The Emulation Game:

Modelling and Machine Learning for the Epoch of Reionization

William David Jennings

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Physics and Astronomy

University College London

November 2, 2019

I, William David Jennings, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

The Epoch of Reionization (EoR) is a fascinating time in the Universe's history. Around 400,000 years after the Big Bang, the Universe was full of neutral atoms. Over the following hundred million years or so, these atoms were slowly ionised by the first luminous objects. We have yet to make precise measurements of exactly when this process started, how long it lasted, and which types of luminous sources contributed the most. The first stars and galaxies had only just started to form, so there were precious few emission sources. The 21cm emission line of neutral hydrogen is one such source. The next generation of radio interferometers will measure for the first time three-dimensional maps of 21cm radiation during the EoR. In this thesis I present four projects for efficient modelling and analysis of the results of these EoR experiments. First I present my code for calculating higher-order clustering statistics from observed or simulated data. This code efficiently summarises useful information in the data and would allow for fast comparisons between theory and future observations. Secondly I use machine learning techniques to determine how physical EoR properties are related to the three-point clustering of simulated EoR data. Thirdly I fit an analytic clustering model to simulated 21cm maps. The model gives approximate predictions for the start of the EoR, but is unable to account for the widespread overlap of ionised regions for later times. Finally I use and compare machine learning techniques for replacing the semi-numerical simulations with trained emulators. My best emulated model makes predictions that are accurate to within 4% of the full simulation in a tiny fraction of the time.

Impact Statement

Cosmologists have yet to find compelling answers to a number of questions about the Epoch of Reionization (EoR). Upcoming experiments such as the Square Kilometre Array will help shed some light on these questions. The projects in this thesis are part of a global effort to lay the groundwork for analysis of these experiments' results. The datasets from these experiments will be so large that compression will be needed. My high-order clustering calculation code will be useful for this, not only allowing faster comparisons between data and theories, but also retaining more information than other methods such as the power spectrum. This has the potential to break degeneracies which would otherwise lead to ambiguous interpretations of the data.

Efficiently making testable predictions from our theories is vital. Current predictions come either from analytic models or from simulations. Analytic models are faster but less detailed than simulations. My halo model is a middle-ground between semi-numerical simulations and analytic models. My final model makes use of intermediate results from pre-calculated simulations to accelerate the modelling speed. The resulting approximate predictions could be useful for comparing data from upcoming experiments, but would likely need to be improved before giving decent comparisons.

In two other projects I use machine learning to extract information about the EoR from simulated data. In one of these projects, I train models to extract information about the progress of the EoR directly from simulated data. The good accuracies of these models show that the 3PCF is a useful summary statistic for EoR data, and should certainly be used in future analysis of the 21cm signal. In the

other project I emulate the full behaviour of the simulations directly, from inputs to outputs, without fitting any intermediate steps. This project has the greatest potential for immediate impact, and if EoR data were available now then my published best surrogate model would be able to replace simulations efficiently and accurately. It is likely that advances in computing power will give even better surrogate models in the near future. The major contribution of this project is as a road map for future researchers: I highlight many of the common stumbling blocks for training such EoR emulators. The project also demonstrates the power of surrogate models in the wider context of any high-cost simulations such as the engineering and design sector.

Contents

List of Figures

List of Tables

Chapter 1

Introductory material: Cosmology

Two and half thousand years ago Greek philosophers stood in fields and discussed the origins of the Universe. They drew pictures in the sand and argued over observations made by looking at the night sky with the naked eye. Today we still seek answers to the same great cosmic riddle — how the Universe came to exist — but over time our questions have evolved and, crucially, we have found ways to quantitatively test the plausibility of our answers. Hundreds of millions of pounds are spent on next-generation multinational experiments which will churn out terabytes of raw data every second. As our questions and experiments have become more sophisticated, so too have the tools that we use to analyse and interpret their results. We have no hope of exploiting these data without continued advances in statistical methods for data analysis. In this thesis I present my work on two crucial aspects of this analysis for Epoch of Reionization (EoR) experiments: compression techniques to extract the most useful information from data; and fast theoretical modelling including mathematical models and machine learning techniques. Armed with these tools we can tackle the formidable datasets from future experiments, and find answers to the remaining unsolved questions about the EoR.

