The importance of recombinations throughout reionization could be quantified by the (dimensionless) mean rate of recombinations per hydrogen atom per Hubble time:

$$\left\langle \dot{\mathrm{N}}_{\mathrm{rec}} \right\rangle = \frac{\langle \mathcal{R} \rangle}{\mathrm{t_H(z)}\ \langle \mathrm{n_H} \rangle\ \mathrm{V_{cell}}} = 0.72\, \frac{\alpha_{\mathrm{B}}\, (1+\mathrm{z})^3}{\mathrm{H(z)}}\, \frac{\bar{\rho}_{\mathrm{c,0}}\, \Omega_{\mathrm{b}}}{\mu_{\mathrm{H}}\, \mathrm{m_p}}\, \mathrm{C_{HII}}\, \langle \mathrm{x_{HII}} \rangle^2 \tag{2.3.2}$$

In figure 2.7 (top left) we show the evolution of the mean of this quantity over the full simulation volume (solid lines), as well as averaged only over the over-dense (dashed lines) and under-dense (dot-dashed lines) regions. Colours indicate the model used, as per legend. We also show (bottom left) the relative percentage difference compared to the BHC model. As could be expected, the number of recombinations grows strongly over time, starting close to zero, departing from BHC model around $z \sim 12$, and then all reaching $\sim 15$ at late times, as more and more structures form over time. Although all models end up at similar values by $z \sim 6$, the BHC model lags behind throughout the evolution. The IC and SC models yield very similar values at all times. The over-/under-dense volumes yield much higher/lower number of recombinations, respectively, demonstrating the wide variety of outcomes dependent on the local conditions. Interestingly, the over-dense average for the BHC model results in very similar recombinations to the full-volume averages of SC and IC models, showing that at least on average the clumping in these last models behaves the same way as the over-dense regions in BHC. Overall, the SC model shows a few percent higher recombination rate ($\sim 1-5\%$) compared to the IC model. This is most likely due to the stochastic nature of the realization process, also related to the broader scatter in figure 2.4 (shaded areas).

In figure 2.7 (right panels) we compare the (non-equilibrium) photoionization rates $\Gamma_i$ computed during the run by the `C`$^2$`-Ray code`. Just as above, all mean photoionization rates are essentially the same until $z \sim 12$, after which the BHC model one rises more slowly, lagging behind the other two cases by about factor of 2.5 throughout most of the evolution, eventially catching up by $z \sim 6$. The average rates in the overdense regions are higher than the mean (reflecting the inside-out nature of reionization) by a similar amount, while the mean photoionization rates in the under-dense regions lag behind by larger factors, up to several hundred, before again rising steeply and catching up with the mean by $z \sim 6$. Interestingly, the mean rate in BHC overdense regions is again very close to the whole volume means of IC and SC models. The average values in the under-dense

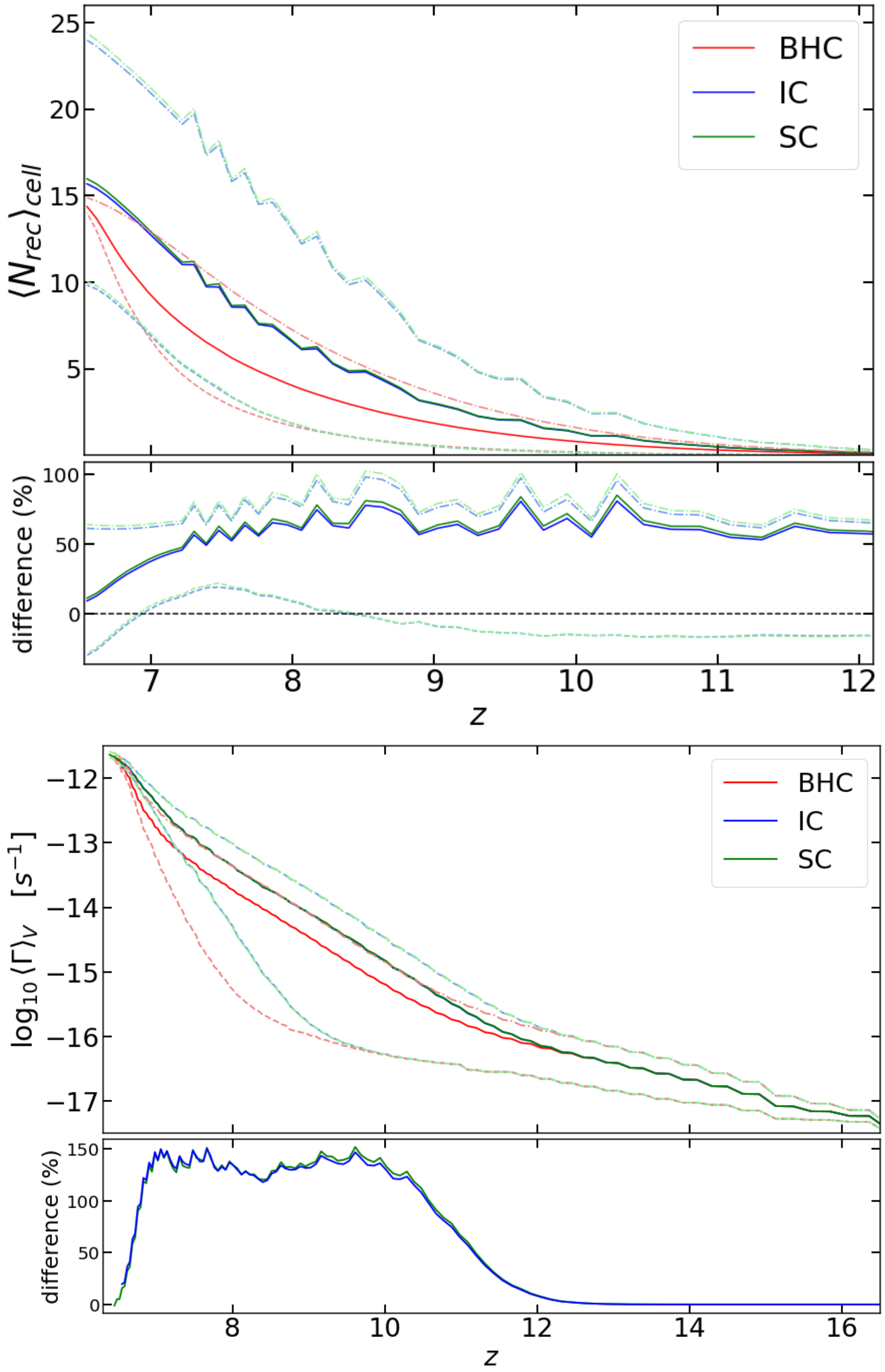

Figure 2.7: Evolution of the number of recombinations per hydrogen atom and per Hubble time throughout reionization (top) and the volume-averaged photoionization rate (bottom). The bottom plots show the relative difference compared to BHC in each case. Dashed and respectively dashed-dotted lines of the same colour indicate the relative quantity in under-dense and over-dense regions.

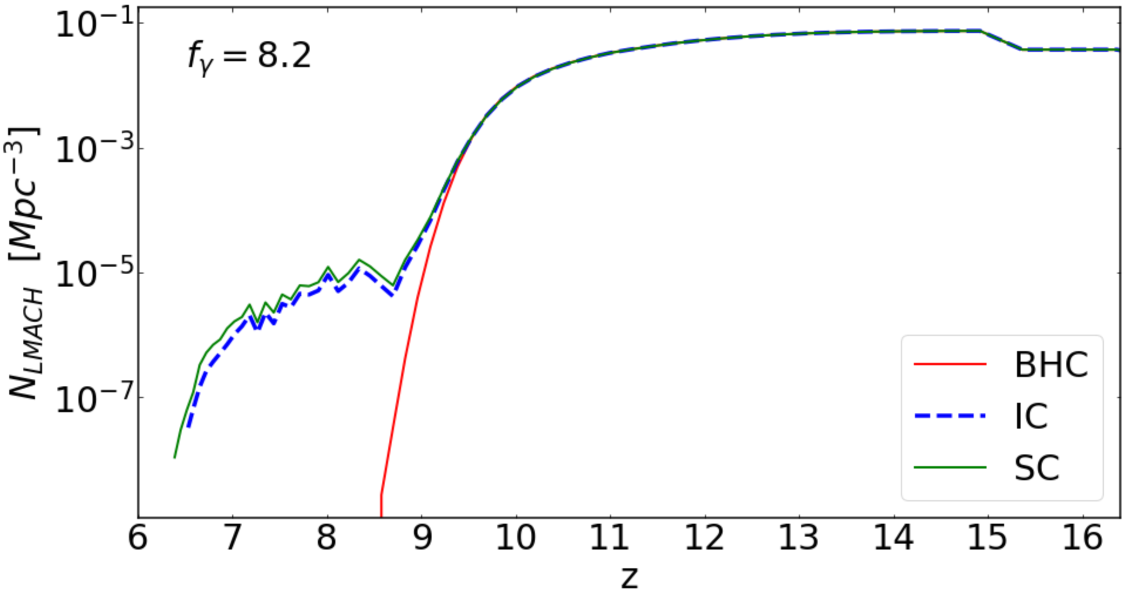

Figure 2.8: The number density evolution of unsuppressed LMACHs ($f_\gamma = 8.2$). Solid red line the BHC model, dashed blue line the IC model and in gree the SC model.

regions remain the same for all models until much later, $z \sim 9.5$, indicating that the specific clumping model has little influence before that redshift. At first glance, it seems somewhat counter-intuitive that reionization proceeds faster in the denity-dependent models IC and SC, despite their notably higher recombination rates. The reason for this is that in the former cases also the suppression of low-mass galaxies (LMACHs) due to radiative feedback is weaker than in the BHC case, as illustrated in figure 2.8. In the BHC case essentially all such galaxies are suppressed by $z \sim 8.5$, while in the density-dependent models the suppression is slowed down, allowing LMACHs to last longer in high density regions. This is further clarified in figure 2.9 where we show the number density distribution of ionized fraction of cells at five different reionization stages, $\bar{x}_i = 0.1, 0.3, 0.5, 0.7, 0.9$, approximately corresponding to redshift between $z \simeq 12 - 6$ (see table 2.2). The vertical line indicates the partial suppression threshold for LMACHs. Early on ($\bar{x}_i = 0.1$) the gas clumping has yet had very litte effect, due to the still small ionized fraction and the short time available for recombinations, thus all models yield very similar results, with only BHC showing slightly fewer highly ionized cells. As reionization progresses ($\bar{x}_i = 0.3$), IC and SC models remain very similar, while BHC is gaining more ionized cells, and at the same time it is starting to show a lack of neutral regions. Starting from roughly mid-point of reionization ($\bar{x}_i = 0.5$), the dearth of neutral cells becomes ever more prominent whereas the peak of highly ionized cells stays roughly similar for all models. A faint difference between SC and IC is visible at late times, where slightly more cells remain neutral in SC.

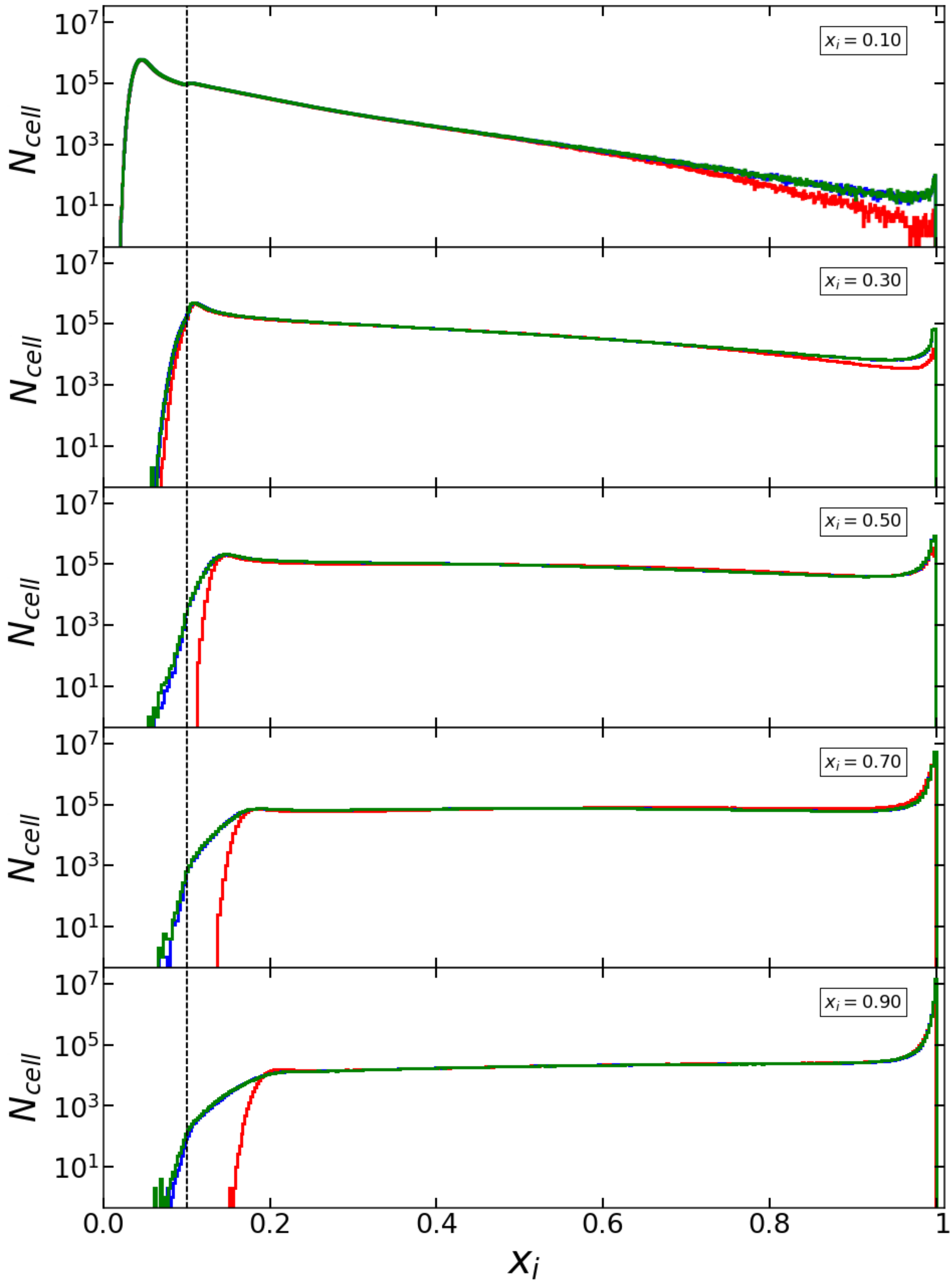

Figure 2.9: Ionized cell number density at reionization milestone $\bar{x}_i = 0.1$, 0.3, 0.5, 0.7, 0.9, top to bottom, for model BHC (red), IC (blue) and SC (green). Vertical line (black dashed) indicates the ionization threshold, $x_i = 0.1$, for partial suppression of LMACHs.

### 2.3.2 Reionization Morphology

The globally-averaged quantities discussed above (figure 2.5, 2.6 and 2.7) give an overall idea of the reionization history. Next step is to understand how the sub-grid gas clumping model affects the propagation of radiation and the local features of reionization. In figure 2.10 we show box slice of LB-2 and compare simulation snapshots with similar globally averaged ionized fraction and the three gas clumping models. From top to bottom row we have $\bar{x}_i = 0.3$, 0.5, 0.7, 0.9 (in table 2.2 we list the corresponding redshift at which this occurs and its consequent time delay compared to the BHC model) and from left to right column we have the different models BHC, IC and SC. Red/crimson regions indicate highly ionized cells $x_\mathrm{i} > 0.9$, in dark blue neutral regions $x_\mathrm{i} < 0.1$, and in green/aquamarine the transition phase $x_\mathrm{i} \approx 0.5$. Within each image we embed a zoom-in region, of 85Mpc per side, to better appreciate the morphological changes of a randomly selected under-dense neutral clump, as ionized fronts expand (bluer blob, right column plots).

Our simulations reproduce the general reionization features found in previous simulations (e.g. Iliev et al., 2014). In high density regions LMACH are the first halos to form. In our simulations they make their first appearance at redshift $z = 21$, and by $z \sim 12$ every volume element contains at least one ionizing source. At first, a modest number of isolated sources, highly clustered on small scale but homogeneously distributed on large scale, start to ionize their surrounding gas, forming small regions of a few Mpc size. The presence of sub-grid gas clumping slows down the propagation of the I-fronts and yields somewhat smaller, more fragmented H II regions. Throughout reionization, these HII bubbles grow and eventually overlap, at which point the ionization process accelerates and many of the smaller bubbles percolate to much larger connected volumes.

The side-by-side comparison shows some notable differences between BHC and the two density-dependent models, with the latter starting at a faster pace, with earlier local percolation, then slowing down compared to the former case. Modest differences appear between the three models in terms of large scale morphology, with a higher degree of ioniziation around early sources in the density-dependent models IC and SC (respectively central and right panel). From around the mid-point of reionization (50% ionization by volume, second row of images) we can see neighbouring growing regions connecting to each other and starting to highly ionize the linking filament. At this point, accordingly to figure 2.9, all cells in BHC have surpassed the threshold limit $x_\mathrm{i} = 0.1$ for the partial suppression of low mass haloes. For IC (middle) and SC (right column) the degree of ionization around sources is visibly more intense compared to BHC, in fact we can dis-

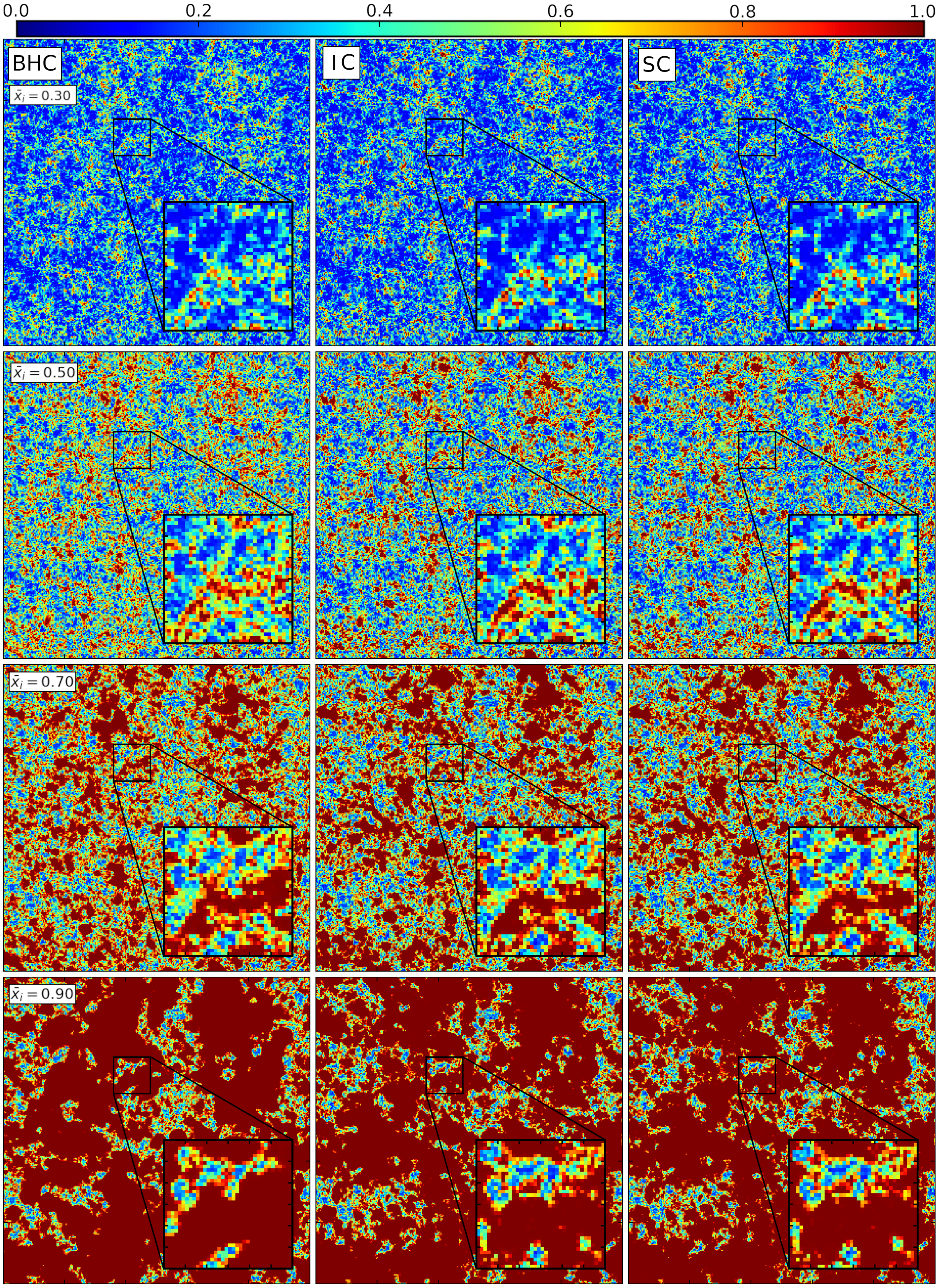

Figure 2.10: Box slice comparison of LB-2 ionization fraction for different clumping models. In red/crimson highly ionized regions $x_{\rm i} > 0.9$, in green/aquamarine transition $x_{\rm i} \approx 0.5$ and in dark blue neutral regions $rmx_{\rm i} < 0.1$. The zoom-in covers an area of 85Mpc per side and each pixel represent a volume element of 2.381Mpc per side. We compare slices at same global average ionization fraction, from top to bottom row we have $\bar{x}_{\rm i} = 0.3,\ 0.5,\ 0.7,\ 0.9$ (see table 2.2 for corresponding redshifts). From left to right column respectively we show the models BHC, IC and SC.

tinguish highly ionized cells clustered around the high density peak, whereas under-dense regions are kept fairly neutral. This diversity is due to the higher recombination rate in inhomogeneity dependent model, shown in figure 2.7 (left), that effectively reduces the number of photons able to escape the cells of origin and spread into the neighbour grid elements. This is not the case for BHC, to which clumping factor in high density regions is underestimated and ionizing photons are free to percolate and been absorbed elsewhere in the surrounding IGM, therefore interconnecting filament cells between sources clearly appears extended and in a more advanced neutral-ionized transition (blue/aquamarine).

Later on ($x_\mathrm{i} = 0.7$, third row of images), ionized regions have grown substantially and become strongly ionized. A first look suggests similar structure patchiness on large scale, although from the zoom-in we can observe that BHC has a wider and smoother transition between the ionized/neutral phases, whereas IC and SC show a narrower front, allowing more cells that host under-density to stay neutral. When the same transition region dwell across the three model, density dependent model show more irregularity with occasionally one or few cells appearing slightly more ionized then their surrounding.

The morphology differences are more evident at late times ($x_\mathrm{i} = 90\%$, bottom row of images), whereby HII bubbles connect together to form one vast interconnected highly ionized region. At this stage the vast ionized IGM in IC and SC show variations that follow the higher recombinations due to density fluctuations, which is not the case in BHC model and therefore the same regions appear uniformly highly ionized, $x \approx 1$. On the other hand there are no striking difference between IC and SC, except for small variations, of a few pixels of size, on the ionized/neutral boundaries. We suspect that this is numeric artefact due to the stochastic nature of SC. We are developing a more complete clumping model, that we will present in future work, to exclude this uncertainty.

### 2.3.3 Bubble Size Distribution

One of the key characteristics of reionization, which directly affects all observables is the normalized distribution of bubble sizes $R\,dN/dR$ or volume sizes $V^2\,dN/dV$ of ionized regions (Furlanetto et al., 2004). A number of complementary approaches to calculate these distributions have been proposed (e.g. Friedrich et al., 2011; Lin et al., 2016a; Giri et al., 2018a). Here we employ the Mean-Free-Path (MFP) method to calculate $R\,dN/dR$, and the Friends-of-Friends (FOF) algorithm (Iliev et al., 2006) to obtain $V^2\,dN/dV$ bubble size distributions (BSD). For both methods we employ the `TOOLS21CM`[3] python package

[3] https://github.com/sambit-giri/tools21cm

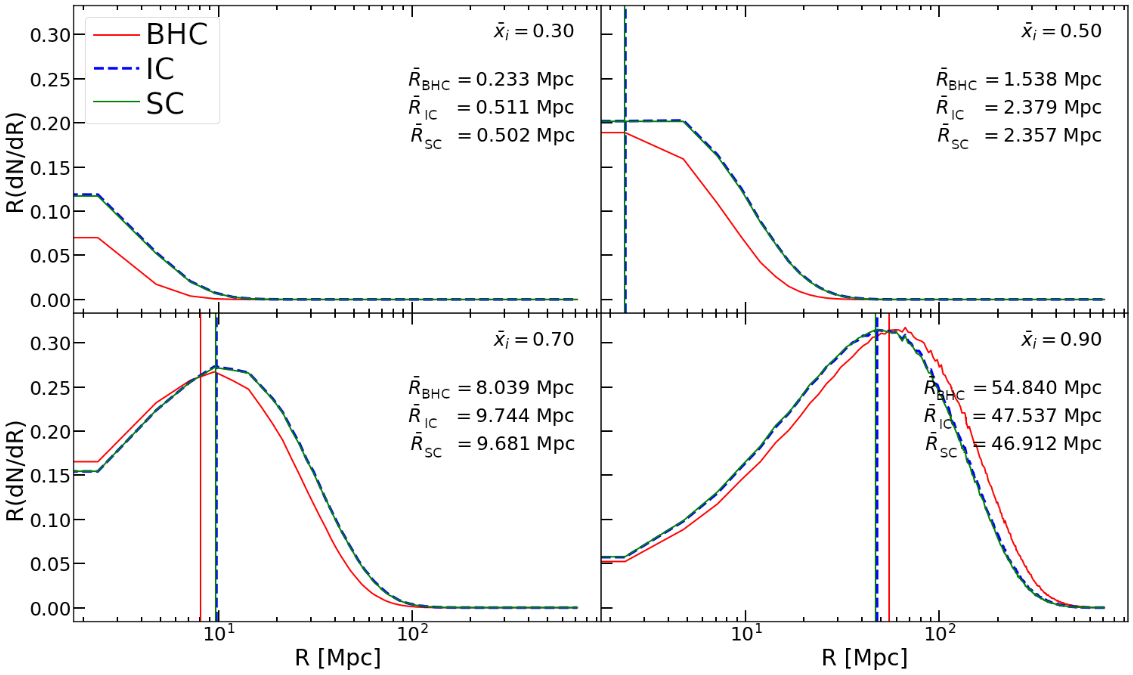

Figure 2.11: Ionized bubble size distribution for simulation LB-2 and the three gas clumping models BHC (red, solid), IC (blue, dashed) and SC (green, solid) at volume averaged ionized fractions $\bar{x}_\mathrm{i} = 0.3, 0.5, 0.7, 0.9$, as labelled. Vertical lines indicate the mean bubble radius $\bar{\mathrm{R}} = \int(\mathrm{R}\,\mathrm{dN/dR})\mathrm{dR}$ for the respective models.

for EoR simulations analysis (Giri et al., 2020). In both cases, we apply a threshold value of $x_{th} = 0.9$, since we want to highlight differences in distribution of highly ionized regions that develop around sources.