1.1 History of the Universe

Our understanding of the Universe's history is constantly changing with every new experiment. In this section I review the current best model for the evolution of the Universe. In the following section I will review the mathematical aspects of this best model. Throughout both sections I will present the main observations which have led to this model being so widely accepted by modern cosmologists.

The idea that the Universe is expanding has been popular since Edwin Hubble's observations of nearby spiral galaxies (Hubble, 1929). Hubble saw that several of the nearby galaxies are moving away from us and, crucially, that the speed of the movement increased linearly with distance. The most compelling explanation for these observations is that the Universe is expanding. This belief has been compounded by other even stronger evidence, such as the abundance of light elements in the Universe (see Burles et al. 1999 for a review) and irregularities of relic radiation from the Big Bang (Planck Collaboration, 2018). It is safe to say that most modern cosmologists would disregard any theory which does not include an expanding universe, describing such models as worthy of being sent "from the pages of physics journals to the far reaches of radical Internet chat groups" (Dodelson, 2003).

If the Universe has been continually expanding then, in the past, it was much smaller than it is today. By measuring the rate of expansion over the past few billion years, we can extrapolate the Universe's size back until it took up a tiny point in space. This event — the appearance of all the Universe's energy — is what we call the Big Bang. Until around 10^{-43} seconds after the Big Bang (hereafter written 'at 10^{-43}s'), the energy levels were so high that our current theories of gravity and the other fundamental forces are effectively useless. This time scale is known as the *Planck time* after Planck (1899) suggested a fundamental quantum of time. Figure 1.1 shows a schematic of the main epochs of the Universe from the Big Bang until the present, according to the current prevailing model.

Starting at around 10^{-36}s it is extremely likely that the Universe underwent a sudden and dramatic period of expansion known as *inflation*, first proposed by

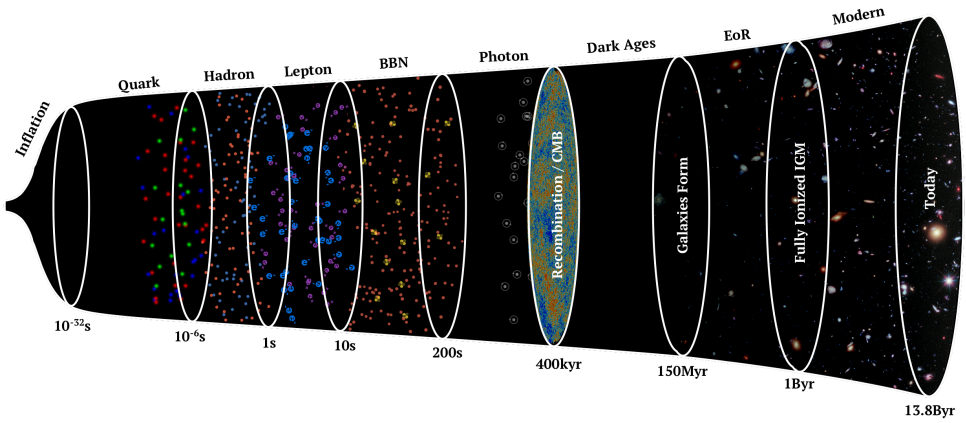

Figure 1.1: Schematic of the main epochs during the history of the Universe, according to the ΛCDM model. Labels above the schematic show the names of each epoch. Labels below show times after the Big Bang.