Results are shown in figure 2.11 and figure 2.12, respectively we see the typical traits of the percolation process, with volume ranges that roughly corresponding to what is expected from large simulated box (Iliev et al., 2014). We present our results at four different reionization milestones, $\bar{x}_i = 0.3, 0.5, 0.7, 0.9$, see table 2.2 for corresponding redshifts. In the case of MFP-BSD, we calculate the mean bubble size by $\bar{R} = \int(R\,dN/dR)\,dR$, represented by the corresponding vertical lines for each simulation. The sharp cut-off at small scales 2.381Mpc, for MFP-BSD, and 13.498Mpc$^3$ for FOF-BSD correspond to the simulation cells size and volume respectively. Early-on ($\bar{\mathrm{x}}_\mathrm{i} = 0.3$, top left panel, figure 2.11), LB-2 hosts small H II bubbles with radius smaller then 10Mpc. For inhomogeneity-dependent models IC and SC, distributions present many more highly-ionized regions, indication of a faster radiation propagation around sources. All three distributions peak at the size corresponding to one cell. The same trend is confirmed by the topologically-connected FOF volumes (figure 2.12), which are however typically larger than MFP, with volumes

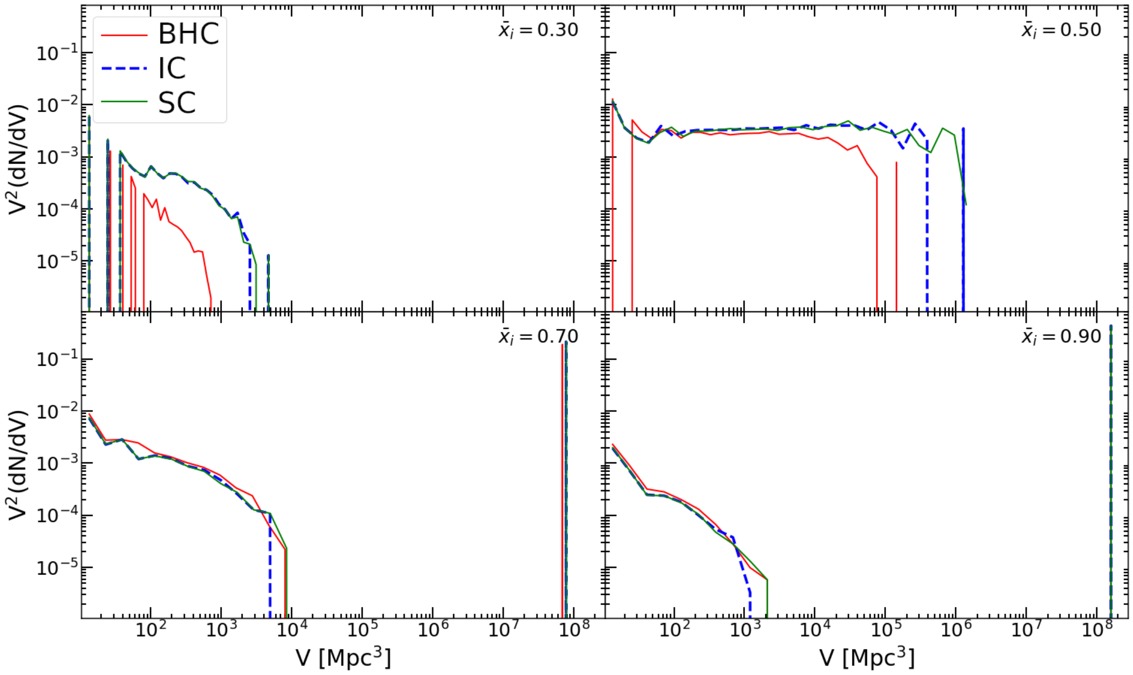

Figure 2.12: Ionized volume size distribution for LB-2. In red the result for BHC, in blue for IC and in green for SC. Distribution represents stages where the volume averaged ionized fraction is $\bar{x}_i = 0.3, 0.5, 0.7, 0.9$.

between $30 - 700\,\mathrm{Mpc}^3$ for BHC and a wider distribution for IC and SC, from one cell up to a few thousand Mpc$^3$.

Even though the number of bubbles increase as reionization progress, at $\bar{\mathrm{x}}_\mathrm{i} = 0.5$ (top right), the MFP-BSD remain similar. However, the FOF-BSD shows a qualitative transition when the small H II regions start to percolate into much larger, connected one. Their sizes vary widely, with a broad flat distribution (plateau) at smaller scales (V $<$ $10^5 - 10^6$Mpc$^3$). However, BHC and IC also show a bifurcated distribution, with a second peak at large scales, at $10^5$Mpc$^3$ for BHC and $10^6$Mpc$^3$ for IC, indicating that percolation process has started (Iliev et al., 2006, 2014; Furlanetto & Oh, 2016a). Compared to BHC, the IC distribution is shifted toward larger sizes, such that the limit for the plateau and the percolation cluster are up to one order of magnitude higher. A narrower separation between these two volume range indicates that the merging of ionized region in BHC has just started (Iliev et al., 2014; Furlanetto & Oh, 2016a; Giri et al., 2018a), whereas in the case of IC this process is already ongoing. On the other hand, IC and SC distribution show similarity at small scale but they differ for larger volumes. The former distribution shows a constant and continuous range of scales from large volumes V $\sim 10^6$Mpc$^3$ down to one cell, sign that ionized regions are in principle less interconnected and therefore the presence of one dominant super cluster has not yet occurred.

During the later stages of the reionization process ($\bar{x}_\mathrm{i} = 0.7$, bottom left) this bifurcation of the FOF-BSD continues and strenghtens, with ever more small patches merging into the large one, while smaller patches become fewer and on average ever smaller. At this stage the three models present similar volume distributions, whereas their MFP-BSD varies. BHC distribution starts to show a clear characteristic size peak. Albeit of similar shape, the BHC size distribution is clearly shifted to smaller scales, with the average bubble size smaller by a few Mpc and the distribution peak at scale about a factor of 2 smaller (8 vs 15 Mpc).

Towards late reionization ($\bar{x}_\mathrm{i} = 0.9$, bottom right), the volume limit for isolated regions to grow before merging is further reduced to $\mathrm{V} \sim 10^3\,\mathrm{Mpc}^3$, while the percolation cluster surpass volumes of $10^8\mathrm{Mpc}^3$ (i.e. close to the full simulation volume) in all the three cases. In figure 2.11, the sizes distribution in the BHC model has surpassed the other two, with average radius of 54.84Mpc. IC and SC show again similar distribution but with an increasing, although still minor, difference in the mean radius. Volume distribution in figure 2.12 present a similar situation, the only difference between IC and SC consists in the value of the volume merging limit, with a difference up to $1\mathrm{Mpc}^3$.

### 2.3.4 21-cm Signal Statistics and Power Spectra

The hyperfine transition of neutral hydrogen redshifed into meter wavelengths is a key observable of reionization. Its characteristic emission/absorption line has rest-frame wavelength $\lambda_0 = 21.1\,\mathrm{cm}$ and corresponding frequency 1.42 GHz. Radio interferometry telescopes measure the intensity of this signal by quantifying the differential brightness temperature $\delta\mathrm{T_b} \equiv \mathrm{T_b} - \mathrm{T_{CMB}}$ signal from patches of the sky, given as:

$$\delta\mathrm{T_b} \approx 28\,\mathrm{mK}(1+\delta)\mathrm{x_{HI}}\left(1 - \frac{\mathrm{T_{CMB}}}{\mathrm{T_S}}\right)\left(\frac{\Omega_\mathrm{b}\mathrm{h}^2}{0.0223}\right)\sqrt{\left(\frac{1+\mathrm{z}}{10}\right)\left(\frac{0.24}{\Omega_\mathrm{m}}\right)} \tag{2.3.3}$$

here $x_\mathrm{HI}$ is the fraction of neutral hydrogen and $1 + \delta = \langle \mathrm{n_{N,IGM}} \rangle / \mathrm{\bar{n}_{N,IGM}}$ is the local IGM overdensity. The differential brightness is characterized by the relation between the CMB temperature $\mathrm{T_{CMB}}$ and spin temperature $\mathrm{T_S}$ (see e.g. Furlanetto & Oh 2006 and Zaroubi 2012 for extended discussion). Equation 3.2.2 saturates when the neutral hydrogen decouples from CMB photons and couples with the IGM gas heated by X-ray sources (e.g. Ross et al., 2019), so that $\mathrm{T_S} \gg \mathrm{T_{CMB}}$, which is the approximation we adopt here. This is known as the heating-saturated approximation where the signal is for the majority observable in emission, $\delta\mathrm{T_b} > 0$, true only at low redshift $z < 15$. Thus in our simulation the approximated differential brightness is dependent on the density distribution of the neutral gas and redshift, such that $\delta \mathrm{T_b} \propto \sqrt{1+\mathrm{z}}\,(1+\delta)\,\mathrm{x_{HI}}$.

From the RT and N-body simulation outputs we calculate the differential brightness coeval cube at each time step. The cube is then smoothed in the angular direction by a Gaussian kernel with a FWHM of $\lambda_0\,(1+\mathrm{z})/\mathrm{B}$, where $\mathrm{B} = 2\,\mathrm{km}$ corresponds to the maximum baseline of SKA1-Low core. Smoothing along the frequency axis is done by a top-hat kernel with the same width and the above Gaussian kernel. SKA1-Low will not observe the coeval cube. Instead it will observe a lightcone, in which the signal evolves along the line of sight direction. We construct lightcones from our simulation results using the method described in Giri et al. (2018a). This method is also implemented in `TOOLS21CM`. In figure 2.13 we show the smoothed lightcone for the three different clumping models, BHC, IC and SC, respectively from top to bottom. This type of data maps the 21-cm differential brightness evolution at the observed frequency $\nu_{\mathrm{obs}} = \nu_0/(1+\mathrm{z})$, where $\nu_0 = 1.42\,\mathrm{GHz}$ is the rest frame frequency when the signal was emitted at redshift z. We then express the comoving box length in corresponding angular aperture of 4.65° at $\mathrm{z} = 6.583$.

Early on, the IGM remains mostly neutral, the average signal largely follows section 3.2.2 ($\delta \mathrm{T_b} > 30\,\mathrm{mK}$) and the fluctuations are driven by the density distribution. The gas clumping also remains low and therefore at low frequencies, $\nu_{\mathrm{obs}} > 120\,\mathrm{MHz}$, there is no visible difference between simulations. As radiation escapes the host halos, it starts to form small isolated transparent regions around sources and gradually suppresses the average signal. The H II regions are still small and thus are smoothed over by the observation beam. Figure 2.13 shows very similar evolution for the three simulations at frequency higher then 130 MHz ($z < 10$), but with different intensity of signal suppression. For example the appearances of the first transparent regions, due to lack of neutral hydrogen, at $\nu_{obs} \simeq 147\,\mathrm{MHz}$ and angular position 3.2° and 1.1° shows that ionization around sources are more consistent for the simulation with inhomogeneity dependent clumping. This is the case even at higher frequency $\nu_{obs} > 180\,MHz$ ($z < 7$), during the final phases of reionization the morphology and size of the *percolation cluster* strongly depends on the clumping model employed by the simulation. BHC model has large regions of feeble emission $\sim$ 3 mK that are extensively linked together. The IC model shows the same morphology but with considerably smaller and more isolated regions of signal. The SC model, in the other hand, shows a conspicuous lack of signal and regions of emission have only of a few Mpc size. These differences between models are more clearly observed in

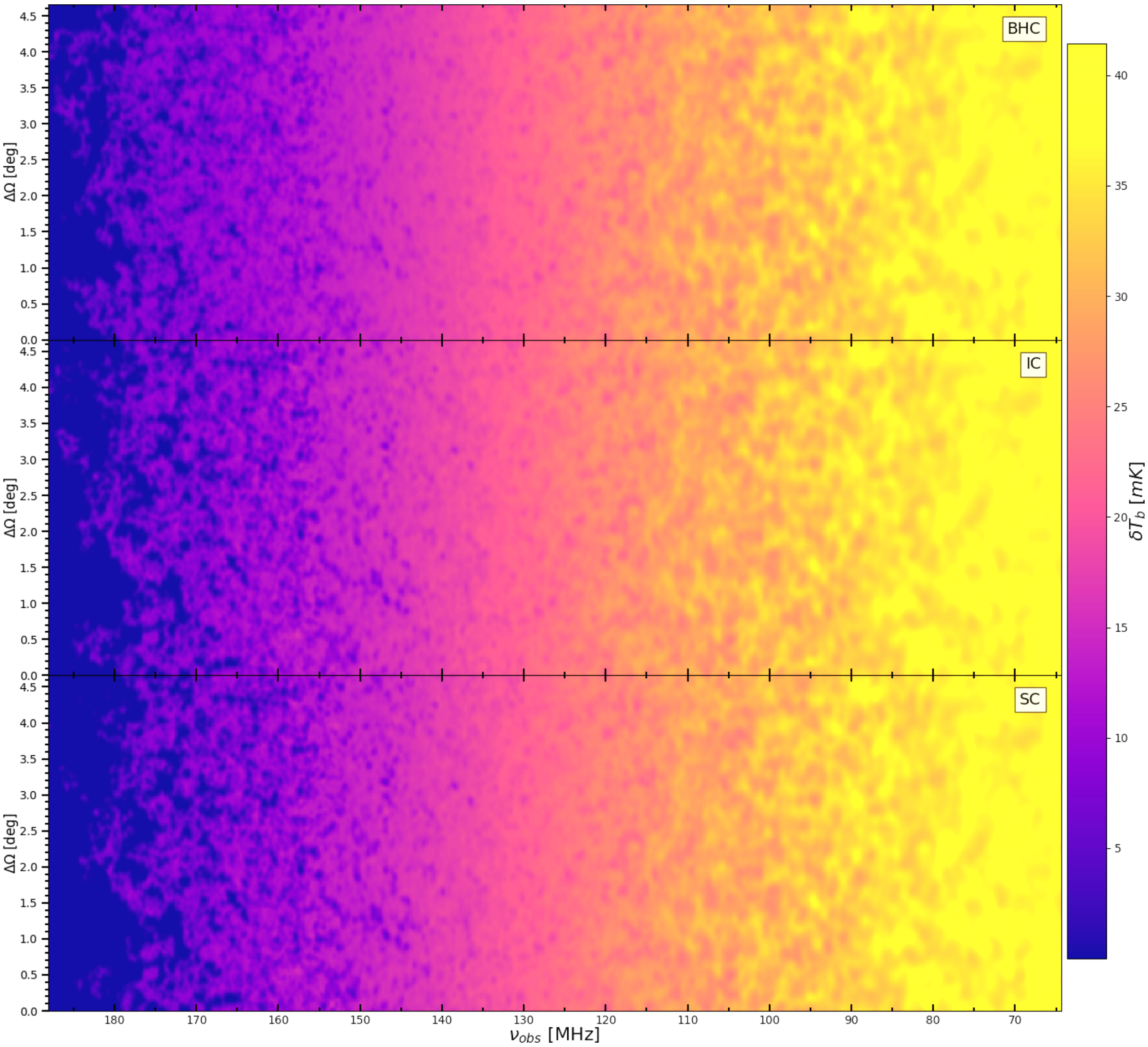

Figure 2.13: Smoothed differential brightness temperature lightcones, the colour map that shows the smoothed differential brightness $\Delta T_b$ intensity as a function of redshifted *21 cm* signal frequency $\nu_{obs}$ and aperture $\Delta\Omega$. The angular smoothing is performed by a Gaussian Kernel with FWHM $\Delta\theta$, on frequency direction is done by a top-hat kernel with same width, we use a baseline of $B = 2\,$km (maximum baseline of the core of SKA1-Low). The figure shows slice through the simulation and a comparison between BHC (top), IC (middle) and SC (bottom).

the statistics of the 21-cm differential brightness temperature fluctuations. are significant variation in the statistics of the differential brightness temperature - rms, PDFs, skewness and power spectra - shown in figure 2.14, figure 2.15, and figure 2.16.

The low frequency cut-off is chosen for range where differences between models becomes noticeable. The high density peaks get ionized early, and the corresponding H II regions are smaller then the interferometer resolution, thus their effect on rms (figure 2.14, top) is to

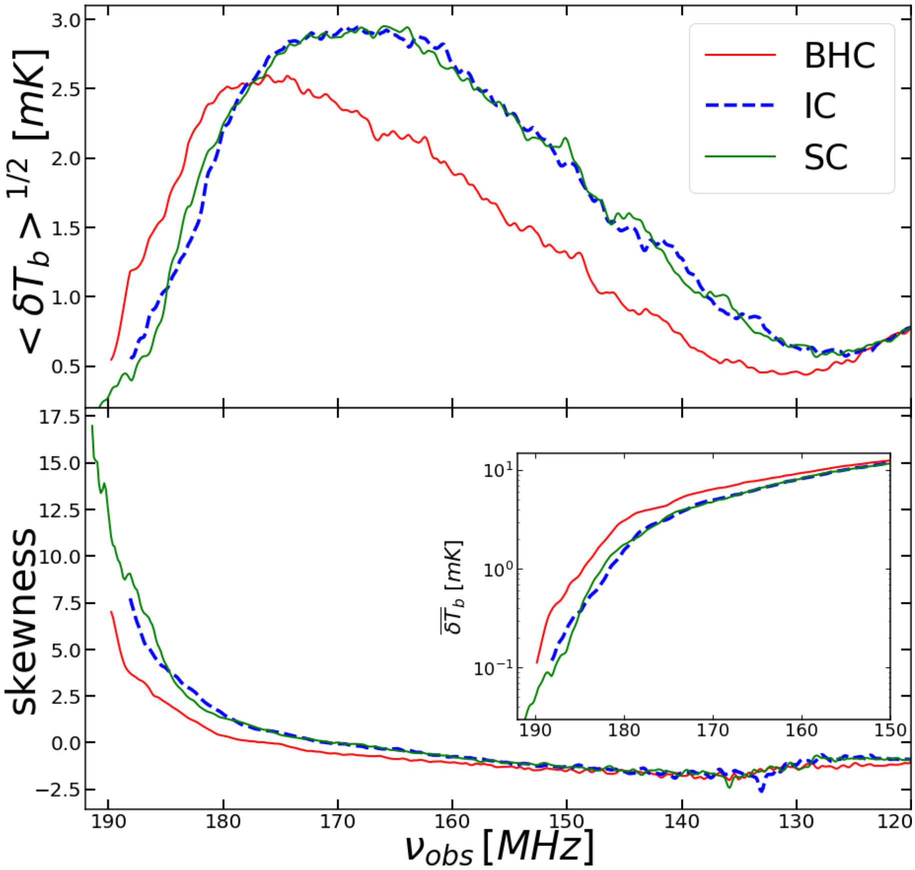

Figure 2.14: Differential brightness statistic quantities derived from the lightcones data smoothed on the core baseline of SKA1-Low (B = 2 km). Plot on top shows the frequency evolution of the signal root mean squared (RMS). Bottom plot shows the skewness and an inset panel show the frequency evolution of the averaged differential brightness in logarithmic scale.

diminish the averaged $\delta \mathrm{T_b}$ without increasing fluctuations. At this stage the signal mostly follows the underlying density field, apart from the peaks and there is little difference between the models. The observed frequency of the RMS dip indicates the timing at which HII regions become larger then the interferometry smoothing scale and eventually start to overlap locally. This is the case at frequency larger then 120 MHz ($z \approx 11$). For the IC and SC models, the turnover occurs earlier and with a steeper slope than the BHC model, indication that signal fluctuations increase faster and stronger. Moreover the peak value of the RMS fluctuations varies, in the case of IC and SC models the amplitude is 14% higher, despite having a lower averaged brightness temperature then the homogeneous case, indicating that the signal is sensitiive to a more physical treatment of the clumping factor. This is the consequence of a lower clumping factor values in under-dense regions, consistent with the conclusion in section 2.3.1. The faster propagation of I-fronts, in the vast low density regions, leads to a earlier second peak in the RMS of the two former approaches. In order of appearance at $\nu_{\rm obs}$ = 165 MHz (z = 7.56) for SC, 169 MHz (z = 7.34) for IC and slightly later at 176 MHz (z = 7.06) for BHC, respectively when the average neutral fraction is $\bar{\rm x}_{\rm n} = 0.33$, 0.28 and 0.25. The subsequent decline is the results of reionization reaching its final stage, with almost complete ionization.

The averaged 21-cm fluctuations level at different scales is reflected in the power spectra (figure 2.16), where we compare the results for models BHC, IC and SC at epochs at which the mean ionization fractions are $x_{\rm i} = 0.1, 0.3, 0.5, 0.7, 0.9$, as well as around reionization completion $x_{\rm i} = 0.99$. At first, the 21-cm signal follows the underlying density distribution of neutral hydrogen and the power spectra are very similar and approximatively a power law in all three cases. The flattening of the power spectra is an indication of the expanding ionized region, shifting the signal toward larger scales while suppressing small structures. Interestingly, this characteristic appears at the same scale regardless of the clumping model but modest difference in amplitude of signal. The BHC model yields systematically lower power at all scales and at all redshifts except close to overlap. The stochastic relation between local overdensity and clumping factor does not have a large effect throughout most of reionization, and is noticeable predominantly at small scales later on. The most significant differences between IC and SC models emerges at the end of reionization ($x_{\rm i}$ = 0.99), where the SC model has less power on all scales, by factor of up to a few. In fact, at that time the SC model has less power than even the BHC, except at the small scales $k > 0.3 {\rm Mpc}^{-1}$. The 21-cm signal fluctuations are strongly non-Gaussian (e.g. Mellema et al., 2006b; Giri et al., 2019c) and therefore are not fully described by the power spectra. We therefore also present the 21-cm differential brightness temperature distribution moments of first (PDFs; figure 2.15) and second order (skewness; figure 2.14, lower panel). For all the models and all times, 21-cm PDFs are bimodal in nature, which is a clear signature of non-Gaussianity (e.g. Ichikawa et al., 2010b; Giri et al., 2018b). Even though all the models show non-Gaussinity, there are significant variations between models. The SC and IC models are much more non-Gaussian, with many more pixels at both high low values. Particularly, they show a very strong tail at high values. This is somewhat stronger for the SC model at all redshifts, indicating that the clumping scatter yields more high brightness temperature peaks, by factor of a few. The signal skewness confirms these observations. It is going from negative to positive symmetry at $\nu_{\rm obs} \simeq 170$ MHz, when the volume ionized fraction is close to $x_{\rm i} = 0.6 - 0.7$ and the RMS fluctuations reach maximum. Differences between models are noticeable only later, once

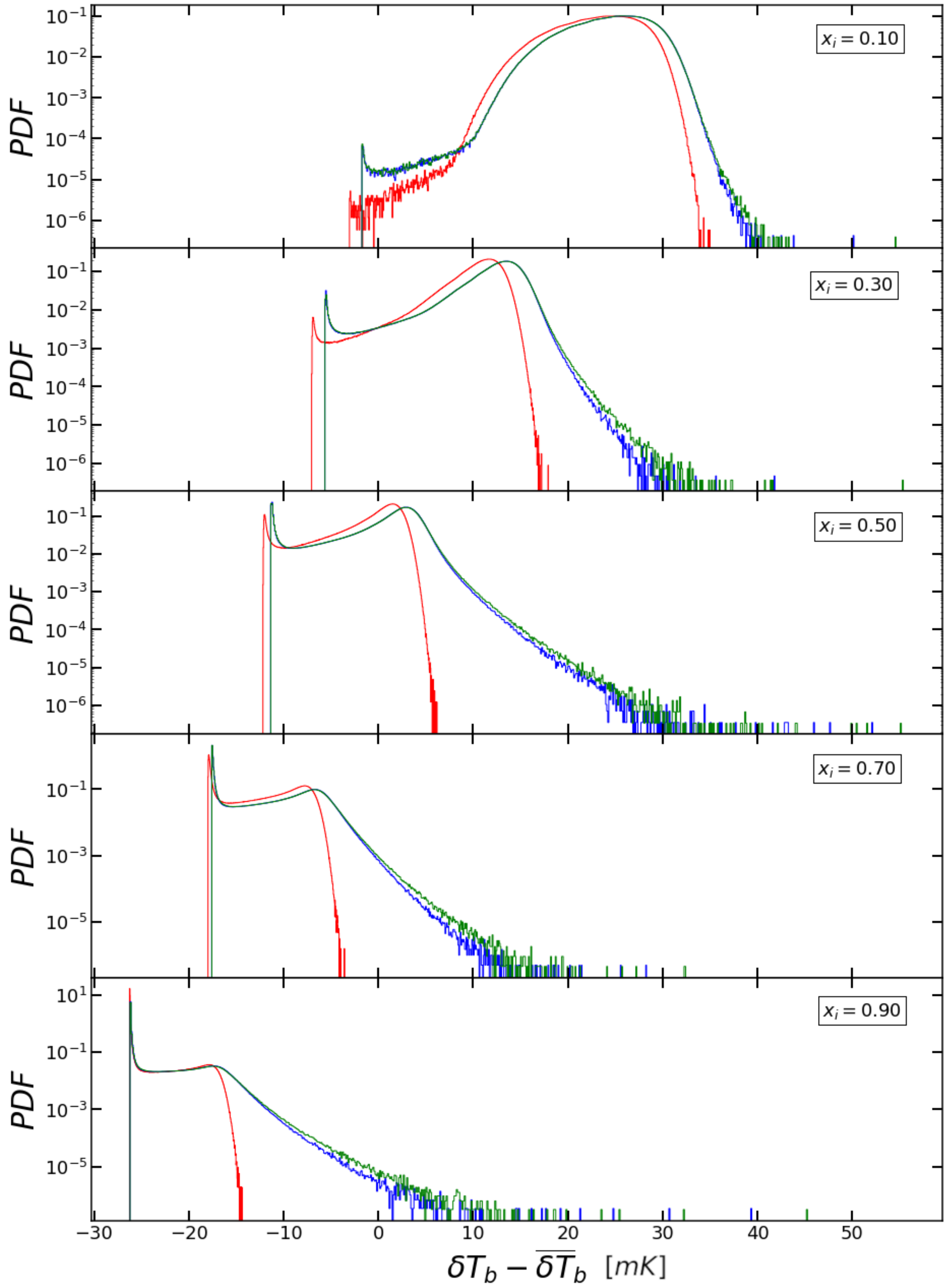

Figure 2.15: Probability distribution functions of the differential brightness temperature at ionized fractions $x_\mathrm{i} = 0.1, 0.3, 0.5, 0.7$ and 0.9, for the three clumping models, as labelled.

the simulation overpass the peak in fluctuations, at frequency larger then 180 MHz. At this point the skewness increases exponentially.