Guth (1981). In a tiny fraction of a second the Universe grew by at least sixty 'e-folds' or sixty factors of e (Liddle and Leach, 2003) which is roughly 10^{26}. The most compelling evidence for inflation is the startling homogeneity of the Universe, as measured by relic radiation from the Big Bang. The near-perfect black body spectrum of this radiation strongly suggests that the entire Universe was in thermal equilibrium (Mather et al., 1994). The edges of the observable Universe are too far apart to have been in thermal equilibrium unless it underwent a period of superluminal expansion in its past – which we call inflation. Inflation also helps resolve two otherwise inexplicable observations: the low abundance of magnetic monopoles, and the geometric flatness of the Universe.

At around 10^{-12}s the Universe consisted of a plasma of quarks and their exchange particles known as gluons. This is known as the Quark Epoch and is the closest time to the Big Bang that we have been able to recreate with particle accelerators, see for example O'Luanaigh (2015). At around 10^{-6}s the continued expansion and cooling of the Universe allowed the quarks to coalesce into individual hadrons (protons, neutrons) and their antiparticles. This era is known as the Hadron Epoch. Protons are slightly lighter than neutrons which meant that around seven protons were formed for each neutron (Ghosh et al., 2015). The particles and anti-particles quickly annihilated until around 1s when only matter remained,

probably due to a very slight under-abundance of anti-matter (the matter-antimatter asymmetry problem, see Canetti et al. (2012) for a review). Around this same time, the ultra-lightweight particles known as neutrinos decoupled from baryonic matter. These neutrinos are still travelling today since they interact weakly with the rest of the Universe but, for the same reason, direct observation is unlikely without dramatic increases in equipment sensitivity, see for example Faessler et al. (2016). Instead, we have indirect evidence for these neutrinos based on their interaction with the Big Bang relic radiation.

Between 1s and 10s, the same annihilation process occurred for leptons during the Lepton Epoch: electrons and anti-electrons annihilated, leaving behind only electrons. From 10s to 1000s, the Universe cooled enough that protons and neutrons were able to bind and form the first atomic nuclei. This process has been modelled by Alpher et al. (1948) from which it is possible to predict the relative abundance of various different nuclei: 75% hydrogen; 25% helium-4; and less than 0.01% of deuterium and lithium.

For many hundreds of thousands of years the Universe continued to expand and cool, with efficient scattering of photons maintaining thermal equilibrium between the soup of nuclei and electrons. After 380,000 years the Universe became cold enough that electrons could bind with the nuclei: the first atoms came into existence. This event is known as recombination. After recombination, photons scattered less and streamed freely in the same direction as their last scattering event. These photons form a spherical *last scattering surface* which we observe today as the Cosmic Microwave Background (CMB) radiation. This radiation was first observed by Penzias and Wilson (1965) who published the result as an "Excess Antenna Temperature" of around 3 Kelvin. Many discoveries have since been made by comparing observations of the CMB with predictions from theory. In particular analysis of its spectrum (Mather et al., 1994) and inhomogeneities in its temperature and polarisation (Planck Collaboration, 2018). Figure 1.2 shows the spectrum of the CMB from several observations including the COBE satellite (NASA Goddard Space Flight Center, 2019) with the fitted black body spectrum.

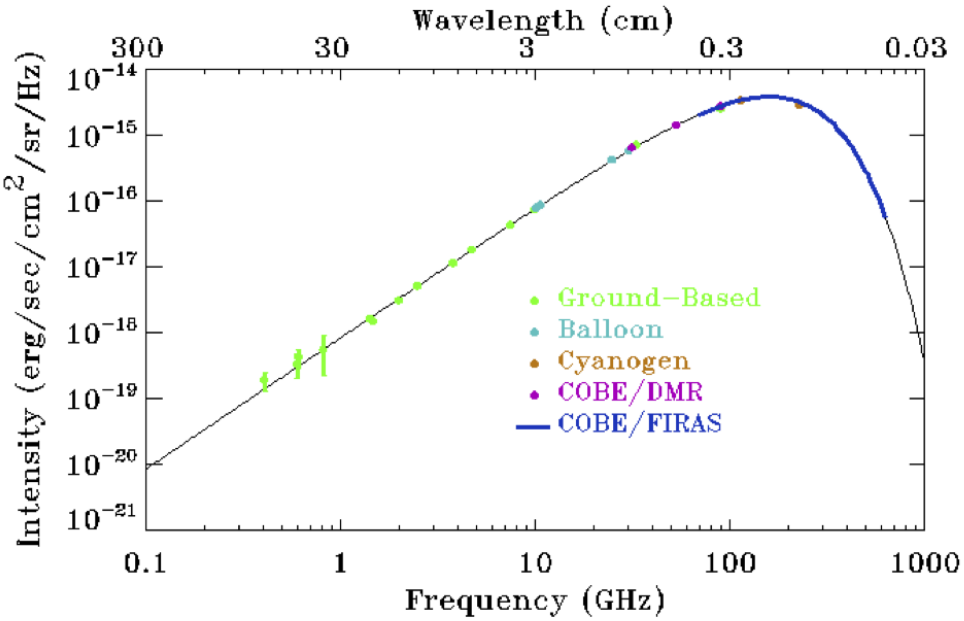

Figure 1.2: Spectrum of the CMB as measured by several instruments including the COBE satellite. This spectrum is strong evidence in favour of the Big Bang model as it shows the Universe was much hotter and denser in the past.

For around 150 million years after recombination the Universe went through a period known as the Dark Ages. Although the photons from the CMB were still able to travel freely, the first stars and galaxies had yet to form and the Universe remained dark because precious few new photons were emitted. New emissions came almost exclusively from neutral hydrogen and the spontaneous spin-flip transition between the electron and proton. This event has a very low transition rate of around 10^{-15} seconds and, given the presence of many other sources at the same frequencies, observing these rare photons requires highly sensitive radio telescope arrays. The Dark Ages finished at the start of the Epoch of Reionization (EoR), when the action of gravity had caused hydrogen to collapse into the first stars and galaxies. The radiation from these first luminous structures caused the remaining surrounding hydrogen to ionise. Bubbles of ionised hydrogen grew around ionising sources over time and, around 1 billion years after the Big Bang, eventually overlapped to fill the entire Universe. From 1 billion years onwards the Universe looked much as it does today, with stars clustered into groups or galaxies, and galaxies into clusters.

1.2 ΛCDM cosmology

In this section I review the status of the prevailing cosmology model known as ΛCDM ("Lambda Cold Dark Matter"). ΛCDM is a Big Bang model which parameterises the expansion of the Universe and everything it contains. The model includes both dark matter and dark energy.

1.2.1 Expanding Universe

The expansion of the Universe causes the physical distance between objects to increase over time. The scale factor $a(t)$ is used to parameterise this expansion: $a(t)$ gives the ratio of the physical distance between objects as a function of time. By convention, the scale factor is normalised so that the current value of $a(t)$ is unity. For cosmological models in which the Universe is continually expanding, values of $a(t)$ in the past are then less than one, and values in the future will be greater than one. Expansion causes nearby galaxies to move away from us. Any emitted light from these galaxies will be stretched out when we observe it. The extent of stretching depends on the speed of recession. The redshift of an observed source is defined as the extent of stretching relative to the emitted wavelength. Formally, redshift is the fractional increase in wavelength between the observed and emitted wavelengths,

$$z \equiv \frac{\lambda_{\text{obs}} - \lambda_{\text{emitted}}}{\lambda_{\text{emitted}}}, \tag{1.1}$$

$$\text{or } 1 + z = \frac{\lambda_{\text{obs}}}{\lambda_{\text{emitted}}}. \tag{1.2}$$