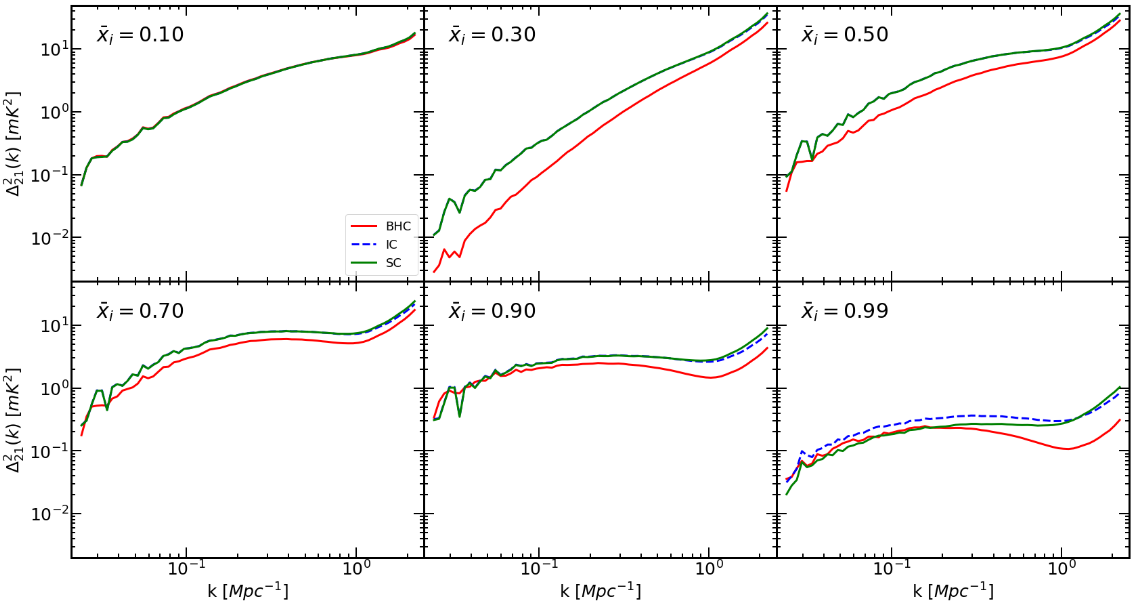

Figure 2.16: The effect of clumping factor on the 21-cm power spectra compared at volume ionization fraction $x_{\rm i} = 0.1$, 0.3, 0.5, 0.7, 0.9 and 0.99 for the models under study: BHC (red, solid), IC (blue, dashed) and SC (green, solid).

## 2.4 Conclusion

Studies of the large scale reionization morphology and its imprint on the observable signatures requires large simulated volumes of a several hundred cMpc per side. Due to computational limitations which limit the dynamic range, uniformly high resolution cannot be achieved in such a volume. Therefore no general model of the local recombinations on scale below the resolution of large numerical simulation exists. Typically a constant value of clumping factor is used, but recently we presented a more general model (Paper I), that depends on the local density, and we demonstrated how an over-simplistic treatment of the clumping factor can have a strong effect on the simulated reionization timescale, topology and size distributions of the ionized region.

In the current work we extend and improve this method by including an empirical stochastic sub-grid gas clumping (SC) model (see §2.2.3) based on the results from high-resolution N-body simulation, where the full range of relevant fluctuations is fully resolved. Our approach considers a novel parametrization of the correlation between local IGM overdensity and clumping factor, which take into accounts the scatter due to e.g. tidal forces. We employ a high resolution N-body simulation SB, of spatial resolution 260 pc per side, that resolve the Jeans length of the cold IGM and structure evolution on scale much smaller then the resolution of EoR simulations. The density-binned scatter is then modelled with a log-normal distribution. Those distributions are then randomly sampled to create a realization of the scatter. We then apply our method to the density fields of larger volumes LB-1 (714Mpc per side) and LB-2 (349 Mpc) to infer its sub-grid clumping factor (see §2.2.4). Subsequently we post-process the large scale N-body snapshot with $\texttt{C}^2\texttt{Ray}$ radiative transfer cosmic reionization simulation code, in order to present the impact of various modeling approaches for gas clumping on reionization observables (see §2.3). We then compare our stochastic model SC with the inhomogeneous clumping model, IC, which is a simple deterministic density-dependent fit, and a globally redshift dependent averaged clumping factor BHC, whereby the subgrid gas clumping is independent of the local density.

We find that density-dependent clumping models, IC and SC, exhibit similar behaviour for globally averaged quantities, meanwhile there is a tangible difference when compared to the volume-averaged model BHC. For instance, the reionization history (figure 2.5) is delayed by as much as $\Delta \mathrm{z} \sim 0.3$ at $x_\mathrm{i} = 0.7$ ($\mathrm{z} \sim 7.5$) and the average neutral fraction decrease swiftly for $\mathrm{z} < 10$. The evolution of ionized regions in IC and SC models is a bit faster due to the on average lower gas clumping factor that decreases gas recombination in the under-dense regions. Meanwhile, as structure formation advance, the higher clumping factor $\mathrm{C} > 20$ in high-density regions considerably increase the recombination rate, such that recombination is twice as effective as in the BHC model case for $\mathrm{z} < 12$. We find that the increase of rate in these regions, due to the different density-dependent gas clumping approach, is responsible for the divergence in the simulated observables. Despite the fact that the over-dense medium constitute a minor fraction of the box volume, compared to the vast under-dense IGM, it is responsible for the majority of recombinations. Our model and the IC method behave similarly, with only 5% of relative error to each other. This difference is mainly due to the broad scatter at high density in the clumping-density correlation plot (figure 2.2). The clumping factor for IGM in the proximity of sources, is extremely high $\mathrm{C} \sim 100$ and the introduced stochasticy can extend it to a factor of few hundreds more. Moreover, the simulated electron scattering optical depth is very similar in IC and SC models and the choice of the clumping model has little effect on the feedback of sources.

The density-dependence of the subgrid gas clumping accelerates the propagation of ionizing fronts in the low density IGM (figure 2.13), By $\mathrm{z} < 10$ ($\nu_\mathrm{obs} > 130\,\mathrm{MHz}$) the regions with low 21-cm signal around the sources are more pronounced than in the BHC case. The differences between the new stochastic approach and the IC model are minor, mostly appearing at late times ($z < 7$, $\nu_{obs} > 175$), where the SC scenario presents considerably less residual neutral gas then the other two models. These last region of neutral gas are mostly in large voids and distant from any ionizing sources, therefore our interpretation is that at lower redshift the empirical stochastic model becomes predominant in under-dense IGM, accelerating the propagation of ionizing radiation in these regions. Meanwhile, at early stages of reionization the gas recombination in high density region drives the reionization process, resulting in reduction of the ionizing photons propagating into the neutral surroundings.

We compared the simulation-derived observables at the same reionization milestones, $x_{\rm i} = 0.1, 0.3, 0.5, 0.7, 0.9$. Compared to our previous work, the bubble-size distributions (based on both mean free path and FOF methods) do not show large variation, as an indication that the SC model does not increase the recombination rate in a way that significantly alters the morphology and sizes of the ionized regions. The same conclusion can be deduced from statistics of the 21-cm differential brightness temperature. As we demonstrated in Paper I, the density-dependent model increase the amplitude and shift the fluctuations peak position to lower frequency with a difference of approximately 20 MHz compared to BHC model, and just a few MHz of difference when compared to the SC model. Hence, the peak occurs at stage of reionization that differ only of few percentage $\bar{\rm x}_{\rm n} \approx 0.3$ for SC and IC models and 0.25 for BHC.

The PDFs of the redshifted 21-cm distributions show some notable differences between our models. While all distributions are non-Gaussian, the IC and SC yield significantly more non-Gaussianity, with long tail of bright pixels, which is very different from the BHC model. The bright tail is longer for the SC model compared to IC, predicting many more and brighter pixels at all redshifts.

The power spectra of the 21-cm signal (see figure 2.16) show that in early phase of reionization, the BHC scenario yields a weaker signal, when compared to density dependent models on all scales. IC and SC differ somewhat at large scale k $<$ 0.1Mpc$^-$1 for $x_{\rm i} = 0.3 - 0.5$. This largely disappears by $\bar{x}_{\rm i} = 0.7$. Towards the final stages of reionization ($x_{\rm i} = 99\%$) results for three models differ. The IC model predicts the highest signal at all scales, higher by a feactor of a few compared to SC. The BHC model signal is intermediate between them for most except the smallest scales.

The results presented here are not intended as a detailed prediction of the reionization observables, but rather a demonstration that an over-simplistic treatment of the clumping factor can have strong effect on the reionization morphology and thus on simulated observables. The widely-used BHC model, overestimates the rate at which the ionized IGM recombines, and therefore have a strong influence on the timescale of reionization, morphology of the ionized region and the intensity of the expected 21-cm signal. We demonstrated that density dependent model takes better account the cumulative effect of the clumping factor on the gas recombination rate. On the other hand, we have also shown that accounting for the scatter around the average, deterministic local density-clumping relation has only modest effects on the reionization morphology and observables, predominantly towards the end of the reionization process. This indicates that the deterministic IC model is usually sufficient except possibly around and after overlap.

The gas clumping factors presented here should be considered as an upper limit to the actual clumping since they are derived based on high-resolution N-body simulations and thus do not capture the photo-ionization feedback that would suppress small-scale density fluctuations. Consequently it overestimates the recombination rate throughout reionization. We leave a more realistic approach, that follows the feedback effects, and the complex physics of the cold gas (T $< 10^4$ K) in IGM, for future work.

## Acknowledgements

ITI was supported by the Science and Technology Facilities Council (grant numbers ST/I000976/1 and ST/T000473/1) and the Southeast Physics Network (SEPNet). MB was supported by PhD Studentship from the Science and Technology Facilities Council. HP was supported by the World Premier International Research Center Initiative (WPI), MEXT, Japan and JSPS KAKENHI Grant Number 19K23455. YM is supported by the National Key R&D Program of China (Grant No. 2020SKA0110401, 2017YFB0203302, 2018YFA0404502) and the NSFC Grant (No. 11761141012, 11673014, 11821303). We acknowledge that the results of this research have been achieved using the DECI resources Kay based in Ireland at ICHEC and Cartesius based in Netherlands at SURFSara with support from the PRACE aisbl. KA was supported by NRF-2016R1D1A1B04935414 and NRF-2016R1A5A1013277, and appreciates APCTP and KASI for hospitality during completion of this work. We acknowledge PRACE for awarding us access to Piz Daint facility hosted by the Swiss National Supercomputer Centre (CSCS) and the MareNostrum IV hosted by the Barcelona Supercomputing Centre (BSC). The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for partly funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS at Juelich Supercomputing Centre

(JSC).

## Data Availability:

The data and codes underlying this article are available upon request, but can also be regenerated from scratch using the publicly available `CUBEP`$^3$`M` and `C`$^2$`Ray` code. The code and table of parameters for equation (2.2.8), 2.2.9 and 2.2.11 presented in §2.2.3 are available on the author's Github page: `https://github.com/micbia/SubgridClumping`.

## 2.5 Appendix

### 2.5.1 Comparison Between Old and New Version of `CUBEP3M`

In the N-body simulations used in our Paper I (Mao et al., 2019), we employed the version 1 of the `CUBEP`$^3$`M` code, the most recent version of the code at the time. Meanwhile in this paper we employed the updated version 2 of that code, that reduces the error of the near-grid point interpolation by extending the particle-particle (PP) force calculation for a particle out to arbitrary number of cells. With the latest version, the user can therefore choose how far outside the hosting cell the PP-force is active. A detailed discussion of this update can be found in §7.3 of Harnois-Déraps et al. (2013).

As an illustration of the effect of that change, in figure 2.17, we show the IC model of the correlation between coarse IGM over-density and coarse clumping factor at z = 7.305 for the SB simulation. In red, the interpolation obtained from N-body simulation run with first version of the code, in blue, the updated code with PP-force that extend for 2 neighbour cells. In both cases, we kept the same cosmology, initial condition and simulation parameters. In the lower panel of the figure, we show the ratio between the two old and the new result. The result of this more precise gravity forces calculation is that the gas clumping is somewhat boosted, while the curve retains the same shape, which has no significant effect on our method and results.

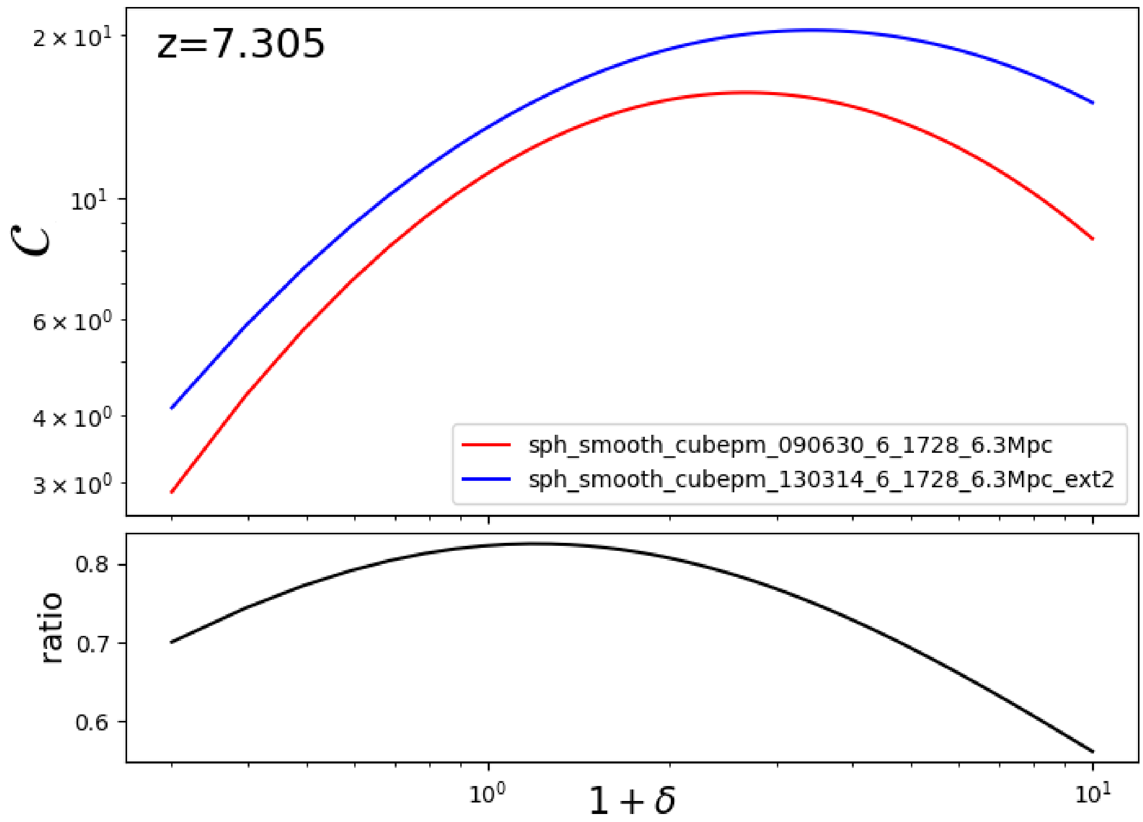

Figure 2.17: Correlation between local coarse IGM over-density and coarse clumping factor at redshift z = 7.305 for the SB simulation. In red, the IC model interpolation ran with the version 1 of the N-body code, with the solid blue line the same quantity but with the updated code. Lower panel, the ratio between the old and new quantity.

# Chapter 3

# Deep learning approach for identification of H II regions during reionisation in 21-cm observations

In this chapter, we present `SegU-Net`, a stable and reliable method for identifying neutral/ionised regions in simulated 21-cm tomographic data images, presented in section 1.2.2, for the upcoming in SKA-Low radio interferometry telescope.

The content in this chapter can be found in Bianco et al. (2021a), which has been accepted for publication on the MNRAS *as is*. The author developed and wrote the network code and ran the semi-numerical simulation to create the dataset. Sambit K. Giri provided the Python package tools for the statistical analysis of the outputs, and the author adopted them to our results. The author implemented the presented deep learning method to the same package tool, made all the figures and wrote the text with help from Sambit Giri. Ilian T. Iliev and Garrelt Mellema provided valuable comments on the methods and results.

## 3.1 Introduction

SKA-Low will observe a sequence of 21-cm images from different redshifts that will constitute a three-dimensional set of data known as a tomographic dataset. The evolution of the 21-cm signal can be seen along the redshift axis. See for example Giri (2019) for more description about tomographic 21-cm images. The reionization process is driven by growing H II regions, often referred to as bubbles (e.g. Furlanetto et al., 2004). As the sources of ionizing photons reside inside them, observing these bubbles and their evolution will be interesting. Numerous studies have provided various methods to detect and study properties of H II bubbles (e.g. Datta et al., 2007; Zackrisson et al., 2020; Mason & Gronke, 2020). We can also study the properties of reionization with 21-cm images (Giri et al., 2018a; Giri et al., 2019a). However, tomographic images from SKA-Low will be prone to instrumental limitations, such as noise, limited resolution and foreground contamination (e.g. Koopmans et al., 2015; Ghara et al., 2017). In the field of image processing, methods that can classify objects or features in images into meaningful segments are known as 'image segmentation' methods. Giri et al. (2018a) implemented an image segmentation method to classify neutral and ionized regions in 21-cm images in the presence of instrumental limitations and demonstrated that key properties of reionization can be derived from such observations.

Artificial intelligence (AI) and deep learning methods are capable of learning patterns in image data and identifying interesting regions. Image segmentation based on AI is quite popular in the field of data analysis and has been applied to study objects with complex visual form contained in big data (Long et al., 2014). In recent years, several papers made use of machine learning techniques for a range of problems in astrophysics (e.g. Lee, 2019; Giri et al., 2019b; Yoshiura et al., 2020; Chen et al., 2020b) and cosmology (e.g. Jeffrey et al., 2020; Sadr & Farsian, 2020; Guzman & Meyers, 2021). In the case of reionization, several of these methods are aimed to either remove foreground emission (Li et al., 2019; Makinen et al., 2021; Villanueva-Domingo & Villaescusa-Navarro, 2021), emulate reionization simulations (e.g. Kern et al., 2017; Schmit & Pritchard, 2018; Jennings et al., 2018; Cohen et al., 2020; Ghara et al., 2020) or constrain reionization history (e.g. Shimabukuro & Semelin, 2017; Chardin et al., 2019; Mangena et al., 2020; Shimabukuro et al., 2020) and its astrophysical inputs (e.g. Sullivan et al., 2017; Gillet et al., 2019; Hassan et al., 2020).

In this work, we present a new approach for the identification of the distribution of H II regions in 21-cm images using a deep learning method named U-shaped convolutional neural network (U-Net), which is specially designed for image segmentation and feature extraction (Ronneberger et al., 2015). In our case, we adapt this network for processing our image data, which are mock observations of the 21-cm signal during the EoR. The method will segment the images into ionized and neutral regions. We call this framework SegU-Net.

This paper is organised as follows. In § 3.2 we present how we generate the simulated data sets used for this work. In § 3.3 we describe the design of our neural network, including the error estimation. In § 3.4 we discuss its application to our simulated SKA-Low data sets, considering a range of summary statistics such as the mean ionization fraction, power spectra and topological quantities such as size distributions and Betti numbers. In § 3.5 we test our framework on various instrumental noise levels, and in § 3.6 we test it on a data set produced from a fully numerical reionization simulation. We discuss and summarize our conclusions in § 3.7.

## 3.2 21-cm Signal

For any deep learning based method, we need a data set containing a sample of all the possible scenarios, known as the training set. In § 3.2.1, we describe the reionization simulation code that we use to create the training set. The observable for radio telescopes observing the 21-cm signal is defined in § 3.2.2. Finally, in § 3.2.3 we give the methodology we use to mimic the observations expected with SKA-Low.

### 3.2.1 Reionization Simulation

To train our network, we require a large set of simulations that represent the 21-cm radio signal for a wide range of redshift during reionization and different assumptions about the astrophysical sources of ionizing radiation. To do so we employ py21cmFAST, the Python wrapped version of the semi-numerical cosmological simulation code 21cmFAST (Mesinger et al., 2010; Murray et al., 2020). The code computes the evolution of the matter density field using the Zel'dovich approximation (Zel'Dovich, 1970). The ionization field and the corresponding 21-cm differential brightness temperature are then calculated from the matter density distribution based on the excursion set formalism (Furlanetto et al., 2004; Mesinger & Furlanetto, 2007), which considers a region to be ionized when the fraction of collapsed matter fluctuation exceeds a mass threshold. The ionization fraction $x_{\mathrm{HII}}(\boldsymbol{r})$ at a position $\boldsymbol{r}$ is given as,

$$x_{\mathrm{HII}}(\boldsymbol{r}) = \begin{cases} 1 & \text{if } f_{\mathrm{coll}} \geq 1/\zeta \\ 0 & \text{otherwise} \end{cases} \tag{3.2.1}$$

where $\zeta$ is the ionizing efficiency of high redshift galaxies and $f_{\mathrm{coll}}(R_{\mathrm{s}}, M_{\mathrm{min}})$ is the fraction of collapsed matter within radius $R_{\mathrm{s}}$ that can form haloes with mass greater than $M_{\mathrm{min}}$. $f_{\mathrm{coll}}$ is calculated at every pixel varying $R_{\mathrm{s}}$ within 0 and $R_{\mathrm{mfp}}$. The maximum value of $f_{\mathrm{coll}}$ is used in equation (3.2.1). $R_{\mathrm{mfp}}$ implements the effect of a finite mean free path for ionizing photons in the ionized IGM.

The cosmological parameters considered in this work are based on WMAP 5 years data observation (Komatsu et al., 2009) and consistent with Planck Collaboration (2019) results. We assume a flat $\Lambda$CDM cosmology with the following parameters, $\Omega_\Lambda = 0.73$, $\Omega_m = 0.27$, $\Omega_b = 0.046$, $H_0 = 70\, km\, s^{-1}\mathrm{Mpc}^{-1}$, $\sigma_8 = 0.82$, $n_s = 0.96$.

### 3.2.2 Differential Brightness Temperature

Radio interferometry based telescopes record the differential brightness temperature $\delta T_{\mathrm{b}}$ while observing the redshifted 21-cm signal. $\delta T_{\mathrm{b}}$ depends on position on the sky $\boldsymbol{r}$ and redshift $z$ and can be given as (e.g. Mellema et al., 2013),

$$\delta T_{\mathrm{b}}(\boldsymbol{r}, z) \approx 27 x_{\mathrm{HI}}(\boldsymbol{x}, z)\big(1 + \delta_{\mathrm{b}}(\boldsymbol{r}, z)\big) \left(\frac{1+z}{10}\right)^{\frac{1}{2}} \left(1 - \frac{T_{\mathrm{CMB}}(z)}{T_{\mathrm{s}}(\boldsymbol{r}, z)}\right) \left(\frac{\Omega_{\mathrm{b}}}{0.044}\frac{h}{0.7}\right) \left(\frac{\Omega_{\mathrm{m}}}{0.27}\right)^{-\frac{1}{2}} \mathrm{mK} \tag{3.2.2}$$

where $x_{\mathrm{HI}}$, $\delta_{\mathrm{b}}$, $T_{\mathrm{CMB}}$ and $T_{\mathrm{s}}$ are neutral fraction, baryon density contrast, CMB temperature and spin temperature respectively.