A distinction is made between redshift due to the peculiar radial movement of galaxies (as described above), and cosmological redshift which arises from the expansion of space itself. The cosmological redshift is directly related to the scale factor by

$$1 + z(t) = \frac{1}{a(t)}. \tag{1.3}$$

The expansion rate of the Universe is usually quantified using the Hubble pa-

rameter,

$$H(z) = \frac{da(z)/dt}{a(z)}. \tag{1.4}$$

The present value of the Hubble parameter is written H_0, with the subscript '0' being a frequent shorthand in cosmology for the current value of a quantity which changes with time. Current best values for H_0 differ depending on which observational probe is used. Planck Collaboration (2018) observations of the CMB give

$$H_0 = 67.4 \pm 0.5 \text{kms}^{-1}\text{Mpc}^{-1}; \tag{1.5}$$

supernovae measurements from the Dark Energy Survey (Macaulay et al., 2018) give

$$H_0 = 67.77 \pm 1.30 \text{kms}^{-1}\text{Mpc}^{-1}; \tag{1.6}$$

and measurements of Cepheid variables from the Hubble space telescope Riess et al. (2018) give

$$H_0 = 73.24 \pm 1.7 \text{kms}^{-1}\text{Mpc}^{-1}. \tag{1.7}$$

There is a surprisingly large inconsistency between the last value and the other two values. This 'tension' is the subject of much debate among cosmologists. In July 2019, a workshop was held at the Kavli Institute for Theoretical Physics specifically to discuss this topic. After the workshop, Verde et al. (2019) compiled the different values for H_0 and summarised the values in Figure 1.3. The tension is quoted as being between roughly 4σ and 6σ, a significant result that certainly warrants continued investigation. Indeed, many attempts have been made at explaining the discrepancy. Bernal et al. (2016) consider the effect of early- and late-time physics on this tension, particularly noting that the CMB-inferred value for H_0 depends on both early- and late-time assumptions. Bernal et al. (2016) also consider possible non-ΛCDM physics such as extra relativistic particles other than the three standard neutrinos, but state that these models are not favoured by Planck Collaboration

(2018). Mörtsell and Dhawan (2018) investigated whether the tension could indicate a new or different type of dark matter. Extra dark energy components at early times cause a faster pre-CMB expansion which, in turn, would cause over-estimates of H_0. They conclude that extra components are unlikely to resolve the tension, although current data cannot completely rule this out. Data from Planck Collaboration (2018) strongly indicates that the Universe has a near-zero spatial curvature, but Bolejko (2018) use a ray-tracing simulation to show that allowing a time-evolving spatial curvature could ease the tension.

The value H_0 is often parameterised using the h parameter (pronounced 'little h'), with $H_0 = 100.0h\text{kms}^{-1}\text{Mpc}^{-1}$. For instance the first value above would then be quoted as $h = 0.674 \pm 0.005$. The units for the Hubble parameter indicate the method of original measurement, using the recession speed of galaxies (in kms^{-1}) and dividing by their distance (in Mpc).

1.2.2 Energy in the Universe

Much of modern cosmology is based on solutions to the Einstein field equations. Einstein suggested that the geometry of spacetime is related to the contents of the Universe. This statement is most succinctly summarised by

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}, \tag{1.8}$$

a simple equation which hides a wealth of information in the four-dimensional tensors on each side. The left-hand side represents the geometry of spacetime. The Einstein tensor $G_{\mu\nu}$ summarises the curvature and expansion of space over time. $G_{\mu\nu}$ is related to the geometry of spacetime through a series of connected equations: starting with the metric $g_{\mu\nu}$ for the coordinate system for an expanding universe, this tensor is transformed and combined into the Christoffel symbols, which are then combined again into the Ricci tensor and finally the Einstein tensor above.