Previous studies have shown that our Universe will be heated before reionization begins (e.g. Pritchard & Furlanetto, 2007; Ross et al., 2017, 2019). Therefore we assume $T_{\mathrm{s}} \gg T_{\mathrm{CMB}}$ throughout this work, which is known as the spin saturated approximation and is relevant at lower redshift $z \lesssim 12$ (e.g. Furlanetto et al., 2004; Furlanetto, 2006b). In the spin saturated approximation scenario, the differential brightness signal is always in emission ($\delta T_{\mathrm{b}} \geq 0$ mK) and locations with $\delta T_{\mathrm{b}} = 0$ mK correspond to H II regions.

### 3.2.3 Mock 21-cm Observation

In order to train `SegU-Net` for application to actual observations, we need a training set of mock observations. We create these mock observations by simulating the $\delta T_{\mathrm{b}}$ using the methods described in previous sections and adding instrumental effects, such as the

Table 3.1: The parameters used in this study to model the telescope properties.

| Parameters | Values |
|---|---|
| System temperature | $60(\frac{\nu}{300\text{MHz}})^{-2.55}$ K |
| Effective collecting area | 962 $\text{m}^2$ |
| Declination | -30° |
| observation hour per day | 6 hours |
| Signal integration time | 10 seconds |

absence of zero baselines, limited resolution and noise. We follow the methods in Ghara et al. (2017) and Giri et al. (2018a) for mimicking the expected effects of SKA1-Low.

We consider a simulation volume of $(256\,\text{Mpc})^3$ and an intrinsic resolution of $\Delta x = 2\,\text{Mpc}$ for simulating the signal. This intrinsic resolution corresponds to an angular aperture of $\Delta\theta = 0.777$ arcmin and a frequency depth of $\Delta\nu = 0.124\,\text{MHz}$ along the line of sight at $z = 7$. As an example, in figure 3.1, we show a coeval cube slice of the neutral fraction field and $\delta T_\text{b}$ field in the top left and bottom left panels, respectively. These slices are taken from the epoch when the universe was about 50 per cent ionized. For each $\delta T_\text{b}$ coeval cube, we assume one axis as the line of sight or frequency direction and subtract the mean signal from each frequency channel, such that this could be considered as a sub-volume from the 3D tomographic data set. We consider this simulation as our reference throughout the results analysis in §3.4, its astrophysical parameters are given in table 3.2. We simulate the instrumental noise using the method given in Giri et al. (2018a) and implemented in `Tools21cm`[1] (Giri et al., 2020). We change the noise seed for each new member of the training set so that the network is trained on different noise realisations and we list our assumed parameters for the telescope setup in table 3.1. In the top right panel of figure 3.1, we show a slice from the simulated noise cube produced from 1000 hours of observation with SKA1-Low at simulation resolution. When we add this noise to our simulated signal at the simulation resolution, we cannot discern any feature of the signal as the noise is several orders of magnitude higher than the signal. Therefore we reduce the resolution of the noisy signal in the field-of-view direction by smoothing with a Gaussian kernel with full-width at half maximum (FWHM) of $\lambda_0(1+z)/B$, where $B$ is the maximum baseline. For example, $B = 2\,\text{km}$ corresponds to a resolution of 2.905 arcmins at redshift $z \approx 7$ and 3.631 arcmins at redshift $z \approx 9$ respectively. In the frequency direction we reduce the resolution by convolving with a top-hat bandwidth filter of a width matching

[1] A python package for EoR simulations analysis. https://github.com/sambit-giri/tools21cm

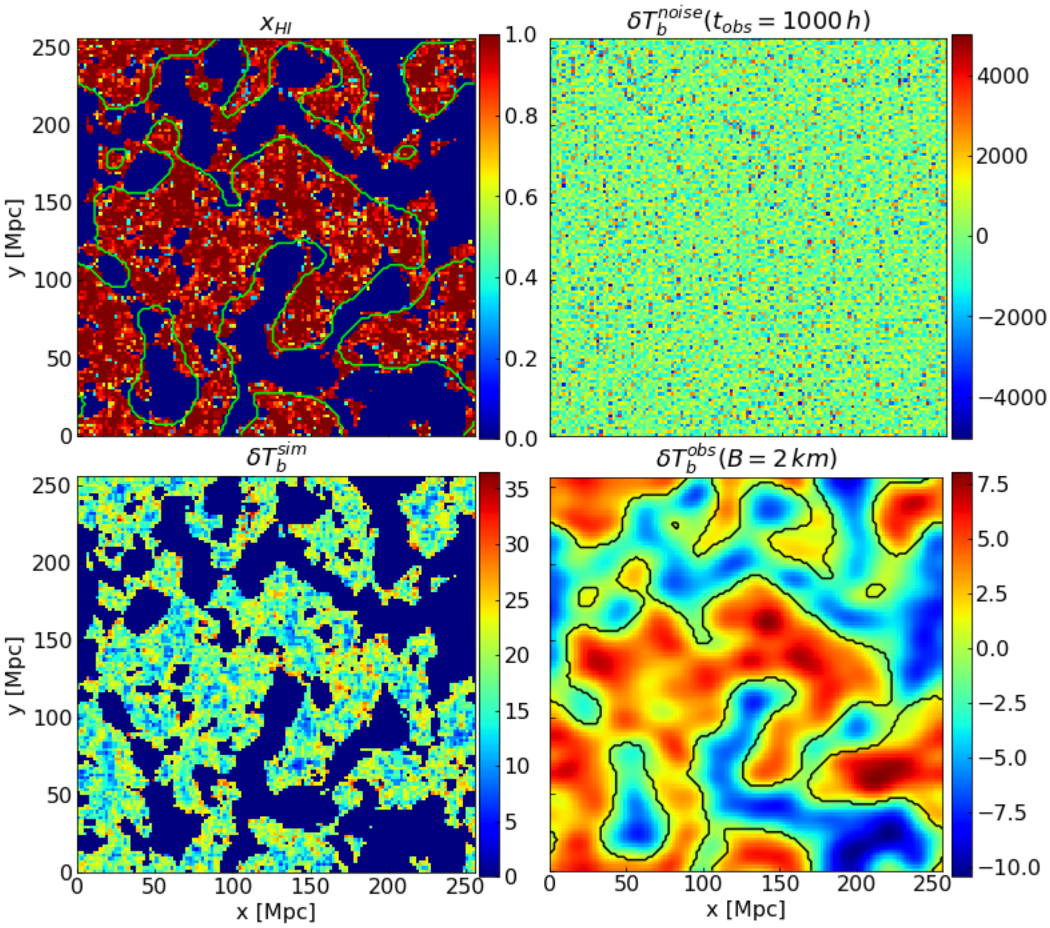

Figure 3.1: *Top left*: the neutral hydrogen fraction at simulation resolution. Green contours indicate the boundary between neutral and ionized regions after reducing the resolution to an observation with a maximum baseline of $B = 2\,\text{km}$ and matching frequency resolution. *Bottom left*: the 21-cm signal at simulation resolution. *Top right*: The 21-cm signal plus noise realisation at simulation resolution for an observing time of 1000 hours. To mimic the effect of the lack of a zero baseline, the mean signal has been subtracted. *Bottom right*: The noisy 21-cm image after smoothing to the resolution to an observation with a maximum baseline of $B = 2\,\text{km}$ and matching frequency resolution. This is an example of a smoothed box slice used during the network training. The solid black line shows the same contour as in the top left panel.

the FWHM of the angular smoothing in comoving units. This width corresponds to 0.462 MHz at redshift $z \approx 7$ and 0.551 MHz at redshift $z \approx 9$ respectively. In the bottom right panel of figure 3.1 we show a slice from our noisy signal at this reduced resolution. At this resolution, the smallest H II regions seen in the top left panel of figure 3.1 can no longer be discerned. However, we can still identify the larger H II regions.

To illustrate what we can achieve with these images, we apply the same smoothing

Table 3.2: Astrophysical parameters used for our fiducial simulation.

| Parameters | Values |
|---|---|
| $\zeta$ | 39.204 |
| $R_{\mathrm{mfp}}$ | 12.861 Mpc |
| $T_{\mathrm{vir}}^{\mathrm{min}}$ | $3.46 \times 10^4$ K |

to the neutral fraction field and apply a threshold of $x_{\mathrm{th}} = 0.5$ to label neutral/ionized regions. We refer to the smoothed and then binarised neutral fraction field as the *ground truth*. We use this field to compare the accuracy of the recovered binary field throughout our paper. We want to point out that this is different from the ground-truth of the original reionization simulation as the limited resolution of the radio telescope will limit the observation of small scale features. Then, we over-plot the boundaries of these ionized regions, the neutral fraction slice and signal slice in top-left and bottom-right panels of figure 3.1 respectively.

### Training and Testing Set

For our training set, we randomly sample the astrophysical simulation parameters by a normal distribution, such that the ionizing efficiency of high-redshift galaxies $\zeta$ is sampled with $\mathcal{N} \sim (52.5, 20)$, the mean free path of ionizing photons $R_{\mathrm{mfp}}$ with $\mathcal{N} \sim (12.5\,\mathrm{Mpc}, 5\,\mathrm{Mpc})$ and the (logarithmically-spaced) minimum virial temperature for halos to host star-forming galaxies $T_{\mathrm{vir}}^{\mathrm{min}}$ with $\mathcal{N} \sim (4.65, 0.5)$. The choice of these values is such that for a majority of the samples most of the reionization history ($x_{\mathrm{HI}}^{\mathrm{V}}$ from 0.9 to 0.1) falls within the redshift interval 9 to 7. The redshift is randomly sampled with a uniform distribution $\mathcal{U} \sim [7, 9]$. The initial conditions of the cosmological density field are changed for each simulation. This helps us avoid the impact of cosmic variance on our trained model. With the list of all the parameter values, we produce 10,000 mock observations of the 21-cm signal. Out of these mock observations, we use 15 per cent as the so-called network validation set. This validation set is used during the training method to provides an unbiased evaluation of the network model fit.

Eventually, we will use `SegU-Net` on actual 21-cm image observations. Here we rely on an additional 300 mock observations as the testing or prediction set. Just as for the training set, the parameter values are randomly chosen. We call this the ‘random’ testing set. The training process is blind to the prediction set. Apart from the above testing set, we create an additional simulation with fixed values of astrophysical parameters (given in table 3.2). We have chosen these values such that between $z = 9$ and 7 reionization proceeds from $x_{\rm HI}^{\rm V} \approx 0.9$ to 0.1. We call this set the 'fiducial' testing set. Since the signal evolves as reionization progresses in this testing set, it better mimics the upcoming 21-cm observations. With this testing set, we will test `SegU-Net`'s capability to capture the evolution of structures and recover the binary field from untrained data in § 3.4.

### Fully-Numerical Simulations Testing Set

To train `SegU-Net`, we relied on `21cmFAST` for creating the training set. However, our Universe may not exactly be described by this semi-numerical code. If our neural network has learnt to find structures in `21cmFAST` simulations only, then we cannot use it for SKA observations. To ensure that the neural network is not over-fitted, we consider a different reionization simulation code to build the mock observations.

We first simulate the matter density field and track the evolution of cosmic structures by using the `CUBEP`$^3$`M` $N$-body code (Harnois-Déraps et al., 2013). The simulation is carried out in a volume of $(349{\rm Mpc})^3$ with 64 billion particles. Dark matter haloes down to a mass of $10^9 M_\odot$ is found at various redshift using the spherical average halo-finder (Watson et al., 2013), meanwhile haloes with masses between $10^8$ and $10^9\ M_\odot$ are implemented with a sub-grid method (Ahn et al., 2015b). We use the same cosmology that is given in § 3.2.1.

We then employ the `C`$^2$`RAY` radiative transfer (RT) code (Mellema et al., 2006b) to simulate the cosmic reionization. `C`$^2$`RAY` requires the matter density field in a 3D grid. Therefore the distribution $N$-body particles are put in 3D grids with a smoothed particle hydrodynamic method (e.g. Shapiro et al., 1996; Mao et al., 2019). This grid has spatial resolution of $\Delta x = 2.1$ Mpc and a $166^3$ mesh-grid. Sources ionizing photon production rate per unit time is proportional to the mass of the hosting halo $M_{\rm halo}$ such that.

$$\dot{N}_\gamma = f_\gamma \frac{M_{\rm halo}\, \Omega_b}{\Delta t_s\, m_p\, \Omega_m} \tag{3.2.3}$$

where $m_p$ is the proton mas and $\Delta t_s = 11.53\,{\rm Myr}$ is the stars lifetime. The efficiency factor of sources is defined as $f_\gamma = f_\star\, f_{\rm esc}\, N_i$ where $f_\star$ is the star formation efficiency, $f_{\rm esc}$ is the photons escape fraction and $N_i$ is the stars ionizing photon production efficiency per stellar atom. The efficiency factor for halos with masses $M_{\rm halo} < 10^9 M_\odot$ is set to $f_\gamma = 2$. For the the lower mass halos it is initially set to $f_\gamma = 8.2$. When their environment becomes ionized (above 10 percent), their efficiency is reduced to $f_\gamma = 2$ to account for radiative feedback. `C`$^2$`RAY` outputs the hydrogen ionization field at a time interval of 11.5 million years. For more details on the RT and $N$-body simulations methods, see **?** and Bianco et al. (2021b). We derive the differential brightness temperature $\delta T_{\rm b}$ from the ionization

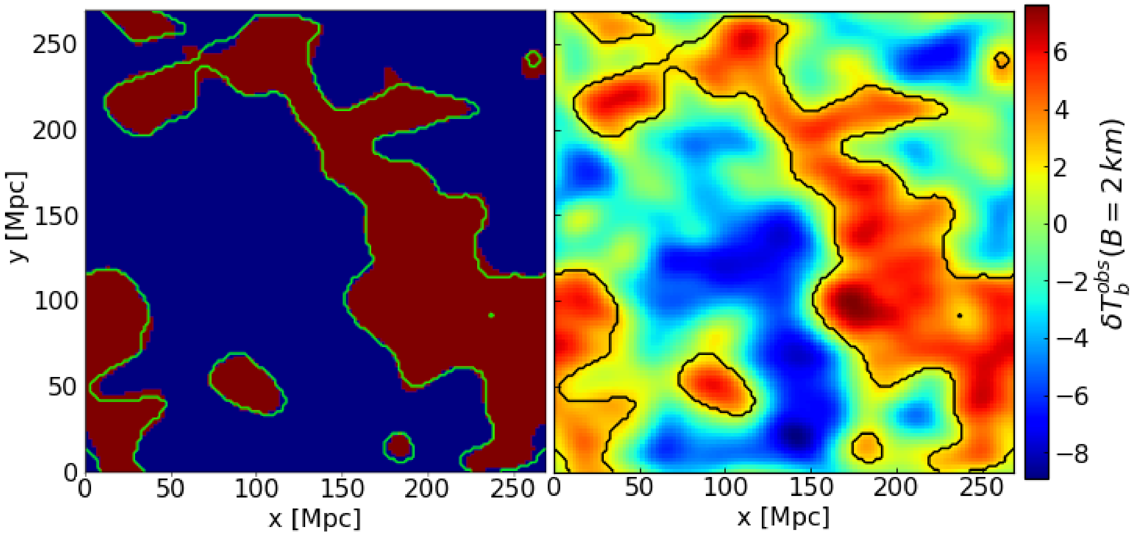

Figure 3.2: *Right panel*: An example of a smoothed cube slice from the C$^2$RAY simulation on the right employed to test the stability and reliability of the network. This slice is for $z = 8.06$ and corresponds to a volume-averaged neutral fraction of $x_{\rm HI}^{\rm V} = 0.38$. As for the training set, at simulation resolution, we subtracted the mean signal in the frequency direction from the differential brightness temperature. We then added simulated instrumental noise for the observed time of 1000 hours and smoothed the signal with the same baseline as SKA1-Low. *Left panel*: the binary field recovered with our neural network. In red/blue, the prediction performed with our network and the green contour shows the boundary between neutral and ionized region. The same contour is shown with a solid black line on the right panel for comparison.

field and the density using section 3.2.2. We select four outputs, which are at redshifts $z = 7.96, 8.06, 8.17, 8.28, 8.40, 8.52, 8.64$, corresponding to a volume averaged neutral fraction of $x_{\rm HI}^{\rm V} = 0.17, 0.29, 0.42, 0.57, 0.70, 0.81, 0.90$, respectively. The simulated $\delta T_{\rm b}$ from these epochs are converted into mock observations using the procedure outlines in § 3.2.3. We use these mock observations as a testing set.

The right panel of figure 3.2 shows a slice of the calculated $\delta T_{\rm b}$ for redshift $z = 8.06$ ($x_{\rm HI}^{\rm V} = 0.38$ at simulation resolution). Similar to the bottom right panel of figure 3.1, we add the instrumental noise corresponding to a 1000 h observation and smooth the signal to a resolution corresponding to a maximum baseline $B = 2\,\rm km$. The black contours correspond to the boundary between neutral and ionized regions. These boundaries are derived from the simulated neutral fraction field at the same resolution as the $\delta T_{\rm b}$ data set.

## 3.3 U-Net for 21-cm Image Segmentation

Here we describe our machine learning method for identifying ionized and neutral regions in noisy 21-cm images and our approach to estimate the uncertainty of its results in § 3.3.1 and § 3.3.2 respectively.

### 3.3.1 Our Network, `SegU-Net`

Our segmentation network[2] is based on the U-Net framework first introduced by Ronneberger et al. (2015). U-Net consists of two likewise symmetric paths, an encoder operator that contracts the image and a decoder operator that expands the extracted features. The encoder corresponds to a classical convolutional neural network (CNN). This CNN aims to reduce the size of the input image in such a way that only information of the most interesting features remains. A series of concatenated convolution operations (layers) returns a low dimensional latent space (or latent vector) that contains information about these extracted features. In Appendix 3.8.1 we provide a visual representation of the low dimensional latent space for the example case of a sphere. We show a schematic representation of the U-Net in figure 3.3. The left part of the U-shape in the diagram and the bottom layer represents the encoder and the low dimensional latent space respectively. For a detailed discussion of CNNs, we refer the reader to Mehta et al. (2019), and for examples of employing CNNs to infer cosmological and astrophysical parameters in the context of reionization to Gillet et al. (2019) and Hassan et al. (2020). In our case, the information in the latent space (or latent vector) of U-Net (bottom layer) is expanded by a decoder into a binary map of the same size as the input image. The right part of the U-shape of the diagram in figure 3.3 represents the decoder. The decoder gradually increases the spatial resolution of the latent vector with an up-sampling operation (transposed convolution) until we obtain the same dimension of the input image. After each up-sampling step, the output is combined with the corresponding encoder layer with the same dimension. We illustrate this further in Appendix 3.8.2 with an example.

Even though each of our image data sets is 3D, `SegU-Net` is trained on 2D slices. We identify structures in 3D image data by running on every slice along the third axis. Tests show that the method is not sensitive to the choice of the third axis. When compared to a neural network trained on 3D data, we found that our approach is computationally less expensive without loss of accuracy. Therefore the U-Net architecture described in this work is only applied to 2D image data. The structure of the encoder layers consists of

[2] `https://github.com/micbia/SegU-Net`

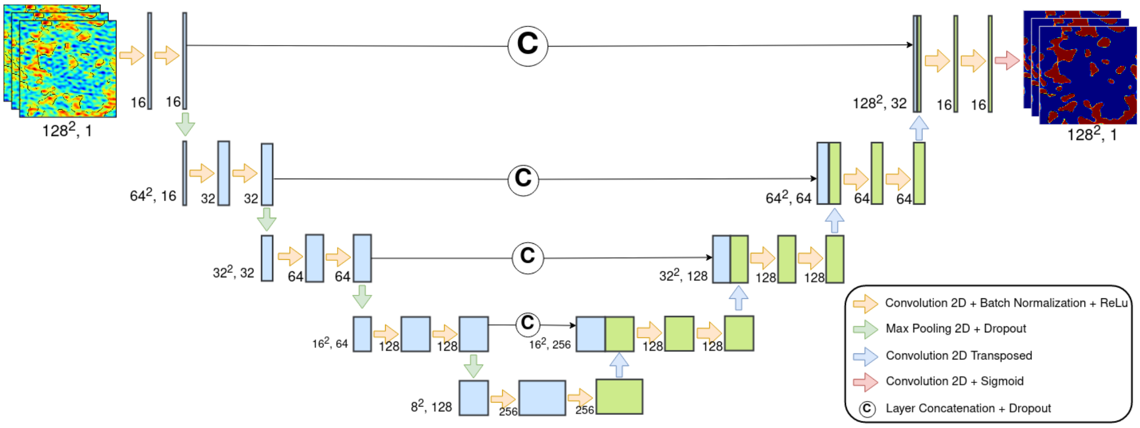

Figure 3.3: Schematic representation of `SegU-Net` network architecture. The orange arrow indicates a 2D convolutional layer, followed by batch normalization and ReLU activation. Pooling operations followed by dropout layer are indicated with green arrows. The blue arrow indicates an up-sampling layer by transposed 2D convolutional layer and with a red arrow the closing layer, a 2D convolution followed by a sigmoid activation function. The descending path on the left side divides the resolution of the image after each pooling operation and doubles the channel dimension after each convolution. On the other hand, the expansion path doubles the spatial dimension at each up-sampling operation and decreases the channel dimension after concatenation with its counterpart layer in the descending path.

two convolutional blocks followed by a 2D max-pooling layer (`MaxPool`) of size 2x2 and a 5 per cent rate dropout layer (`Drop`). This regularization technique randomly shuts down a portion of the layer neurons to avoid over-fitting (Hinton et al., 2012; Srivastava et al., 2014). The convolutional block (`ConvBlock`) consists of a 2D convolution layer (`Conv2D`) with 3x3 kernel size. We add a layer that normalizes the previous input layer over the batch sample to avoid over-fitting (`BN`) (Ioffe & Szegedy, 2015) and as an activation function we employ a Rectified Linear Unit (`ReLU`) activator (Jarrett et al., 2009; Glorot et al., 2011), `ConvBlock=Conv2D+BN+ReLU`. This layer structure is repeated for a total of four levels (`Encoder-Level`). At each step, the dimension of the input image is halved by the max-pooling operation. The number of feature channels is doubled by the convolutional layer, `Encoder-Level=2*ConvBlock+MaxPool+Drop`. The decoder structure is somewhat similar to the encoder. We replace the pooling operation with a transposed 2D convolution (`TConv2D`) (Dumoulin & Visin, 2016; Zeiler & Fergus, 2013), that has an opposite scaling effect on the resolution and channel size. This layer output is then concatenated (`CC`) with the corresponding encoder level to preserve the features extracted in the contracting path. This step is followed by a dropout layer and two convolutional blocks, `Decoder-Level=TConv2D+CC+Drop+2*ConvBlock`. The final output consists of a 2D convolutional layer followed by a sigmoid activation. Our network has a total of 23 2D convolutional layers distributed on four down- and up-sampling scaling levels and a bottom layer, for a total of approximately 2.5 million trainable parameters. In figure 3.3, we show our best performing network and label the shape of the output from each intermediate hidden layer of this network. More details are provided in Appendix 3.8.1 and 3.8.2.

During our training process, the hyperparameters of the network are learnt by minimizing a loss function. We employ the balanced cross-entropy (BCE) (Salehi et al., 2017),

$$\mathcal{L}(\mathrm{y},\hat{\mathrm{y}}) = -\frac{1}{\mathrm{N}}\sum_{\mathrm{i=0}}^{\mathrm{N}}\left(\beta\,\mathrm{y_i}\log_{10}(\hat{\mathrm{y}}_\mathrm{i}) + (1-\beta)(1-\mathrm{y_i})\log_{10}(1-\hat{\mathrm{y}}_\mathrm{i})\right) \tag{3.3.1}$$

where $y_i \in \{0;1\}$ is the pixel-wise ground truth, $\hat{y}_i$ the predicted value, $N$ the batch size, which is our case is of size 32 and the parameter $\beta = \frac{1}{N}\sum_{i=0}^{N}(1-y_\mathrm{i})$ is the average volume ionised fraction of the batch. In our context, at early/late stage of reionization the statistical weight of the ionized/neutral pixels are under-represented. This situation is known in data science as a problem affected by *"class unbalanced" data*. To deal with this we use the above loss function which has been shown to be well suited for segmentation problems that are affected by class unbalanced data (Cui et al., 2019). We further used the Adaptive Moment Estimator Adam (Kingma & Ba, 2014), an optimized stochastic gradient descent algorithm for error minimization. The initial learning rate, the step size of the rate of convergence that minimizes the loss function, is set to $10^{-3}$. We trained the network using 2 GPUs, and it took approximately 1,500 wall clock hours.

### 3.3.2 Uncertainty Estimation On `SegU-Net`

One of the main drawbacks of machine learning is that it is unable to quantify uncertainties and confidence intervals for its predictions, and only recently attempts have been made (Charnock et al., 2020; Hortúa et al., 2020) to include error estimation. However, this has not yet been generally implemented for U-Nets. Additionally, if not well optimized, neural networks are prone to over-fitting and tend to be biased. Therefore, we have developed an error estimation procedure to be used during the prediction process. This procedure gives our network additional power by providing a pixel by pixel error map.

Image manipulations, such as zooming, shifting along an axis, flipping axes and rotation along an axis, are commonly performed on 2D or 3D image training data to increase

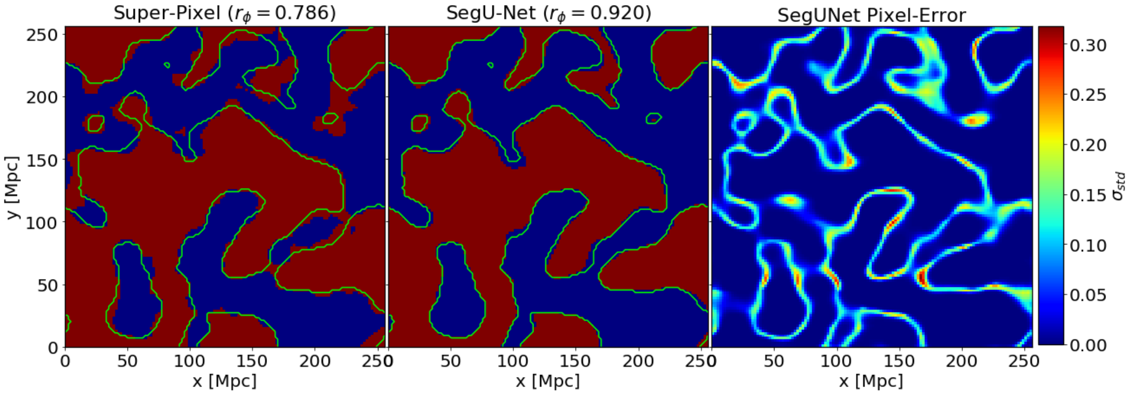

Figure 3.4: Slice comparison of the binary field, in blue ionized regions and in red neutral. *Left panel*: binary field recovered by the Super-Pixel method. *Central panel*: binary field recovered by our neural network. Green lines indicate the true separation between ionized/neutral regions, derived from a smoothed version of the simulated neutral hydrogen distribution. *Right panel*: the per-pixel error as calculated by `SegU-Net`. The color-bar indicates the intensity of the network uncertainty.

the number of samples (Simonyan & Zisserman, 2015; Szegedy et al., 2015). This technique is known as time-test augmentation (TTA) of data (Perez & Wang, 2017; Wang et al., 2020). Here we use this approach to estimate the error on the final result. We perform several copies (∼ 100) of the same sample during the prediction process through image manipulations. These manipulated copies are then independently processed by `SegU-Net`. Each of the recovered binary fields is transformed back. We calculate the final result as the average of these fields and the per-pixel standard deviation to estimate the error for each pixel.

An example of the pixel per pixel error map can be seen in figure 3.4 (right-most panel). We will discuss this figure further in § 3.4.1. This simple method provides our neural network with an uncertainty estimation for each labelled pixel.

## 3.4 Results

Once the network is trained, we want to estimate how well it recovers the binary field from noisy 21-cm images. To do so, we include in our analysis the state-of-the-art Super-Pixel method presented in Giri et al. (2018a). The Super-Pixel method is based on an advanced image processing technique called the Simple Linear Iterative Clustering (SLIC) (Achanta et al., 2012). SLIC groups similar pixels in images into *"super-pixels"*. These Super-Pixels are then classified into neutral and ionized ones to get the final map containing the identified features. In previous studies, this method has been shown to be superior compared to other methods, such as putting a simple threshold to the mean signal (e.g. used in Kakiichi et al., 2017), the k-means method (e.g. used in Giri et al., 2018a) or the maximum deviation method (e.g. used in Gazagnes et al., 2021). The Super-Pixel method proves to be quite efficient in recovering the binary fields from noisy 21-cm images. The summary statistics extracted from those are accurately reproducing the ones obtained using the simulation data sets. As shown by Giri et al. (2018a), the choice for the number of super-pixels depends on the simulation box size and resolution. In our case, we tested for a few values between 500 and 7000. We noticed that above the value of 5000, the algorithm becomes more computationally expensive without yielding a substantial increase in the segmentation accuracy. Hence we employ 5000 super-pixels.

### 3.4.1 Visual Comparison

To start, we show a visual comparison of slices in figure 3.4. We compare the predicted binary field recovered by the Super-Pixel method (left-most panel) and `SegU-Net` (central panel) with the ground truth (green contours in both panels). As explained in § 3.2.3 the ground truth is the boundary of ionized regions extracted from the simulation neutral fraction field at the same resolution by putting a threshold of 0.5. The red and blue pixels represent neutral and ionized pixels, respectively. In the right-most panel, we show the pixel-error estimated from `SegU-Net` with a colour bar. The error is determined by calculating the standard deviation of the same pixel from the different version of the same mock observation produced with TTA (see § 3.3.2).

`SegU-Net` shows better precision in recovering shapes of the ionized regions compared to the Super-Pixel method. As expected, most of the network uncertainty is located at the boundaries of neutral regions or between two large ionized bubbles when these are percolating, and the gap is getting narrower. This uncertainty has a direct bearing on small neutral islands of a few Mpc scale, residing in vast ionized regions. Moreover, larger uncertainties, $\sigma_{\rm std} \geq 0.25$ are located around narrow ionized regions protruding into large neutral regions (e.g. in figure 3.4 right-most panel, at coordinates $x \sim 140\,{\rm Mpc}$ and $y \sim 125\,{\rm Mpc}$). This behaviour suggests that the uncertainty mainly depends on the contrast between the local neutral and ionized regions. The network selects regions in the image based on the largest gradient in the 21-cm signal intensities to recover the binary field. Therefore, we expect larger uncertainties for reionization scenarios in which the contrast in the 21-cm intensities are relatively small.

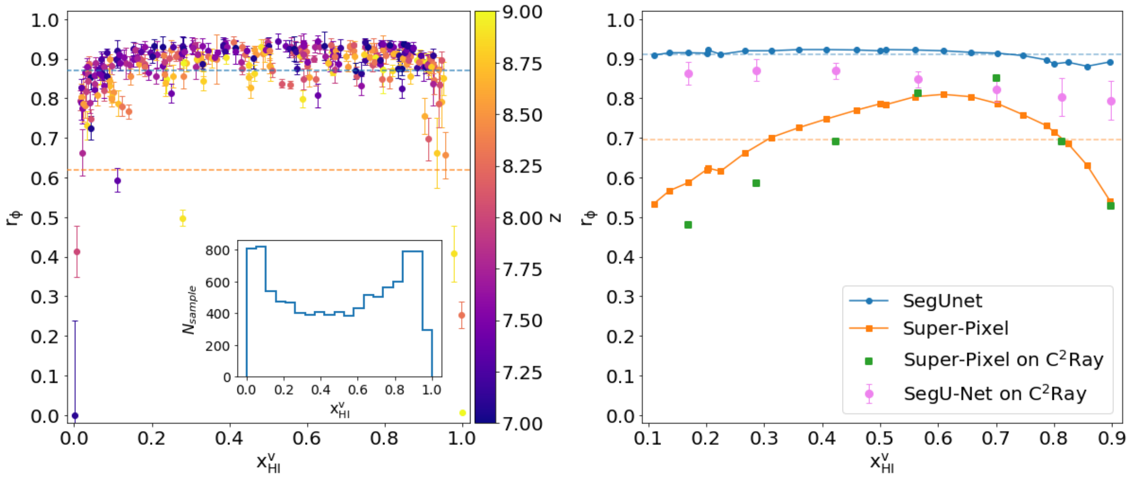

Figure 3.5: *Left panel*: the Matthews correlation coefficient $r_\phi$ of the recovered binary field for the prediction set, against its volume-averaged neutral fraction. Error-bar indicates the network confidence interval, and colours indicate the redshift of the simulated coeval cube. On the inset panel, we show the distribution of the training set (blue histogram) against the volume average neutral fraction. *Right panel*: comparison of the same correlation coefficient for recovery performed on the fiducial simulation with our neural network (blue circle line) and the Super-Pixel method (orange square line). We also include the result from the test on the C$^2$RAY simulation, from left to right, redshift $z = 7.96$, 8.06, 8.17, 8.28, 8.40, 8.52, 8.64 corresponding to a volume-averaged neutral fractions of $x^{\rm V}_{\rm HI} = 0.17$, 0.29, 0.42, 0.57, 0.70, 0.81, 0.90. The violet dots with relative confidence intervals are predictions performed with SegU-Net for these cases and the green squares are the corresponding results from the Super-Pixel method. Horizontal dashed lines in both panels indicate the overall average $r_\phi$ coefficient for the entire data set and the fiducial simulation respectively, in blue for `SegU-Net` and orange for Super-Pixel method.

### 3.4.2 Correlation Coefficient

To compare the predicted ionized fields from the 21-cm images mathematically, we use the Matthews correlation coefficient (MCC) (also known as $r_\phi$ coefficient) defined as:

$$r_\phi = \frac{N_{\rm TP}\cdot N_{\rm TN} - N_{\rm FP}\cdot N_{\rm FN}}{\sqrt{(N_{\rm TP}+N_{\rm FP})(N_{\rm TP}+N_{\rm FN})(N_{\rm TN}+N_{\rm FP})(N_{\rm TN}+N_{\rm FN})}}\ , \tag{3.4.1}$$

where $N_{\rm TP}$ and $N_{\rm TN}$ are the total numbers of neutral and ionized pixels recovered correctly, respectively. $N_{\rm FP}$ is the total numbers of pixels incorrectly guessed as neutral and $N_{\rm FN}$ is the total numbers of pixels incorrectly guessed as ionized. In our case, a positive/negative result corresponds to the neutral/ionized case since the quantity 1 in our binary fields

Table 3.3: Summary of the Matthews correlation coefficient score (in per cent) of our two test sets for the two feature identification methods.

| | SegU-Net | | Super-Pixel | |
|---|---|---|---|---|
| redshift | random set | fiducial | random set | fiducial |
| $z \leq 7.75$ | 88.9% | 91.7% | 63.7% | 62.6% |
| $z \geq 8.25$ | 85.3% | 90.1% | 60.7% | 71.8% |
| $7 \leq z \leq 9$ | 87.1% | 91.2% | 62.0% | 69.5% |

indicates the neutral condition and 0 the ionized. Thus, MCC is a useful metric to correlate binary fields. In figure 3.5 we show the MCC estimated from the fields segmented into ionized and neutral regions in our testing sets. In the left panel, we provide a scatter plot of MCC values against the reionization history ($x^{\mathrm{v}}_{\mathrm{HI}}$) for the 'random' testing set. We indicate the redshift of the realization by the colour of the points and respective confidence interval with an error bar. We show the number of samples in our training set at a different neutral fraction in an inset panel. After a first attempt, we realized that to overcome the unbalanced class problem requires a better representation of the early ($x^{\mathrm{v}}_{\mathrm{HI}} \approx 1$) and late stages of reionization ($x^{\mathrm{v}}_{\mathrm{HI}} \approx 0$). For this reason, we increased the number of training samples for these stages. Therefore the distribution of samples against neutral fraction has a bimodal shape with peaks at approximately $x^{\mathrm{v}}_{\mathrm{HI}} \approx 0.1$ and 0.9.

As a result, the $r_\phi$ value for the overall prediction data set (figure 3.5, left panel) is about 87 per cent for `SegU-Net` (blue dashed line) and 62 per cent in the case of the Super-Pixel method (orange dashed line). The noise level increases with redshift. Therefore the score is slightly less accurate for redshift $z \geq 8.25$ with an 85 per cent accuracy, meanwhile higher for lower redshift $z \leq 7.75$ with 88 per cent. In the future, we consider increasing the proportion of the training data with high redshift to decrease this performance dissimilarity. The same trend is present in the case of the Super-Pixel method, with an accuracy of 60 per cent and 63 per cent, respectively.

In the right panel of figure 3.5, we compare the MCC values from `SegU-Net` (blue line with circles) with that from the Super-Pixel method (orange line with squares) for our 'fiducial' simulation. As we already know from Giri et al. (2018a), the Super-Pixel method performs best for $x^{\mathrm{V}}_{\mathrm{HI}} \approx 0.5$ and deteriorates towards earlier and later stages of reionization. The reason for this behaviour is that during these stages, structures are usually smaller and, therefore, more difficult to identify. With `SegU-Net` we are able to overcome this problem by employing a specifically designed BCE loss function (equation (3.3.1)) during

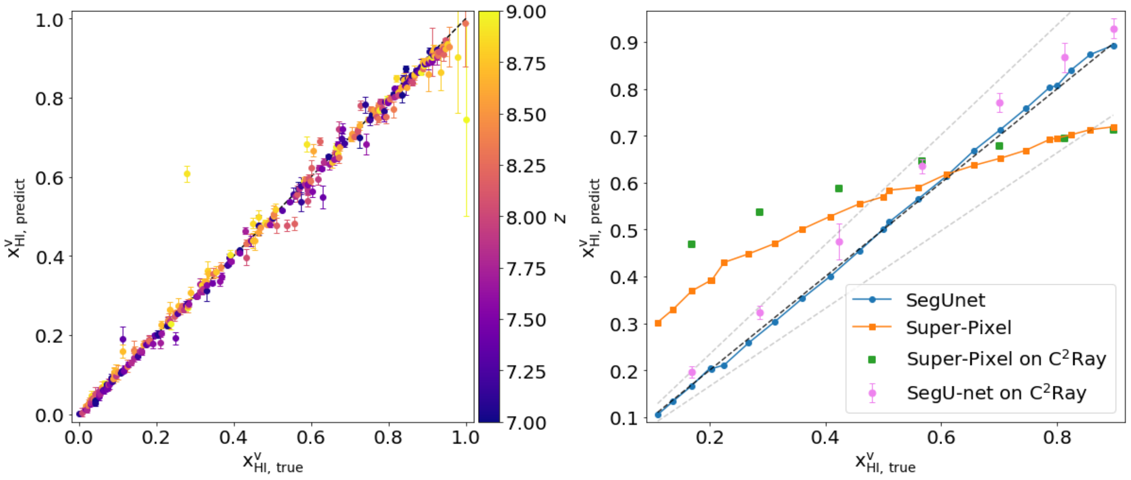

Figure 3.6: *Left panel*: Comparison of the simulated neutral fraction against the recovered one. Error-bar and color-bar are the same as figure 3.5. *Right panel*: The same comparison for the 'fiducial' simulation. We also include the results from C$^2$RAY simulation. The redshift of C$^2$RAY simulation are $z$ = 7.96, 8.06, 8.17, 8.28, 8.40, 8.52, 8.64 corresponding to a volume averaged neutral fraction of $x^{\rm v}_{\rm HI}$ = 0.17, 0.29, 0.42, 0.57, 0.70, 0.81, 0.90. The violet dots with relative confidence interval are predictions performed with SegU-Net and green squares are with the Super-Pixel method.

the validation process after each training epoch. Therefore, the average $r_\phi$ value for the 'fiducial' simulation is about 91 per cent for SegU-Net (blue dashed line) and 70 per cent in the case of the Super-Pixel method (orange dashed line). In table 3.3, we summarise the $r_\phi$ score for the two test sets.

### 3.4.3 Average Neutral Fraction

After identifying the ionized regions, we can determine the volume-averaged neutral fraction $x^{\rm v}_{\rm HI}$, which quantifies the reionization history. In figure 3.6 we show the volume-averaged neutral fraction $x^{\rm v}_{\rm HI,\,predicted}$ as calculated from the recovered binary fields extracted by the two methods. In the left panel we show the $x^{\rm v}_{\rm HI,\,predicted}$ from the SegU-Net outputs against the true volume-averaged neutral fraction $x^{\rm v}_{\rm HI,\,true}$ for our 'random' testing set. The colour of the points indicates the redshifts. The black dashed line indicates $x^{\rm v}_{\rm HI,\,predicted} = x^{\rm v}_{\rm HI,\,true}$. Except for a few points, all the points lie on or near the black dashed line.

In the right panel of figure 3.6 we compare the results of $x^{\rm V}_{\rm HI,\,predicted}$ derived with the Super-Pixel method (orange line with squares) and SegU-Net (blue line with circles) for our 'fiducial' simulation. Again, the black dashed line represent $x^{\rm V}_{\rm HI,\,predicted} = x^{\rm V}_{\rm HI,\,true}$.

In the case of our neural network all results lie within the half standard deviation (0.5-$\sigma$) of the true value (gray dashed lines). With Super-Pixel method, this is true only from $x^{\rm V}_{\rm HI} \approx 0.5$ to 0.85. The recovered neutral fraction is either underestimated at $x^{\rm V}_{\rm HI} > 0.6$ or largely overestimate for $x^{\rm V}_{\rm HI} < 0.4$.

### 3.4.4 Size Distributions

From the 3D tomographic data that will be produced with the upcoming SKA experiment, we will be able to study the size distribution of neutral or ionized region during the EoR. ionized regions are often called bubbles and whereas neutral regions are referred to as islands. The Bubble and Island size distributions (BSDs and ISDs) are useful to derive the properties of reionization and its evolution (Xu et al., 2017; Giri et al., 2019a). Several approaches were presented to calculate this distribution (Lin et al., 2016b; Kakiichi et al., 2017; Giri et al., 2018a). In this work, we employ the Mean-Free-Path method (MFP; Mesinger & Furlanetto, 2007; Giri et al., 2018a) to calculate the size distribution ($R\frac{{\rm d}N}{{\rm d}R}$) of recovered neutral (ISD) and ionized field (BSD). Previous works have demonstrated that this method should be preferred since the calculated size distributions are almost unbiased (Lin et al., 2016b; Giri et al., 2018a).

In the left and right columns of figure 3.7 we show the ISDs, respectively BSDs of the binary fields recovered with `SegU-Net` (blue line) and Super-Pixel method (orange line) compared to the ground truth (black dashed line). The Super-Pixel method performs best when the simulation is halfway through the reionization process $x^{\rm v}_{\rm HI} = 0.5$ (central panel). However, it is considerably less accurate compared to `SegU-Net`. We show the relative difference with the ground truth in the plots below the ISDs and BSDs. The blue shaded region shows the error on each of the size distributions determined by `SegU-Net`. In both the ISD and BSD case, the main difference between the two recovered distribution occurs at the earlier $x^{\rm v}_{\rm HI} = 0.8$ (top) and later $x^{\rm v}_{\rm HI} = 0.2$ (bottom) stages of reionization. `SegU-Net` shows a relative difference of a few per cent while the distributions determined from the Super-Pixel segmentations show relative differences up to 10 per cent for large sizes.

### 3.4.5 Dimensionless Power Spectra

The dimensionless power spectrum of the neutral field is defined as $\Delta^2_{\rm xx} = k^3 P_{\rm xx}(k)/2\pi^2$, where $P_{\rm xx}$ is the auto-power spectrum that quantifies the fluctuations due to the distribution of neutral regions. These fluctuations contribute to the 21-cm power spectrum

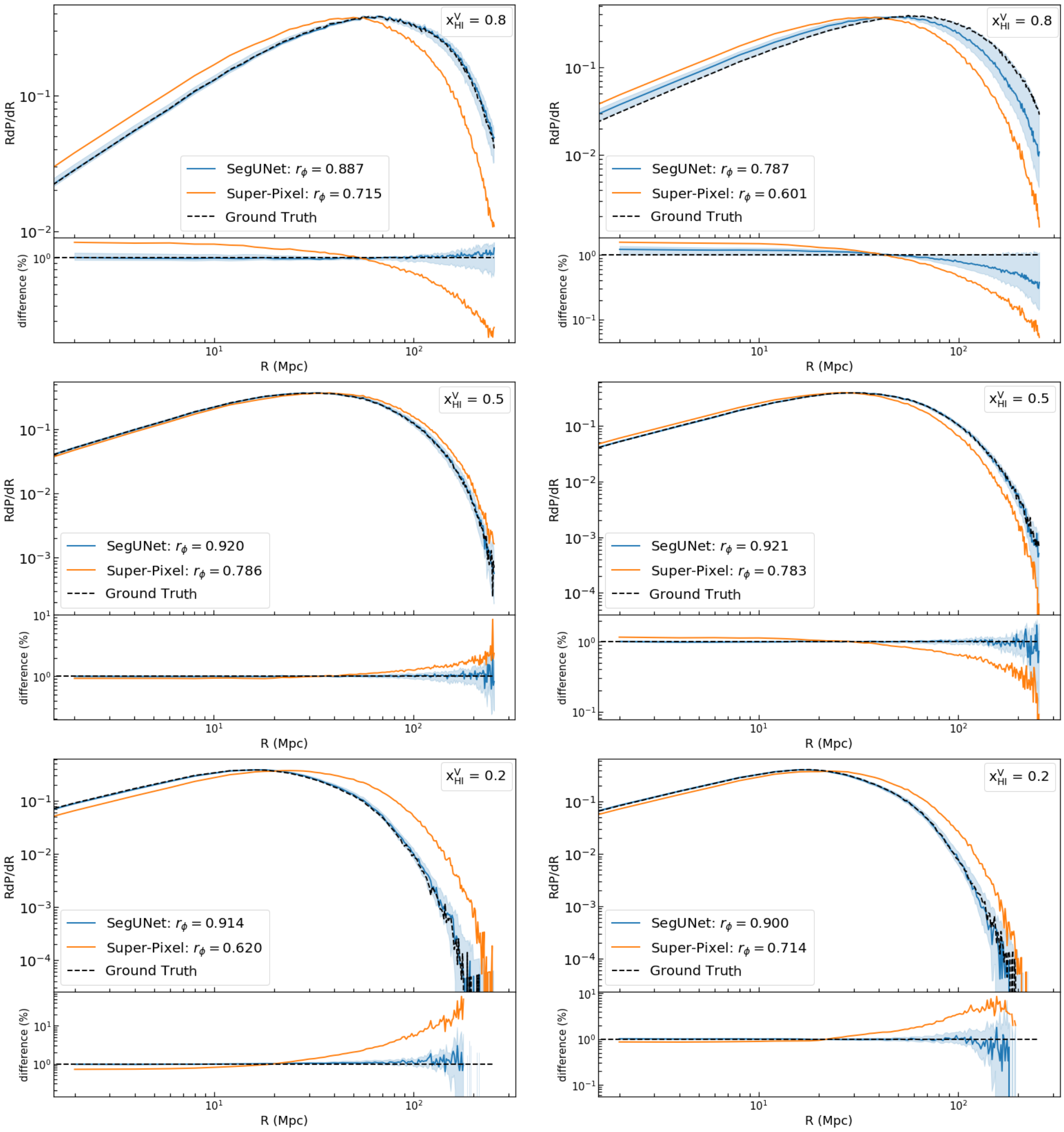

Figure 3.7: *Left column*: the size distribution of neutral regions (ISD). *Right column*: the size distribution of ionized region (BSD). Rows from top to bottom represents early ($x^{\rm v}_{\rm HI} = 0.8$), middle ($x^{\rm v}_{\rm HI} = 0.5$) and late ($x^{\rm v}_{\rm HI} = 0.2$) stages of reionization respectively. On each panel, we show the size distributions from the binary fields of the 'fiducial' simulation recovered by `SegU-Net` (blue line) and its respective confidence interval (blue shadow). Black dashed lines and orange lines give the size distributions of the ground truth and binary field recovered by the Super-Pixel method. At the bottom of each size distribution panel, we show the relative deviation from the true binary field distribution.

that is observed with radio interferometric telescopes. See for example Furlanetto et al. (2006) and Lidz et al. (2007) for descriptions of the fluctuations of the 21-cm signal. In this section, we study the $\Delta^2_{\rm xx}$ estimated from the neutral fields recovered from various methods.

In figure 3.8, we consider the 'fiducial' simulation at three stages of reionization, which are $x^{\mathrm{V}}_{\mathrm{HI}} = 0.8$ (top panel), $x^{\mathrm{V}}_{\mathrm{HI}} = 0.5$ (central panel) and $x^{\mathrm{V}}_{\mathrm{HI}} = 0.2$ (bottom panel). At the mid-point of reionization (central panel), the Super-Pixel method performs well at large scales $k < 0.2\,\mathrm{Mpc}^{-1}$ with a relative difference within 25 per cent for lower k-values. The $\Delta^2_{\mathrm{xx}}$ of the neutral field recovered by the Super-Pixel method at early and late times have the correct shape but differ in magnitude. The $\Delta^2_{\mathrm{xx}}$ of the neutral field recovered by `SegU-Net` match the ground truth well at all three stages of reionization. The network maintains a maximum difference compared with the ground truth, of a few tens of per cent at all scales. For $k \lesssim 0.5\,\mathrm{Mpc}^{-1}$, the network uncertainty interval grows to 25-50 per cent relative difference.

### 3.4.6 Betti Numbers

During reionization ionized bubbles form, grow and connect with each other to form a complex topology (Furlanetto & Oh, 2016b). Various studies have proposed topological descriptors for this distribution, such as Euler characteristics (e.g. Friedrich et al., 2011) and Betti numbers (Elbers & van de Weygaert, 2019; Giri & Mellema, 2021; Kapahtia et al., 2021). Giri & Mellema (2021) pointed out that Betti numbers contain more information compared to the Euler characteristics. Therefore in this section we study the zeroth $\beta_0$, first $\beta_1$ and second $\beta_2$ Betti number (Betti, 1870) of the binary 3D maps recovered by the two feature identification methods.

$\beta_0$, $\beta_1$ and $\beta_2$ describe the number of isolated ionizing regions, tunnels and isolated neutral regions, respectively. In the top, middle and bottom panels of figure 3.9, we show the $\beta_0$, $\beta_1$ and $\beta_2$ values estimated from the recovered binary fields of our 'fiducial' model at $x^{\mathrm{V}}_{\mathrm{HI}}$ between 0.1 and 0.9. The black, blue and orange curves represent the Betti numbers calculated from the ground truth, recovered field with `SegU-Net` and Super-Pixel method, respectively. In line with the results for the other quantities discussed above, we find that the topology recovered with `SegU-Net` is much closer to the ground truth than the one recovered by the Super-Pixel method.

## 3.5 Tests on Different Instrumental Noise levels

We have trained and tested `SegU-Net` for one specific noise level, corresponding to the theoretically expected noise for $t_{\mathrm{obs}} = 1000$ h with the current design of SKA-Low. However, in practice, the noise level may differ from this, either because the observing time or

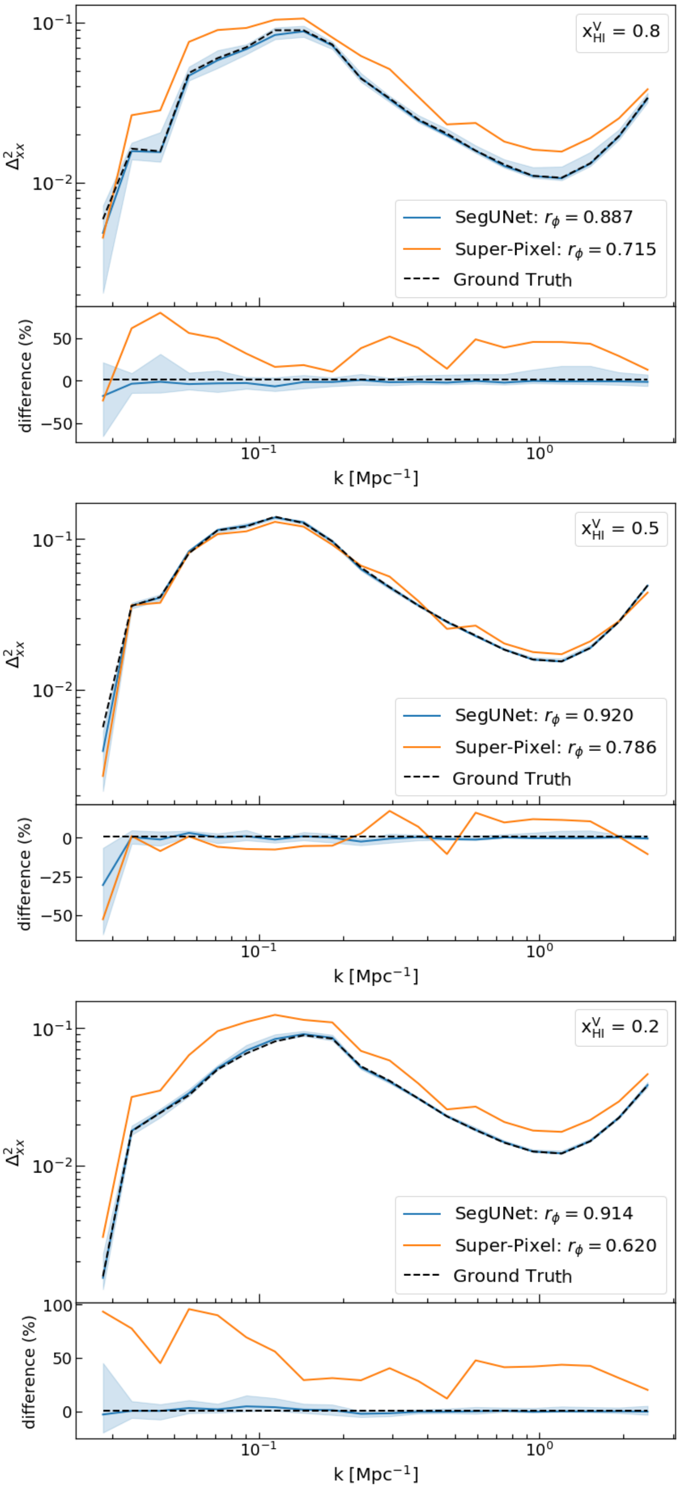

Figure 3.8: Dimensionless power spectra of the neutral field from the fiducial simulation as recovered by our network (blue line) and its respective confidence interval (blue shadow). Compared at early, middle and late stage of reionization (from top to bottom $x^{\mathrm{V}}_{\mathrm{HI}} = 0.8,\ 0.5,\ 0.2$) with the same quantity derived from the ground truth (black dashed line) and the Super-Pixel method (orange line). At the bottom of each panel, we show the relative difference compared to the ground truth for both cases, the network and Super-Pixel method.

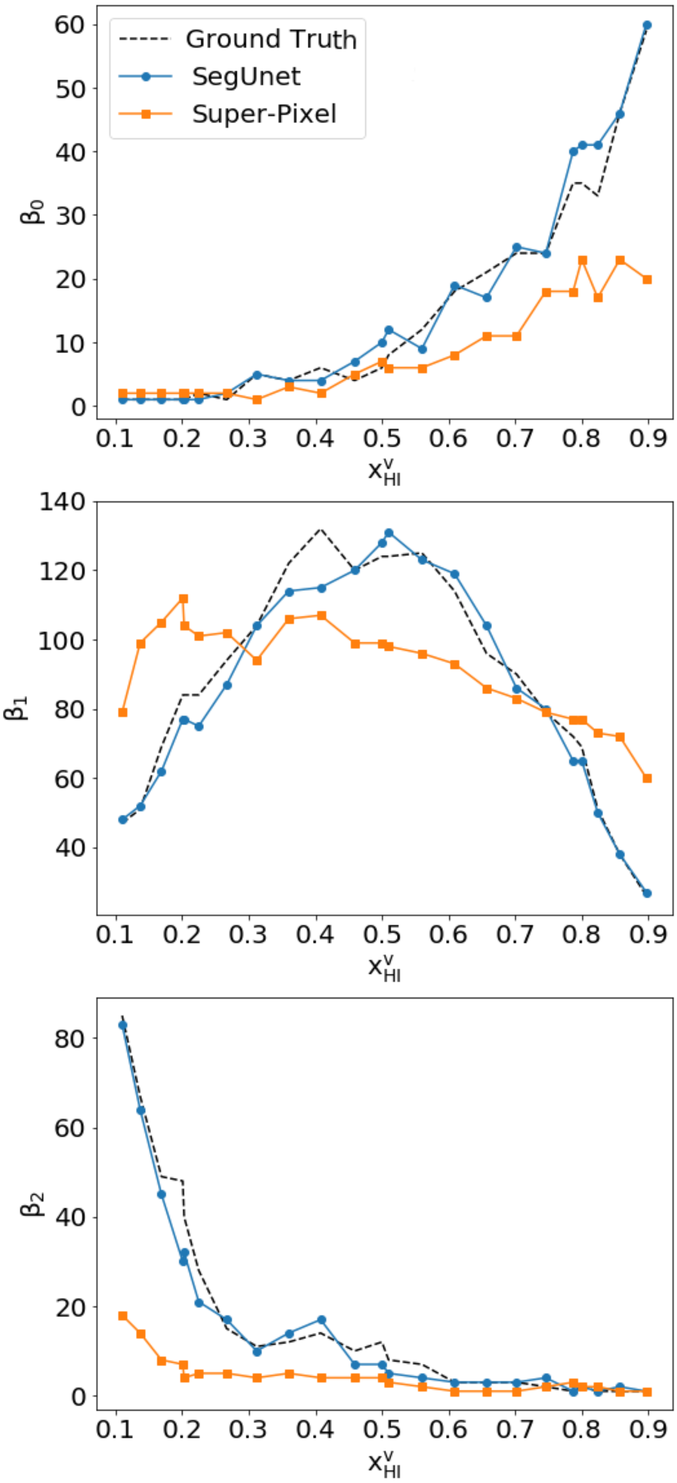

Figure 3.9: Comparison of the topology of the identified regions with Betti numbers estimated from the original neutral field (black dashed line), the `SegU-Net` (blue line with circles) and the Super-Pixel method (orange line with squares), for the case of our 'fiducial' simulation. The top, middle and bottom panels give $\beta_0$, $\beta_1$ and $\beta_2$ respectively. The Betti numbers recovered with `SegU-Net` matches the ground truth better than those recovered with the Super-Pixel method.

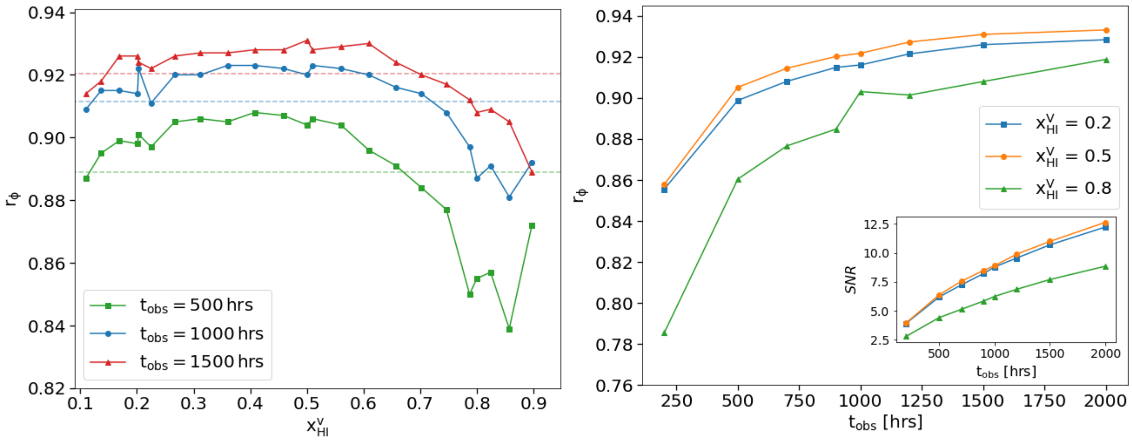

Figure 3.10: *Left panel*: the Matthews correlation coefficient $r_\phi$ of the recovered binary field against its volume-averaged neutral fraction. We compare the prediction set for a high noise level ($t_{\rm obs} = 500\,h$, green line with squares) and low noise level ($t_{\rm obs} = 1500\,h$, red line with triangles) against the noise level employed during the training ($tobs = 1000\,h$, blue line with squares). Horizontal dashed lines of the respective colour represent the MCC average score of the reference simulation. *Right panel*:, the evolution of the MCC score for increasing observation time for a set of mock observations with volume-averaged neutral fraction of $x^{\rm V}_{\rm HI} = 0.2\,(z = 7.310)$, $0.5\,(z = 8.032)$ and $0.8\,(z = 8.720)$, respectively in blue, orange and green color. In the same panel, an inset plot shows the signal-to-noise ratio (SNR $= \sigma_{21}/\sigma_{\rm noise}$) of 21-cm images at a resolution corresponding to a maximum baseline of 2 km we achieve for different observation times.

telescope design is different from our assumptions or simply because the theoretical noise level is not achieved due to complications with the calibration. Therefore, it is important to test to which extent the performance of our network is sensitive to the noise level in the actual data. To change the noise level we choose different observing times, one shorter ($t_{\rm obs}$ = 500 h) and one longer ($t_{\rm obs}$ = 1500 h). The former case corresponds to a noise level $\sqrt{2}$ higher than used in the training set and the latter to a noise level which is $\sqrt{2/3}$ lower. In the left panel of figure 3.10, we show the $r_\phi$ coefficient of the recovered binary field against the volume-averaged neutral fraction $x^{\rm V}_{\rm HI}$. We compare the prediction on the reference simulation for the higher ($t_{\rm obs}$ = 500 h, green line with squares) and lower noise case ($t_{\rm obs}$ = 1500 h, red line with triangles) with the one using the noise level employed during the training and validation process ($t_{\rm obs}$ = 1000 h, blue line with circles). It is evident from the plot that although the noise level does impact the accuracy of the results, we still achieve approximately the same level of precision as in our test case, as commented in § 3.4.2. In fact, the overall average accuracy, indicated with horizontal dashed lines in figure 3.10, on the simulation of reference is 89 per cent for the higher noise case (green dashed line) and slightly better, 92 per cent, for the lower noise case (red dashed line). In both cases, there is a drop in performance down to 88 per cent accuracy during the early stages of reionization $x^{\mathrm{V}}_{\mathrm{HI}} > 0.7$, due to the redshift dependency of the simulated noise.

We also want to test how far we can push our `SegU-Net` trained on data with $t_{\mathrm{obs}} = 1000$ h instrumental noise to identify structures in the presence of a higher or lower noise level. In the right panel of figure 3.10, we plot the $r_\phi$ coefficient at different observation times $t_{\mathrm{obs}}$, for three different stages of reionization in our reference simulation, namely for volume-averaged neutral fractions $x^{\mathrm{V}}_{\mathrm{HI}}$ is 0.2 (blue line with squares), 0.5 (orange line with circles) and 0.8 (green line with triangles), corresponding to redshift $z = 7.310$, 8.032 and 8.720. This plot shows that our network performs well for $t_{\mathrm{obs}} \gtrsim 500$ h, where $r_\phi \gtrsim 0.85$. The spike in the curve for $x^{\mathrm{V}}_{\mathrm{HI}} = 0.8$ at $t_{\mathrm{obs}} = 1000$ h is due to the fact that this is the noise level for which the network was trained.

To put our noise level into perspective, the inset plot in the right panel of figure 3.10 shows the signal-to-noise ratio (SNR) achieved for different observation times. The SNR is defined as $\sigma_{21}/\sigma_{\mathrm{noise}}$ (e.g. Kakiichi et al., 2017), where $\sigma_{21}$ and $\sigma_{\mathrm{noise}}$ are the standard deviations of the 21-cm signal and noise respectively at the resolution corresponding to a maximum baseline of 2 km. From this we conclude that a good performance, with the same accuracy as the 'random' testing set ($r_\phi \gtrsim 0.85$), requires a SNR$\gtrsim 3$.

## 3.6 Tests on a Fully Numerical Simulation

We applied our network to mock $\delta T_b$ cubes calculated with the `C`$^2$`RAY` code, presented in § 3.2.3, with a spatial resolution close to the `21cmFAST` simulations employed in the training process, 2 Mpc. In order to obtain the same level of noise per pixel. A visual comparison of the recovered binary field, similar to the results in § 3.4.1, is shown in figure 3.2. In the left panel, the red/blue colour indicates the network prediction

We go through the same process presented in § 3.4. The $r_\phi$ score with `SegU-Net` is represented by the violet dots with error-bars on the right panel of figure 3.5, from left to right we have redshift $z = 7.96$, 8.06, 8.17, 8.28, 8.40, 8.52 and 8.64 corresponding to a universe with volume-averaged neutral fraction of $x^{\mathrm{V}}_{\mathrm{HI}} = 0.17$, 0.29, 0.42, 0.57, 0.70, 0.81 and 0.90, green squares represent the score obtained with the Super-Pixel method. As we can see, our neural network is performing with similar accuracy as for the prediction set of semi-numerical simulations as discussed in § 3.4.1. For $x^{\mathrm{V}}_{\mathrm{HI}} \approx 0.55$ `SegU-Net` performs slightly better than the Super-Pixel method. The Super-Pixel method shows a drop in accuracy at the late ($x_{\rm HI}^{\rm V} < 0.5$) and early ($x_{\rm HI}^{\rm V} > 0.8$) stages of reionization. We do the same comparison with the recovered volume-averaged neutral fraction $x_{\rm HI}^{\rm V}$, in figure 3.6 right panel, the green error-bar points are the same data as mentioned above. As we can see, also for the `C`$^2$`RAY` simulation, `SegU-Net` recovered quantity resides within the 0.5-$\sigma$ confidence interval (violet dots with error-bars). For the Super-Pixel results, this is true only for $x_{\rm HI}^{\rm V} = 0.57$, 0.70 and 0.81, with approximately the same precision as `SegU-Net` in the case of $x_{\rm HI}^{\rm V} = 0.57$ and slightly better results for $x_{\rm HI}^{\rm V} = 0.70$.

## 3.7 Discussion & Conclusions

This work has developed a convolutional neural network based on the U-Net architecture, which can be used to segment redshifted 21-cm image observations into neutral and ionized regions. We have shown that this application of deep learning provides a fast and stable method that significantly improves the identification of ionized/neutral regions during the epoch of reionization over previously proposed methods. To train our network, we employ a large set of simulated mock observations of the 21-cm signal.

Our image segmentation network, `SegU-Net`, also contains an uncertainty estimator. This uncertainty estimator is a simple but efficient application of the test-time augmented (TTA) technique. With this uncertainty estimator, our network can create a pixel by pixel error map during the prediction process. The pixel by pixel error map can later be used to determine the error in any quantity derived from the segmentation.

Once the network has been trained, the binary field's extraction is swift. In our case for simulations of volume $(256\,\mathrm{Mpc})^3$ and mesh-grid of $128^3$, a run in serial on a `Intel® Core™i7-6500U CPU @ 2.5 GHz` processor and using a 16 Gigabytes of `RAM` takes between 5 to 10 seconds. Including the pixel-error map calculation increases the computing time to approximately 10 minutes. For comparison, the Super-Pixel method typically requires several minutes to extract the binary field, where the actual time depends on the image pixel resolution and the number of Super-Pixels employed. The computational speed of our network thus makes it particularly useful as a component in a Bayesian statistical inference framework to perform EoR parameter estimation using tomographic statistics (e.g. Gazagnes et al., 2021).

We compare the accuracy of our approach with a feature finding method from the field of image processing, the so-called Super-Pixel method, which Giri et al. (2018a) showed to be superior over simple thresholding methods. The results show that our neural network can identify neutral regions in the mock observations at least as well and often much better than the Super-Pixel method. We show a visual comparison and the resulting pixel per pixel error map tested on our 'fiducial' model. This error map gives valuable insight into the parts of the image that are hard to recover and helps us check for over-fitting.

We studied the accuracy of a range of derived quantities from the recovered binary fields, comparing the performance of `SegU-Net` with the Super-Pixel method. These quantities are the volume-averaged ionization fraction — the evolution of which provides the reionization history, the size distribution of the ionized (BSD) and neutral (ISD) regions, the dimensionless power spectra of the recovered binary fields and the three Betti numbers, which quantify the topology of the segmented data sets. For all quantities, we find that the `SegU-Net` results are more accurate than the Super-Pixel results, especially for the early and late stages of reionization, where the Super-Pixel method often struggles to produce accurate results.

Machine learning methods generally are sensitive to the properties of the training set. Therefore, we tested `SegU-Net` on input data with different properties than the training set. First, we analysed the performance on data sets with different noise levels than the network was trained. We found that `SegU-Net` yields accurate results for data sets in which the noise level is characterised by an observing time of $t_{\mathrm{obs}} > 500\,h$, which approximately corresponds to an SNR $\gtrsim 3$. Second, we applied the network to mock observations calculated from the results of a fully numerical reionization simulation, rather than the semi-numerical simulations used to train the network. We find that `SegU-Net` achieves the same level of accuracy when applied to this data set and therefore is not sensitive to the type of simulation employed during the training process.

We want to point out that similar efforts are being made by Gagnon-Hartman et al. (2021). They focus on reconstructing the segmented maps of ionized and neutral regions in the context of foreground mitigation using the foreground avoidance method (e.g. Liu & Tegmark, 2011; ?), and also consider the possibility of doing so with instruments that are not optimised for imaging such as HERA. We include the effect of instrumental noise and study in great detail the summary statistics of the reconstructed binary maps and the dependency of the results on the noise level. In the future, we will extend our study to include the impact of foreground mitigation strategies while recovering the summary statistics.

Here we assumed the spin temperature to be saturated ($T_{\mathrm{S}} \gg T_{\mathrm{CMB}}$). However, it is possible to have such a scenario where this assumption fails, especially during the time when reionization starts. In the future, we will evolve our `SegU-Net` to identify

H II regions in such scenarios. Even though our network is built to identify H II regions, U-Net architecture can be trained to identify any interesting feature. Before reionization started, the luminous sources heated the IGM and left its impact on the 21-cm signal (e.g. Ross et al., 2017, 2019). The U-Net architecture can also be trained to identify these heated regions.

## Acknowledgements

The authors would like to thank Leon Koopmans for useful discussions and comments. We also acknowledge helpful discussion with Adrian Liu and collaborators. MB is supported by PhD Studentship from the Science and Technology Facilities Council (STFC) and appreciates the Oskar Klein Center at Stockholm University for hospitality during the completion of this work. This work was possible thanks to the STFC Long Term Attachment (LTA) travel grant. ITI is supported by the Science and Technology Facilities Council (grant numbers ST/I000976/1 and ST/T000473/1) and the Southeast Physics Network (SEPNet). GM is supported in part by the Swedish Research Council grant 2020-04691. We acknowledge PRACE for awarding us access to the KAY facility hosted by the Irish Centre for High-end Computing (ICHEC) and the GALILEO hosted by the Super Computer Application and Innovation (SCAI) in collaboration with the CINECA consortium. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for partly funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCSSupercomputer JUWELS at Juelich Supercomputing Centre (JSC). The deep learning implementation was possible thanks to the application programming interface of `Tensorflow` (Abadi et al., 2015) and `Keras` (Chollet et al., 2017). The algorithms and image processing tools operated on our data were performed with the help of `NumPy` (Harris et al., 2020), `SciPy` (Virtanen et al., 2020) and `scikit-image` packages (van der Walt et al., 2014). All figures were created with `mathplotlib` (Hunter, 2007).

## Data Availability

The data underlying this article is available upon request, and can also be re-generated from scratch using the publicly available `21cmFAST`, CUBEP$^3$M, C$^2$RAY and `Tools21cm` code. The `SegU-Net` code and its trained network weights are available on the author's `GitHub` page: `https://github.com/micbia/SegU-Net`.

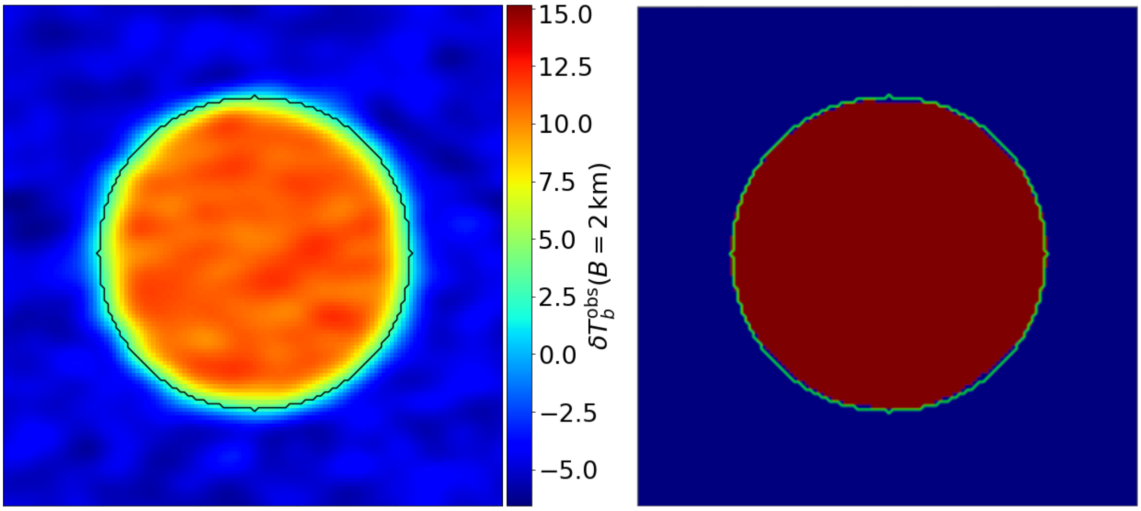

Figure 3.11: Test SegU-Net on a spherical ionized region. *Left panel*: slice through the input image. The colour map shows the differential brightness temperature, and the black contour shows the boundary between neutral and ionized regions. *Right panel*: the recovered binary field with SegU-Net. The green contour represents the same boundary again. The identified neutral and ionized regions are indicated in red and blue, respectively.

## 3.8 Appendix

### 3.8.1 SegU-Net Hidden Layer Outputs

We test SegU-Net to see if it can recover the binary field for a simple case, namely a single spherical neutral region. We assume a uniform density field at $z = 8.032$ and calculate the differential brightness temperature with section 3.2.2, adding noise corresponding to $t_{\rm obs} = 1000\,{\rm h}$ and reducing the resolution to correspond to a maximum baseline of $B = 2\,{\rm km}$. In figure 3.11, we show the input image of the sphere (left panel) and the corresponding recovered binary field by SegU-Net (right panel). The black contour in the left panel and the green contour in the right panel show the true boundary of the sphere. For this test SegU-Net achieves an accuracy of 98 per cent.

In figure 3.12 we show the output of the bottom hidden layer of SegU-Net, which is the last layer of the left part of the U-shaped in figure 3.3. The colour coding is such that blue correspond to negative, red to positive and white to zero values. This output gives a visual representation of the low dimensional latent space of our network encoder. In our case, this consists of 256 images, where each corresponds to a convolutional kernel and contains information about the image's extracted features. The encoder path contracts the input image from an original mesh size of $128^2$ down to $8^2$. The information contained in the latent space is then expanded by SegU-Net decoder into a binary map of the same

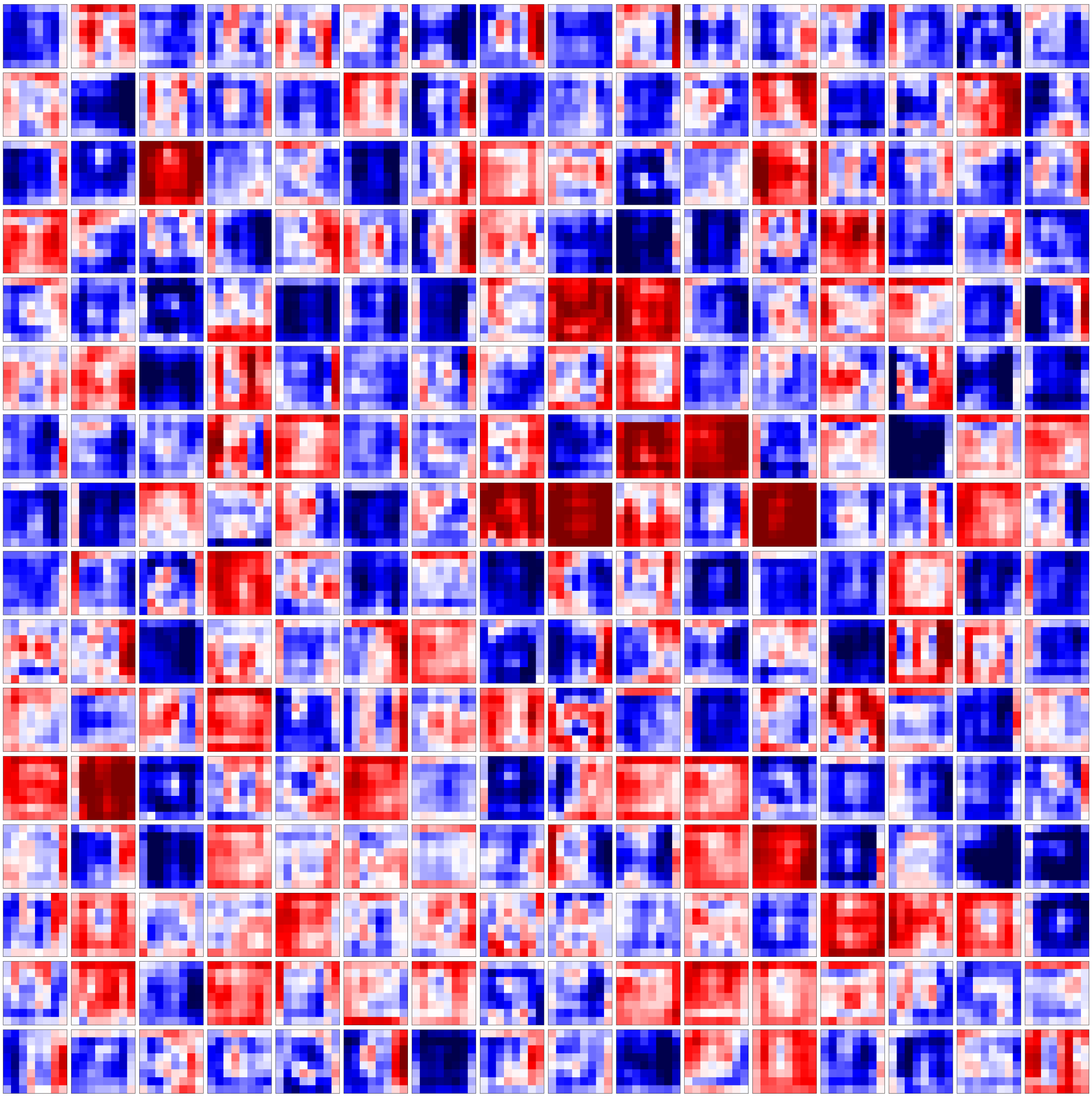

Figure 3.12: Visual representation of `SegU-Net`'s low dimensional latent space (bottom layer), which contains information about the extracted features of our test input image.

size as the input image (see right panel of figure 3.11).

### 3.8.2 Skip Connection Between Encoder and Decoder Levels

The main advantage of a U-shaped network is that it overcomes the bottleneck limitation encountered by auto-encoder networks (a classical encoder/decoder architecture) by adding interconnections between the descending (encoder) and ascending (decoder) paths (Long et al., 2014; Ronneberger et al., 2015). These connections allow feature representations to pass through the bottleneck (bottom layer) and avoid loss of information due to contraction.

In figure 3.13, we show a visual example of interconnections between the encoder (left

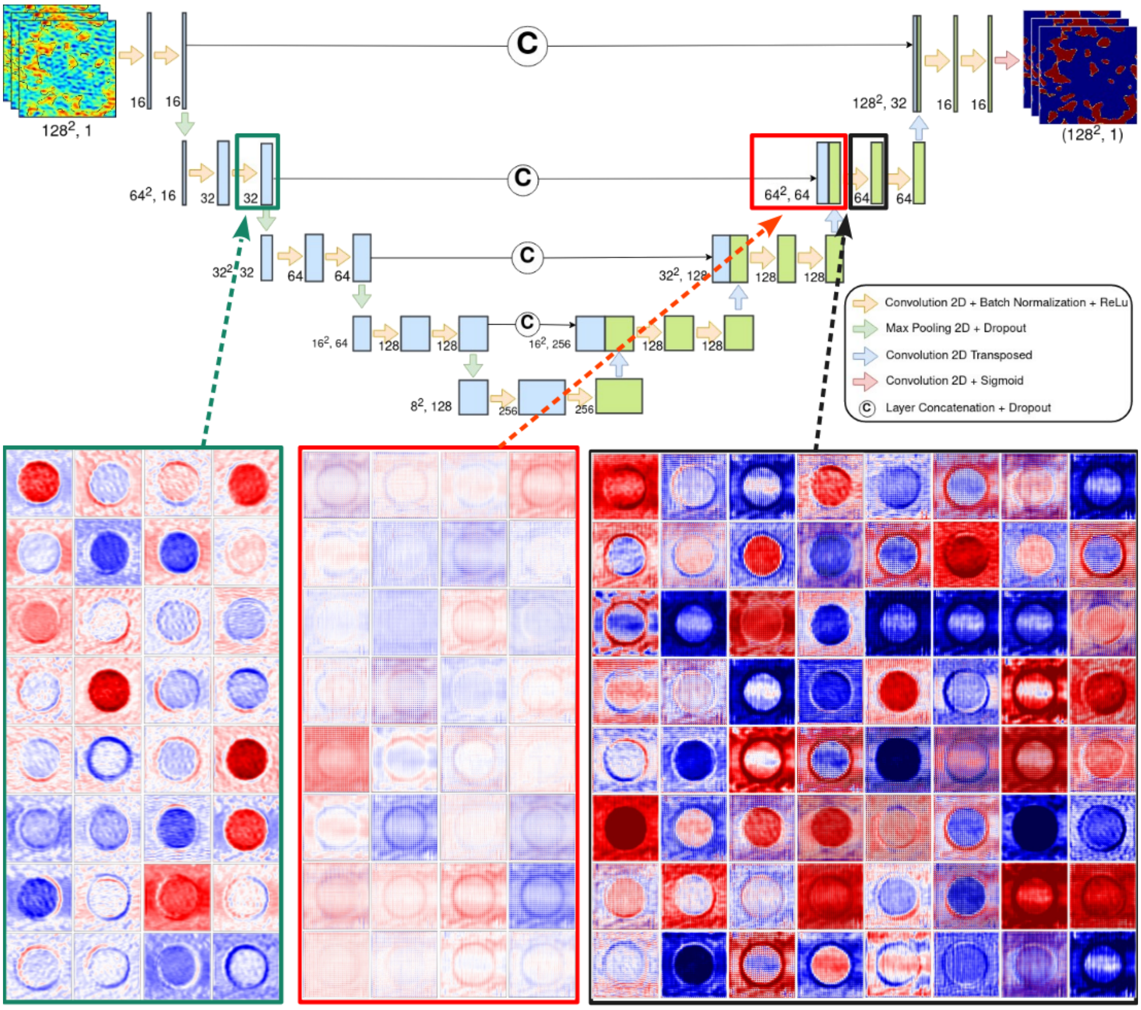

Figure 3.13: Example of skip connection between encoder and decoder levels. The top panel shows the architecture of our network. The bottom panels display the output of three hidden layers. On the left (green dashed line), a convolutional block (`ConvBlock`) output is interconnected with the output of the second to last up-sampling operation (central panel, red dashed line). The right-most panel shows the result of the merge after a convolution block (black dashed line).

part of the U-shape) and the decoder (right part). The top panel shows a schematic representation of our network architecture, and the bottom part displays a visual output of three hidden layers for the test case of a sphere. The left-most panel (connected by a green dashed line) shows the output of the second convolutional block in the encoder's second level. This block consists of 32 kernels with a mesh size of $64^2$. At this level, the shape and form of the input image are still visible. The centre panel (connected by a red dashed line) shows the result of the second to last up-sampling operation of the decoder. The number of kernels and mesh size match with the corresponding encoder layers. The skip connection merges the encoder and decoder-level output for a total of 64 images with mesh $64^2$. The right-most panel (connected with a black dashed line) shows the concatenation after a convolutional block. The effect of the up-sampling operation is still visible.

# Chapter 4

# Thesis Conclusion

This thesis focused on improving state of the art for numerical simulations employed for the Epoch of Reionisation. To our knowledge, we provided the most accurate treatment of the local atomic recombination in the largest reionisation simulation to date.

We then direct our attention to improve the methodical approach employed for interpreting results in numerical simulations that could be employed for future and ongoing 21-cm observations. With the application of advanced deep learning methods, we obtained the most accurate approach to date.

## 4.1 Summary of Results

In chapter 2, we presented an empirical stochastic model for the sub-grid gas clumping. With this project, we continued the discussion started by our collaborator Mao et al. (2019), which presented a local density-dependent deterministic model for the sub-grid gas clumping. Here, we based our model on the results from a high-resolution numerical simulation, which fully resolve all relevant fluctuations. Our model reproduces well both the mean density-clumping relation and its scatter. Then, we applied our stochastic model and created a large-volume realisation (714 Mpc along each side) of the clumping field. For comparison, we reproduced the same field with a mean clumping factor model and the deterministic model. Finally, we used these in radiative transfer simulations of cosmic reionisation and compared our empirical approach with a mean clumping factor model and the deterministic model.

Our results show that the simplistic mean clumping model delays reionisation compared to local density-dependent models, despite producing fewer recombinations overall. This trend is due to the very different spatial distribution of clumping, resulting in much higher photo-ionisation rates in the latter cases. Moreover, the density-dependent model indirectly affects the lifetime of low mass atomically cooling halos. This is because the high-density regions are more resilient to getting ionised due to the enhanced recombination rate. In summary, the primary conclusion of this first part of the thesis is that density-dependent models of the clumping factor are preferable to oversimplified treatment, as sub-grid recombinations play a crucial role in large scale reionisation simulations.

The main focus of the project presented in chapter 3 was to provide a stable a reliable method for future neutral/ionised region identification in tomographic images from SKA1-Low. To do so, we developed a deep learning network, named `SegU-Net`, trained on a vast set of simulated tomographic images, produced with the semi-numerical code `21cmFAST`.

The results show that our network is capable of segmenting simulated 21-cm image data into meaningful features (ionised and neutral regions) with greater accuracy compared to previous methods. We can estimate the ionisation history from noisy mock observations of SKA with an observation time of 1000 hours with more than 87 per cent accuracy. The size distributions and Betti numbers of the recovered field have a relative difference of only a few per cent from the values derived from the original smoothed and then binarised neutral fraction field. Moreover, its stability is demonstrated by successfully recover 21-cm signal for different levels of instrumental noise. Perhaps the most interesting result is that the deep learning approach obtains the same level of accuracy for images with a lower signal-to-noise ratio than the one used for the training.

In conclusion, we find that the deep learning approach is independent of the simulations code used while training the network. Instead, it learns the simulation independent pattern of the ionised regions by comparing the gradient in the 21-cm signal intensities to identify these regions.

# Bibliography

Abadi M., et al., 2015, TensorFlow, `http://tensorflow.org/` 110

Abbott T. M. C., et al., 2021, Dark Energy Survey Year 3 Results: Cosmological Constraints from Galaxy Clustering and Weak Lensing (`arXiv:2105.13549`) 4, 29

Abel T., Norman M. L., Madau P., 1999, The Astrophysical Journal, 523, 66 33

Abel T., Bryan G. L., Norman M. L., 2000, The Astrophysical Journal, 540, 39 9

Abel T., Bryan G. L., Norman M. L., 2001, Science, Volume 295, Issue 5552, pp. 93-98 (2002)., 295, 93 9

Achanta R., Shaji A., Smith K., Lucchi A., Fua P., Süsstrunk S., 2012, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34, 2274 96

Ade P. A. R., et al., 2014, Astronomy & Astrophysics, 571, A16 2, 63

Aghanim N., et al., 2020, Astronomy & Astrophysics, 641, A6 4, 12

Ahn K., Shapiro P. R., Iliev I. T., Mellema G., Pen U.-L., 2009, The Astrophysical Journal, 695, 1430 50

Ahn K., Iliev I. T., Shapiro P. R., Srisawat C., 2015a, Monthly Notices of the Royal Astronomical Society, 450, 1486 32, 50

Ahn K., Iliev I. T., Shapiro P. R., Srisawat C., 2015b, Monthly Notices of the Royal Astronomical Society, 450, 1486 91

Aubert D., Teyssier R., 2008, Monthly Notices of the Royal Astronomical Society, 387, 295 33

Aurélien G., 2019, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, paperback edn. O'Reilly Media 37

Avelino P. P., Barbosa D., 2004, Physical Review D, 70, 067302 11

Barkana R., 2018, Nature, 555, 71 18

Bartelmann M., 2015, Physical Review D, 91 6

Becker G. D., Bolton J. S., Madau P., Pettini M., Ryan-Weber E. V., Venemans B. P., 2014, Monthly Notices of the Royal Astronomical Society, Volume 447, Issue 4, p.3402-3419, 447, 3402 14

Becker G. D., Bolton J. S., Lidz A., 2015, Publications of the Astronomical Society of Australia, 32 xii, 15

Becker G. D., D'Aloisio A., Christenson H. M., Zhu Y., Worseck G., Bolton J. S., 2021, arXiv e-prints, p. arXiv:2103.16610 35

Behroozi P. S., Wechsler R. H., Wu H.-Y., 2012, The Astrophysical Journal, 762, 109 31

Bennett C. L., et al., 2013, The Astrophysical Journal Supplement Series, 208, 20 4

Benson A. J., 2010, Physics Reports, 495, 33–86 6

Betti E., 1870, Annali di Matematica Pura ed Applicata (1867-1897), 4, 140 103

Bianco M., Giri S. K., Iliev I. T., Mellema G., 2021a, Monthly Notices of the Royal Astronomical Society 84

Bianco M., Iliev I. T., Ahn K., Giri S. K., Mao Y., Park H., Shapiro P. R., 2021b, Monthly Notices of the Royal Astronomical Society, 504, 2443 24, 32, 47, 91

Binney J., Tremaine S., 1987, Galactic dynamics xi, 7

Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, The Astrophysical Journal, 379, 440 34

Bouwens R. J., et al., 2014, The Astrophysical Journal, 795, 126 8

Bouwens R., et al., 2016, The Astrophysical Journal, Volume 833, Issue 1, article id. 72, 32 pp. (2016)., 833 9

Bowman J. D., Rogers A. E. E., Monsalve R. A., Mozdzen T. J., Mahesh N., 2018, Nature, 555, 67–70 18

Braun R., Bonaldi A., Bourke T., Keane E., Wagg J., 2019, Technical report, Anticipated Performance of the Square Kilometre Array – Phase 1 (SKA1). (arXiv:1912.12699) 22

Bromm V., Coppi P. S., Larson R. B., 1999, The Astrophysical Journal, 527, L5–L8 9

Bromm V., Coppi P. S., Larson R. B., 2002, The Astrophysical Journal, 564, 23 9

Cain C., D'Aloisio A., Gangolli N., Becker G. D., 2021, arXiv e-prints, p. arXiv:2105.10511 35

Cen R., 2003, The Astrophysical Journal, 591, 12 48

Chandra K., et al., 2019, Gradient Descent: The Ultimate Optimizer (arXiv:1909.13371) 40

Chardin J., Uhlrich G., Aubert D., Deparis N., Gillet N., Ocvirk P., Lewis J., 2019, Monthly Notices of the Royal Astronomical Society, 490, 1055–1065 85

Charnock T., Perreault-Levasseur L., Lanusse F., 2020, Bayesian Neural Networks (arXiv:2006.01490) 95

Chen X.-L., Miralda-Escude J., 2008, The Astronomical Journal, 684, 18 11

Chen H.-W., et al., 2020a, Monthly Notices of the Royal Astronomical Society, 497, 498 35

Chen B. H., et al., 2020b, Monthly Notices of the Royal Astronomical Society, 501, 3951–3961 85

Chollet F., Allaire J., et al., 2017, Keras, https://github.com/rstudio/keras 110

Choudhury T. R., 2009, Current Science, 97, 841 25

Choudhury T. R., Ferrara A., 2006, arXiv e-prints, pp astro–ph/0603149 11, 14

Christlein V., Spranger L., Seuret M., Nicolaou A., Král P., Maier A., 2019, Deep Generalized Max Pooling (arXiv:1908.05040) 42

Ciardi B., Ferrara A., 1997, The Astrophysical Journal, 483, L5 28, 48

Ciardi B., Ferrara A., 2005, Space Sci. Rev., 116, 625 23

Ciardi B., Ferrara A., Marri S., Raimondo G., 2001, Monthly Notices of the Royal Astronomical Society, 324, 381 33

Ciardi B., Bolton J. S., Maselli A., Graziani L., 2012, Monthly Notices of the Royal Astronomical Society, 423, 558 33

Cohen A., Fialkov A., Barkana R., Monsalve R. A., 2020, Monthly Notices of the Royal Astronomical Society, 495, 4845–4859 85

Colless M., et al., 2001, Monthly Notices of the Royal Astronomical Society, 328, 1039 4

Crocce M., Pueblas S., Scoccimarro R., 2006, Monthly Notices of the Royal Astronomical Society, 373, 369 50

Cui Y., Jia M., Lin T., Song Y., Belongie S. J., 2019 (arXiv:1901.05555) 95

Datta K. K., Bharadwaj S., Choudhury T. R., 2007, Monthly Notices of the Royal Astronomical Society, 382, 809 85

Davies F. B., et al., 2018, The Astrophysical Journal, 864, 142 xvi, 14, 61

DeBoer D. R., et al., 2017, Publications of the Astronomical Society of the Pacific, 129, 045001 19

Delyon B., 2000, Stochastic approximation with decreasing gain: Convergence and asymptotic theory 40

Deng J., Dong W., Socher R., jia Li L., Li K., Fei-fei L., 2009, in In CVPR. 35

Dixon K. L., Iliev I. T., Mellema G., Ahn K., Shapiro P. R., 2016, Monthly Notices of the Royal Astronomical Society, 456, 3011 31, 32, 33, 51

Dodelson S., ed. 2003. Academic Press, Burlington, doi:https://doi.org/10.1016/B978-012219141-1/50019-X 3

Dumoulin V., Visin F., 2016 (arXiv:1603.07285) 94

Einstein A., 1916, Annalen der Physik, 354, 769 2

Elbers W., van de Weygaert R., 2019, Monthly Notices of the Royal Astronomical Society, 486, 1523 20, 103

Ewall-Wice A., Chang T.-C., Lazio J., Doré O., Seiffert M., Monsalve R. A., 2018, The Astrophysical Journal, 868, 63 18

Fan X., et al., 2006a, The Astrophysical Journal, 131, 1203 14, 28

Fan X., et al., 2006b, The Astronomical Journal, 132, 117–136 14

Feng C., Holder G., 2018, The Astrophysical Journal, 858, L17 18

Ferland G., Korista K., Verner D., Ferguson J., Kingdon J., Verner E., 1998, Publications of the Astronomical Society of the Pacific, 110, 761 7

Ferrara A., Loeb A., 2013, Monthly Notices of the Royal Astronomical Society, 431, 2826 26

Ferrara A., Pandolfi S., 2014, Proc. Int. Sch. Phys. Fermi, 186, 1 14, 23

Fialkov A., Barkana R., Cohen A., 2018, Phys. Rev. Lett., 121, 011101 18

Field G. B., 1958, Proceedings of the IRE, 46, 240 16

Field G. B., 1959, The Astrophysical Journal, 129, 536 16

Finkelstein S. L., et al., 2015, The Astrophysical Journal, 810, 71 9

Finkelstein S. L., et al., 2019, The Astrophysical Journal, 879 8, 9

Fraser S., et al., 2018, Physics Letters B, 785, 159 18

Friedmann A., 1922, Zeitschrift fur Physik, 10, 377 3

Friedrich M. M., Mellema G., Alvarez M. A., Shapiro P. R., Iliev I. T., 2011, Monthly Notices of the Royal Astronomical Society, 413, 1353 20, 69, 103

Friedrich M. M., Mellema G., Iliev I. T., Shapiro P. R., 2012, Monthly Notices of the Royal Astronomical Society, 421, 2232–2250 33

Fukushima K., 1980, Biological Cybernetics, 36, 193 35

Furlanetto S. R., 2006a, Monthly Notices of the Royal Astronomical Society, 371, 867 25

Furlanetto S. R., 2006b, Monthly Notices of the Royal Astronomical Society, 371, 867 87

Furlanetto S. R., Oh S. P., 2006, The Astrophysical Journal, 652, 849 72

Furlanetto S. R., Oh S. P., 2016a, Monthly Notices of the Royal Astronomical Society, 28, 303 71

Furlanetto S. R., Oh S. P., 2016b, Monthly Notices of the Royal Astronomical Society, 457, 1813 103

Furlanetto S. R., Zaldarriaga M., Hernquist L., 2004, The Astrophysical Journal, 613, 1 21, 34, 69, 85, 86, 87

Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Physics Reports, 433, 181 15, 16, 102

Gagnon-Hartman S., Cui Y., Liu A., Ravanbakhsh S., 2021, Monthly Notices of the Royal Astronomical Society, 504, 4716 109

Gazagnes S., Koopmans L. V. E., Wilkinson M. H. F., 2021, Monthly Notices of the Royal Astronomical Society, 502, 1816 97, 108

Ghara R., Choudhury T. R., Datta K. K., 2016, Monthly Notices of the Royal Astronomical Society, 460, 827 22

Ghara R., Choudhury T. R., Datta K. K., Choudhuri S., 2017, Monthly Notices of the Royal Astronomical Society, 464, 2234 22, 85, 88

Ghara R., Mellema G., Giri S. K., Choudhury T. R., Datta K. K., Majumdar S., 2018, Monthly Notices of the Royal Astronomical Society, 476, 1741 33

Ghara R., et al., 2020, Monthly Notices of the Royal Astronomical Society, 493, 4728 19, 85

Gholamalinezhad H., Khosravi H., 2020, Pooling Methods in Deep Neural Networks, a Review (`arXiv:2009.07485`) 42

Gillet N., Mesinger A., Greig B., Liu A., Ucci G., 2019, Monthly Notices of the Royal Astronomical Society xiv, 43, 44, 85, 93

Giri S. K., 2019, PhD thesis, Stockholm University, Department of Astronomy 85

Giri S. K., Mellema G., 2021, Monthly Notices of the Royal Astronomical Society, 505, 1863 20, 21, 103

Giri S. K., Mellema G., Dixon K. L., Iliev I. T., 2018a, Monthly Notices of the Royal Astronomical Society, 473, 2949 69, 71, 73, 85, 88, 96, 97, 99, 101, 108

Giri S. K., Mellema G., Ghara R., 2018b, Monthly Notices of the Royal Astronomical Society, 479, 5596 76

Giri S. K., Mellema G., Aldheimer T., Dixon K. L., Iliev I. T., 2019a, Monthly Notices of the Royal Astronomical Society, 489, 1590 85, 101

Giri S. K., Zackrisson E., Binggeli C., Pelckmans K., Cubo R., 2019b, Monthly Notices of the Royal Astronomical Society, 491, 5277–5286 85

Giri S. K., D'Aloisio A., Mellema G., Komatsu E., Ghara R., Majumdar S., 2019c, Journal of Cosmology and Astroparticle Physics, 2019, 058 19, 20, 76

Giri S. K., Mellema G., Jensen H., 2020, Journal of Open Source Software, 5, 2363 70, 88

Glorot X., Bordes A., Bengio Y., 2011, in Gordon G., Dunson D., Dudík M., eds, Proceedings of Machine Learning Research Vol. 15, Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, Fort Lauderdale, FL, USA, pp 315–323, `http://proceedings.mlr.press/v15/glorot11a.html` 37, 94

Gnedin N. Y., Abel T., 2001, New Astronomy, 6, 437 33

Gnedin N. Y., Ostriker J. P., 1997, The Astrophysical Journal, 486, 581–598 26, 33, 48

Gnedin N. Y., Kravtsov A. V., Chen H.-W., 2008, The Astrophysical Journal, 672, 765 26

Gong D., Zhang Z., Shi Q., van den Hengel A., Shen C., Zhang Y., 2020, Learning Deep Gradient Descent Optimization for Image Deconvolution (`arXiv:1804.03368`) 40

Greif T. H., Glover S. C. O., Bromm V., Klessen R. S., 2010, The Astrophysical Journal, 716, 510 10

Greig B., Mesinger A., Haiman Z., Simcoe R. A., 2016, Monthly Notices of the Royal Astronomical Society, 466, 4239 14

Greig B., Mesinger A., Bañados E., 2019, Monthly Notices of the Royal Astronomical Society, 484, 5094 14

Greig B., Trott C. M., Barry N., Mutch S. J., Pindor B., Webster R. L., Wyithe J. S. B., 2020a, Monthly Notices of the Royal Astronomical Society, 500, 5322–5335 19

Greig B., et al., 2020b, Monthly Notices of the Royal Astronomical Society, 501, 1–13 19

Gunn J. E., Peterson B. A., 1965, The Astrophysical Journal, 142, 1633 14

Guzman E., Meyers J., 2021 (`arXiv:2101.01214`) 85

Hahn O., Abel T., 2011, Monthly Notices of the Royal Astronomical Society, 415, 2101 29

Haiman Z., Holder G. P., 2003, The Astrophysical Journal, 595, 1 25

Haiman Z., Thoul A. A., Loeb A., 1996, The Astrophysical Journal, 464, 523 9

Hansen S. H., Haiman Z., 2004, The Astronomical Journal, 600, 26 11

Harnois-Déraps J., Pen U. L., Iliev I. T., Merz H., Emberson J. D., Desjacques V., 2013, Monthly Notices of the Royal Astronomical Society, 436, 540 31, 49, 50, 82, 91

Harris C. R., et al., 2020, Nature, 585, 357–362 110

Hartle J. B., 2003, American Journal of Physics, 71, 1086 2

Hassan S., Andrianomena S., Doughty C., 2020, Monthly Notices of the Royal Astronomical Society, 494, 5761–5774 85, 93

He K., Zhang X., Ren S., Sun J., 2015, Deep Residual Learning for Image Recognition (`arXiv:1512.03385`) 42

Heath D. J., 1977, Monthly Notices of the Royal Astronomical Society, 179, 351 5

Hinshaw G., et al., 2013, The Astrophysical Journal Supplement Series, 208, 19 29

Hinton G. E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R. R., 2012 (`arXiv:1207.0580`) 94

Hochreiter S., Schmidhuber J., 1997, Neural Computation, 9, 1735 35

Hopfield J. J., 1982, Proceedings of the National Academy of Sciences, 79, 2554 35

Hortúa H. J., Volpi R., Malagò L., 2020 (`arXiv:2005.02299`) 95

Hu W. T., 1995, Other thesis (`arXiv:astro-ph/9508126`) xi, 12, 13

Huang G., Liu Z., van der Maaten L., Weinberger K. Q., 2018, Densely Connected Convolutional Networks (`arXiv:1608.06993`) 42

Hubble E., 1929, Proceedings of the National Academy of Science, 15, 168 1

Hubble E., Humason M. L., 1931, The Astrophysical Journal, 74, 43 1

Hui L., Gnedin N. Y., 1997, Monthly Notices of the Royal Astronomical Society, 292, 27–42 23

Hunter J. D., 2007, Computing in Science & Engineering, 9, 90 110

Ichikawa K., Barkana R., Iliev I. T., Mellema G., Shapiro P. R., 2010a, Monthly Notices of the Royal Astronomical Society, 406, 2521 19

Ichikawa K., Barkana R., Iliev I. T., Mellema G., Shapiro P. R., 2010b, Monthly Notices of the Royal Astronomical Society, 406, 2521 76

Iliev I. T., Scannapieco E., Shapiro P. R., 2005, The Astrophysical Journal, 624, 491–504 48

Iliev I. T., Mellema G., Pen U. L., Merz H., Shapiro P. R., Alvarez M. A., 2006, Monthly Notices of the Royal Astronomical Society, 369, 1625 28, 33, 69, 71

Iliev I. T., Pen U., Bond J. R., Mellema G., Shapiro P. R., 2007, The Astrophysical Journal, 660, 933 48, 52, 54, 60

Iliev I. T., et al., 2009, Monthly Notices of the Royal Astronomical Society, 400, 1283 28, 33

Iliev I. T., Mellema G., Shapiro P. R., Pen U.-L., Mao Y., Koda J., Ahn K., 2012, Monthly Notices of the Royal Astronomical Society, 423, 2222–2253 60

Iliev I. T., Mellema G., Ahn K., Shapiro P. R., Mao Y., Pen U. L., 2014, Monthly Notices of the Royal Astronomical Society, 439, 725 21, 28, 30, 67, 70, 71

Ioffe S., Szegedy C., 2015 (`arXiv:1502.03167`) 94

Ishigaki M., Kawamata R., Ouchi M., Oguri M., Shimasaku K., Ono Y., 2018, , 854, 73 9

Jarrett K., Kavukcuoglu K., Ranzato M., LeCun Y., 2009, in 2009 IEEE 12th International Conference on Computer Vision. pp 2146–2153, doi:10.1109/ICCV.2009.5459469 37, 94

Jeeson-Daniel A., Ciardi B., Graziani L., 2014, Monthly Notices of the Royal Astronomical Society, 11, 1 52

Jeffrey N., Lanusse F., Lahav O., Starck J.-L., 2020, Monthly Notices of the Royal Astronomical Society, 492, 5023–5029 85

Jenkins A., Frenk C. S., White S. D. M., Colberg J. M., Cole S., Evrard A. E., Couchman H. M. P., Yoshida N., 2001, Monthly Notices of the Royal Astronomical Society, 321, 372 26

Jennings W. D., Watkinson C. A., Abdalla F. B., McEwen J. D., 2018, Monthly Notices of the Royal Astronomical Society, 483, 2907–2922 85

Kakiichi K., et al., 2017, Monthly Notices of the Royal Astronomical Society, 471, 1936 97, 101, 107

Kapahtia A., Chingangbam P., Ghara R., Appleby S., Choudhury T. R., 2021 (arXiv:2101.03962) 20, 103

Kaurov A. A., Gnedin N. Y., 2015, The Astrophysical Journal, 810, 154 26

Kern N. S., Liu A., Parsons A. R., Mesinger A., Greig B., 2017, The Astrophysical Journal, 848, 23 85

Kingma D. P., Ba J., 2014 (arXiv:1412.6980) 95

Knebe A., et al., 2011, Monthly Notices of the Royal Astronomical Society, 415, 2293 31

Knollmann S. R., Knebe A., 2009, The Astrophysical Journal Supplement Series, 182, 608–624 31

Kohler K., Gnedin N. Y., Hamilton A. J. S., 2007, The Astrophysical Journal, 657, 15 48, 54

Komatsu E., et al., 2009, The Astrophysical Journals, 180, 330 4, 87

Komatsu E., et al., 2011b, The Astrophysical Journal Supplement Series, 192, 18 50

Komatsu E., et al., 2011a, The Astrophysical Journal Supplement Series, 192, 18 12

Konno A., et al., 2014, The Astrophysical Journal, 797, 16 14

Koopmans L. V. E., et al., 2015, Proceedings of Science, AASKA14, 001 19, 85

Krizhevsky A., Sutskever I., Hinton G. E., 2012, in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. NIPS'12. Curran Associates Inc., Red Hook, NY, USA, p. 1097–1105 40

Lacey C., Cole S., 1994, Monthly Notices of the Royal Astronomical Society, 271, 676 34

Le Cun Y., Bottou L., Bengio Y., 1997, Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, 1, 151 35, 40

LeCun Y., Cortes C., 2010, BibSonomy 35

Lee C., 2019, Planetary and Space Science, 170, 16–28 85

Lewis A., Challinor A., Lasenby A., 2000, The Astrophysical Journal, 2, 2 29, 50

Li W., et al., 2019, Monthly Notices of the Royal Astronomical Society, 485, 2628 85

Lidz A., Zahn O., McQuinn M., Zaldarriaga M., Dutta S., Hernquist L., 2007, The Astrophysical Journal, 659, 865 102

Lin Y., Oh S. P., Furlanetto S. R., Sutter P. M., 2016a, Monthly Notices of the Royal Astronomical Society, 461, 3361 69

Lin Y., Oh S. P., Furlanetto S. R., Sutter P. M., 2016b, Monthly Notices of the Royal Astronomical Society, 461, 3361 101

Linnainmaa S., 1976, BIT, 16, 146 35

Liu A., Tegmark M., 2011, Physical Review D, 83, 103006 109

Long J., Shelhamer E., Darrell T., 2014 (`arXiv:1411.4038`) xv, 40, 44, 45, 85, 112

Lupi A., Bovino S., Capelo P. R., Volonteri M., Silk J., 2018, Monthly Notices of the Royal Astronomical Society, 474, 2884 55

Maas A. L., Hannun A. Y., Ng A. Y., 2013, in in ICML Workshop on Deep Learning for Audio, Speech and Language Processing. 37

Machacek M. E., Bryan G. L., Abel T., 2001, The Astrophysical Journal, 548, 509 10

Madau P., Dickinson M., 2014, Annual Review of Astronomy and Astrophysics, 52, 415–486 8, 9

Madau P., Meiksin A., Rees M. J., 1997, The Astrophysical Journal, 475, 429 15

Madau P., Haardt F., Rees M. J., 1999, The Astrophysical Journal, 514, 648 48, 54

Maio U., Ciardi B., Dolag K., Tornatore L., Khochfar S., 2010, Monthly Notices of the Royal Astronomical Society, 407, 1003–1015 10

Majumdar S., Pritchard J. R., Mondal R., Watkinson C. A., Bharadwaj S., Mellema G., 2018, Monthly Notices of the Royal Astronomical Society, 476, 4007 19

Makinen T. L., Lancaster L., Villaescusa-Navarro F., Melchior P., Ho S., Perreault-Levasseur L., Spergel D. N., 2021, Journal of Cosmology and Astroparticle Physics, 04, 081 85

Mangena T., Hassan S., Santos M. G., 2020, Monthly Notices of the Royal Astronomical Society, 494, 600–606 85

Mao Y., Koda J., Shapiro P. R., Iliev I. T., Mellema G., Park H., Ahn K., Bianco M., 2019, Monthly Notices of the Royal Astronomical Society, 491, 1600–1621 24, 32, 47, 48, 50, 51, 52, 54, 78, 82, 91, 115

Mapelli M., Ferrara A., Pierpaoli E., 2006, Monthly Notices of the Royal Astronomical Society, 369, 1719 11

Mason C. A., Gronke M., 2020, Monthly Notices of the Royal Astronomical Society, 499, 1395 85

Mather J. C., et al., 1994, , 420, 439 5

Matsuda T., Satō H., Takeda H., 1969, Progress of Theoretical Physics, 42, 219 9

McCulloch W. S., Pitts W., 1943, The Bulletin of Mathematical Biophysics, 5, 115 35

McGreer I. D., Mesinger A., Fan X., 2011, Monthly Notices of the Royal Astronomical Society, 415, 3237–3246 14

McGreer I. D., Mesinger A., D'Odorico V., 2014, Monthly Notices of the Royal Astronomical Society, 447, 499–505 14

McLeod D. J., McLure R. J., Dunlop J. S., 2016, Monthly Notices of the Royal Astronomical Society, 459, 3812 9

McQuinn M., Furlanetto S. R., Hernquist L., Zahn O., Zaldarriaga M., 2005, ] 10.1086/432049 13

McQuinn M., Zahn O., Zaldarriaga M., Hernquist L., Furlanetto S. R., 2006, The Astrophysical Journal, 653, 815 22

Mcquinn M., Lidz A., Zahn O., Dutta S., Hernquist L., Zaldarriaga M., 2007, Monthly Notices of the Royal Astronomical Society, 377, 1043 33

Mebane R. H., Mirocha J., Furlanetto S. R., 2018, Monthly Notices of the Royal Astronomical Society, 479, 4544 10

Mehta P., Bukov M., Wang C.-H., Day A. G., Richardson C., Fisher C. K., Schwab D. J., 2019, Physics Reports, 810, 1–124 93

Mellema G., Iliev I. T., Alvarez M. A., Shapiro P. R., 2006a, New Astronomy, 11, 374–395 33, 60

Mellema G., Iliev I. T., Pen U.-L., Shapiro P. R., 2006b, Monthly Notices of the Royal Astronomical Society, 372, 679 16, 19, 20, 48, 50, 54, 76, 91

Mellema G., et al., 2013, Experimental Astronomy, 36, 235 19, 87

Mertens F. G., et al., 2020, Monthly Notices of the Royal Astronomical Society, 493, 1662 19

Mesinger A., Furlanetto S., 2007, The Astrophysical Journal, 669, 663–675 86, 101

Mesinger A., Furlanetto S., Cen R., 2010, Monthly Notices of the Royal Astronomical Society, 411, 955–972 34, 86

Mesinger A., McQuinn M., Spergel D. N., 2012, Monthly Notices of the Royal Astronomical Society, 422, 1403 13

Meurer G. R., Heckman T. M., Calzetti D., 1999, , 521, 64 9

Miralda-Escude J., 2003, The Astrophysical Journal, 597, 66–73 35

Mirocha J., 2014, Monthly Notices of the Royal Astronomical Society, 443, 1211 18

Monaco P., 1997, PhD thesis, Trieste U. (`arXiv:astro-ph/9710085`) 25

Mondal R., et al., 2020, Monthly Notices of the Royal Astronomical Society, 498, 4178–4191 19

Murray S. G., Greig B., Mesinger A., Muñoz J. B., Qin Y., Park J., Watkinson C. A., 2020, Journal of Open Source Software, 5, 2582 86

Nath B. B., Biermann P. L., 1993, Monthly Notices of the Royal Astronomical Society, 265, 241–249 18

Oesch P. A., Bouwens R. J., Illingworth G. D., Labbé I., Stefanon M., 2018, The Astrophysical Journal, 855, 105 9

Onken C., Miralda-Escudé J., 2004, The Astrophysical Journal, 610, 1 48

Oquab M., Bottou L., Laptev I., Sivic J., 2015, in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp 685–694, doi:10.1109/CVPR.2015.7298668 42

Ota K., et al., 2008, The Astrophysical Journal, 677, 12 xvi, 14, 61

Ouchi M., et al., 2010, The Astrophysical Journal, 723, 869 xvi, 14, 61

O'Shea B. W., Norman M. L., 2007, The Astrophysical Journal, 654, 66–92 9

Park H., Shapiro P. R., Choi J.-h., Yoshida N., Hirano S., Ahn K., 2016, The Astrophysical Journal, 831, 86 23, 60

Patterson J., Gibson A., 2017, Deep Learning: A Practitioner's Approach. O'Reilly, Beijing, `https://www.safaribooksonline.com/library/view/deep-learning/9781491924570/` 40

Pawlik A. H., Schaye J., van Scherpenzeel E., 2009, Monthly Notices of the Royal Astronomical Society, 394, 1812 48

Peacock J. A., 1998, Cosmological Physics. Cambridge University Press, doi:10.1017/CBO9780511804533 3

Peebles P. J. E., 1980, The large-scale structure of the universe. Princeton University Press Princeton, N.J 5

Penzias A. A., Wilson R. W., 1965, The Astrophysical Journal, 142, 419 1

Perez L., Wang J., 2017 (`arXiv:1712.04621`) 96

Perlmutter S., et al., 1999, The Astrophysical Journal, 517, 565 1, 3

Planck Collaboration 2019, Astronomy & Astrophysics 87

Planck Collaboration et al., 2016, Astronomy & Astrophysics, 596, A108 12

Planck Collaboration et al., 2020, Astronomy & Astrophysics, 641, A6 29

Pospelov M., Pradler J., Ruderman J. T., Urbano A., 2018, Phys. Rev. Lett., 121, 031103 18

Press W. H., Schechter P., 1974, The Astrophysical Journal, 187, 425 34

Pritchard J. R., Furlanetto S. R., 2007, Monthly Notices of the Royal Astronomical Society, 376, 1680 87

Pritchard J. R., Loeb A., 2012a, Reports on Progress in Physics, 75, 086901 xii, 17

Pritchard J. R., Loeb A., 2012b, Reports on Progress in Physics, 75, 086901 25

Prochaska J. X., O'Meara J. M., Fumagalli M., Bernstein R. A., Burles S. M., 2015, , 221, 2 35

Raičević M., Theuns T., 2011, Monthly Notices of the Royal Astronomical Society, 412, 1 54

Rasera Y., Alimi J., Courtin J., Roy F., Corasaniti P., Füzfa A., Boucher V., 2010, AIP Conference Proceedings, 1241, 1134 31

Rauch M., 1998, Annual Review of Astronomy and Astrophysics, 36, 267 28

Rees M. J., 1968, , 153, L1 12

Ribaudo J., Lehner N., Howk J. C., 2011, The Astrophysical Journal, 736, 42 35

Riess A. G., et al., 1998, The Astronomical Journal, 116, 1009 1, 3

Riess A. G., et al., 2007, The Astrophysical Journal, 659, 98 1

Riess A. G., Casertano S., Yuan W., Macri L. M., Scolnic D., 2019, The Astrophysical Journal, 876, 85 4

Robertson B. E., Ellis R. S., Dunlop J. S., McLure R. J., Stark D. P., 2010, Nature, 468, 49–55 8

Robertson B. E., Ellis R. S., Furlanetto S. R., Dunlop J. S., 2015a, ] 10.1088/2041-8205/802/2/L19, 802, L19 9

Robertson B. E., Ellis R. S., Furlanetto S. R., Dunlop J. S., 2015b, The Astrophysical Journal, 802, L19 14

Ronneberger O., Fischer P., Brox T., 2015 (`arXiv:1505.04597`) 86, 93, 112

Rosdahl J., Blaizot J., Aubert D., Stranex T., Teyssier R., 2013, Monthly Notices of the Royal Astronomical Society, 436, 2188 28

Rosenblatt F., 1962, in , Brain Theory. Washington, Spartan Books, p. 616, doi:10.1007/978-3-642-70911-1$_2$0 35

Ross H. E., Dixon K. L., Iliev I. T., Mellema G., 2017, Monthly Notices of the Royal Astronomical Society, 468, 3785 87, 110

Ross H. E., Dixon K. L., Ghara R., Iliev I. T., Mellema G., 2019, Monthly Notices of the Royal Astronomical Society, 487, 1101–1119 21, 32, 72, 87, 110

Ruder S., 2017, An overview of gradient descent optimization algorithms (arXiv:1609.04747) 40

Rumelhart D. E., Zipser D., 1985, Cognitive Science, 9, 75 35, 39

Rybicki G. B., Lightman A. P., 1986, Radiative Processes in Astrophysics 33

Sadr A. V., Farsian F., 2020 (arXiv:2004.04177) 85

Salehi S. S. M., Erdogmus D., Gholipour A., 2017 (arXiv:1706.05721) 95

Samuel A. L., 1959, IBM J. Res. Dev., 3, 210–229 35

Saslaw W. C., Zipoy D., 1967, , 216, 976 9

Sazonov S., Sunyaev R., 2015, Monthly Notices of the Royal Astronomical Society, 454, 3464 10, 18

Schlaufman K. C., Thompson I. B., Casey A. R., 2018, The Astrophysical Journal, 867, 98 10

Schmidt B. P., et al., 1998, The Astronomical Journal, 507, 46 3

Schmit C. J., Pritchard J. R., 2018, Monthly Notices of the Royal Astronomical Society, 475, 1213 85

Schroeder J., Mesinger A., Haiman Z., 2013, Monthly Notices of the Royal Astronomical Society, 428, 3058 14

Scoccimarro R., Colombi S., Fry J. N., Frieman J. A., Hivon E., Melott A., 1998, The Astrophysical Journal, 496, 586 30

Shapiro P. R., Martel H., Villumsen J. V., Owen J. M., 1996, The Astrophysical Journal, pp 270–330 32, 52, 91

Shimabukuro H., Semelin B., 2017, Monthly Notices of the Royal Astronomical Society, 468, 3869 85

Shimabukuro H., Mao Y., Tan J., 2020, Beyond power spectrum I: recovering HII bubble size distribution from 21cm power spectrum with artificial neural network (arXiv:2002.08238) 85

Shull J. M., Tumlinson J., Giroux M. L., Kriss G. A., Reimers D., 2004, The Astrophysical Journal, 600, 570–579 23

Shull J. M., Danforth C. W., Tilton E. M., Moloney J., Stevans M. L., 2017, , 849, 106 35

Simonyan K., Zisserman A., 2015 (arXiv:1409.1556) 96

Smith L. N., 2017, Cyclical Learning Rates for Training Neural Networks (arXiv:1506.01186) 40

Springel V., et al., 2005, Nature, 435, 629–636 xiii, 28, 30

Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R., 2014, Journal of Machine Learning Research, 15, 1929 94

Storrie-Lombardi L. J., McMahon R. G., Irwin M. J., Hazard C., 1994, , 427, L13 35

Sudholt S., Fink G. A., 2017, PHOCNet: A Deep Convolutional Neural Network for Word Spotting in Handwritten Documents (arXiv:1604.00187) 42

Sullivan D., Iliev I. T., Dixon K. L., 2017, Monthly Notices of the Royal Astronomical Society, 473, 38–58 85

Sunyaev R. A., Zeldovich Y. B., 1972, Comments on Astrophysics and Space Physics, 4, 173 12

Sutter P. M., Ricker P. M., 2010, The Astrophysical Journal, 723, 1308 31

Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z., 2015 (arXiv:1512.00567) 96

Tacchella S., Bose S., Conroy C., Eisenstein D. J., Johnson B. D., 2018, The Astrophysical Journal, 868, 92 xi, 8, 9

Tegmark M., Silk J., Rees M., Blanchard A., Abel T., Palla F., 1996, The Astrophysical Journal, 12, 1 48

Tegmark M., Silk J., Rees M. J., Blanchard A., Abel T., Palla F., 1997, , 474, 1 6

Tesauro G., 1995, Commun. ACM, 38, 58–68 35

Tingay S. J., et al., 2013, Publications of the Astronomical Society of Australia (PASA), 30, 7 19

Tinker J., Kravtsov A. V., Klypin A., Abazajian K., Warren M., Yepes G., Gottlöber S., Holz D. E., 2008, The Astrophysical Journal, 688, 709–728 26

Tomassetti M., Porciani C., Romano-Díaz E., Ludlow A. D., 2014, Monthly Notices of the Royal Astronomical Society, 446, 3330 55

Totani T., Aoki K., Hattori T., Kawai N., 2016, Publications of the Astronomical Society of Japan, 68, 1 14

Trac H., Cen R., 2007, The Astrophysical Journal, 671, 1–13 33

Trac H., Gnedin N. Y., 2011, Advanced Science Letters, 4, 228–243 30, 34

Trott C. M., et al., 2020, Monthly Notices of the Royal Astronomical Society, 493, 4711 19

Trümper J., Hasinger G., 2008, The Universe in X-Rays, doi:10.1007/978-3-540-34412-4. 28

Umeda H., Nomoto K., 2003, Nature, 422, 871–873 9

Valageas P., Silk J., 2004, Astronomy & Astrophysics, 413, 1087 48

Valentino E. D., Melchiorri A., Silk J., 2020, Journal of Cosmology and Astroparticle Physics, 2020, 013–013 29

Villanueva-Domingo P., Villaescusa-Navarro F., 2021, The Astrophysical Journal, 907, 44 85

Virtanen P., et al., 2020, Nature Methods, 17, 261 110

Vogelsberger M., et al., 2014, Monthly Notices of the Royal Astronomical Society, 444, 1518–1547 28

Wald R. M., 1984, General Relativity. Chicago Univ. Pr., Chicago, USA, doi:10.7208/chicago/9780226870373.001.0001 2

Wang Y., Huang G., Song S., Pan X., Xia Y., Wu C., 2020 (`arXiv:2007.10538`) 96

Watkinson C. A., Pritchard J. R., 2015, Monthly Notices of the Royal Astronomical Society, 454, 1416 19

Watson W. A., Iliev I. T., D'Aloisio A., Knebe A., Shapiro P. R., Yepes G., 2013, Monthly Notices of the Royal Astronomical Society, 433, 1230 26, 31, 50, 91

Wei K., Fu Y., Huang H., 2020, 3D Quasi-Recurrent Neural Network for Hyperspectral Image Denoising (`arXiv:2003.04547`) 40

Weinberg S., 1972, Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity. John Wiley and Sons, New York 2

Weinberger R., Springel V., Pakmor R., 2020, The Astrophysical Journal Supplement Series, 248, 32 28

White M., 2014, Monthly Notices of the Royal Astronomical Society, 439, 3630 30

White M., 2015, Monthly Notices of the Royal Astronomical Society, 450, 3822 30

Widrow B., Hoff M. E., 1960, in 1960 IRE WESCON Convention Record, Part 4. IRE, New York, pp 96–104 35

Wilkins S. M., Feng Y., Di-Matteo T., Croft R., Stanway E. R., Bouwens R. J., Thomas P., 2016, , 458, L6 10

Wong K. C., et al., 2019, Monthly Notices of the Royal Astronomical Society, 498, 1420–1439 4

Wouthuysen S. A., 1952, The Astronomical Journal, 57, 31 16

Wyithe S., Geil P. M., Kim H., 2015, Proceedings of Science, AASKA14, 015 19

Xu Y., Yue B., Chen X., 2017, The Astrophysical Journal, 844, 117 101

Yadav K., 2021, A Comprehensive Study on Optimization Strategies for Gradient Descent In Deep Learning (arXiv:2101.02397) 40

Yoshida N., Abel T., Hernquist L., Sugiyama N., 2003, The Astrophysical Journal, 592, 645–663 9

Yoshiura S., Shimabukuro H., Hasegawa K., Takahashi K., 2020 (arXiv:2004.09206) 85

Zackrisson E., et al., 2020, Monthly Notices of the Royal Astronomical Society, 493, 855 85

Zahn O., Mesinger A., McQuinn M., Trac H., Cen R., Hernquist L. E., 2011, Monthly Notices of the Royal Astronomical Society, 414, 727 28

Zaroubi S., 2012, preprint, 369, 1055 (arXiv:1206.0267) 72

Zaroubi S., Thomas R. M., Sugiyama N., Silk J., 2007, Monthly Notices of the Royal Astronomical Society, 375, 1269 18

Zeiler M. D., Fergus R., 2013 (arXiv:1311.2901) 94

Zel'Dovich Y. B., 1970, Astronomy & Astrophysics, 500, 13 6, 29, 86

Zhang J., 2019, Gradient Descent based Optimization Algorithms for Deep Learning Models Training (arXiv:1903.03614) 40

Zhang J., Hui L., Haiman Z., 2007, Monthly Notices of the Royal Astronomical Society, 375, 324–336 48

van Albada T. S., Baker N., 1973, , 185, 477 10

van Haarlem M. P., et al., 2013, Astronomy & Astrophysics, 556, A2 19

van de Hulst H. C., 1945, Nederlandsch Tijdschrift voor Natuurkunde, 11, 210 15

van der Walt S., et al., 2014, PeerJ, 2, e453 110