The energy tensor $T_{\mu\nu}$ on the right-hand side of Equation (1.8) summarises the different energy components of the Universe. A solution to the Einstein field equations for an expanding Universe model was found by Friedman (1922). The

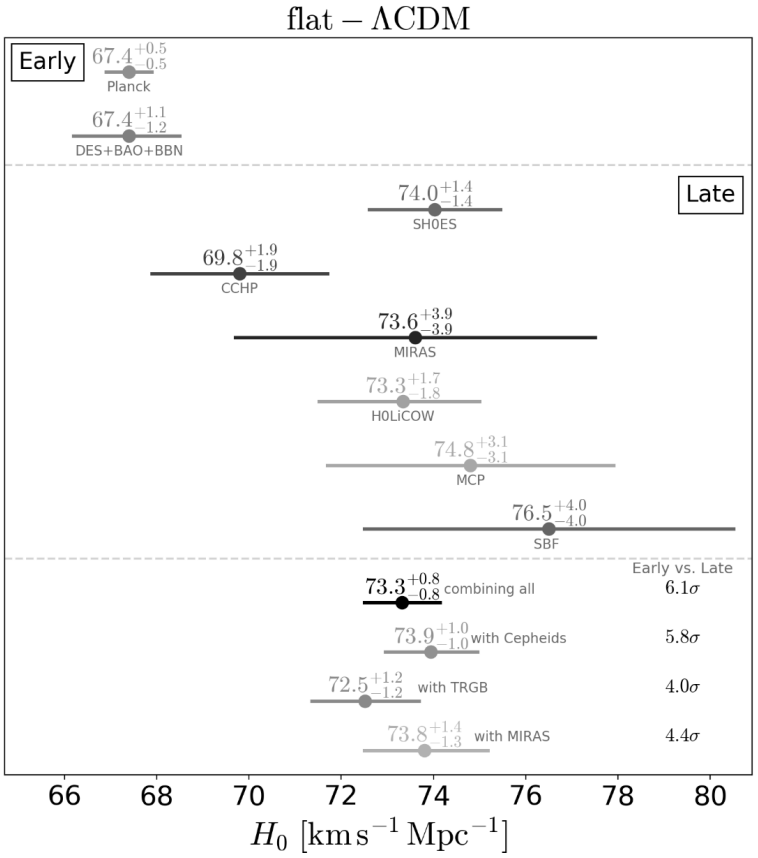

Figure 1.3: Hubble Constant predictions and measurements discussed at the July 2019 Kavli Institute for Theoretical Physics workshop, taken from Verde et al. (2019). Two predicted values using early-Universe measurements are shown at the top, using the results of Planck Collaboration (2018) and Abbott et al. (2017). The measured values using late-Universe measurements are shown in the middle section, and are generally larger than the early-Universe values (see Verde et al. (2019) for the full set of references). The bottom section shows values using combined late-Universe measurements, explicitly listing the tensions of these values with early-time measurements on the right side of the figure.

solution allows us to write the evolution of the Universe's expansion rate as a function of time,

$$H^2(t) = \frac{8\pi G}{3}\Big(\rho(t) - \big[\rho_{\rm crit} - \rho_0(1+z)^2\big]\Big), \tag{1.9}$$

where ρ_0 is the current-day Universe energy density, $G = 6.674 \times 10^{-11}\mathrm{Nkg}^{-2}\mathrm{m}^2$ is Newton's gravitational constant, and $\rho_{\rm crit}$ is the critical density of the Universe, related mathematically to the current day expansion by

$$\rho_{\rm crit} = \frac{3H_0^2}{8\pi G}. \tag{1.10}$$

The critical density is the energy density that gives a geometrically flat universe. If the energy density of the Universe is greater than the critical density then the Universe is closed and will collapse in the future. If the energy density is lower than the critical value, then the Universe is open and parallel lines will slowly diverge from one another. A flat universe lies between these two extremes: parallel lines stay parallel and the Universe will continue to expand. The current best value from Planck Collaboration (2018) is $\rho_{\rm crit} = 1.87847 \times 10^{-29}\mathrm{h}^2\mathrm{gcm}^{-3}$, equivalent to around five protons per cubic meter. The energy stored in the various components of the Universe are usually quoted as density parameters. The density parameter Ω_i of a component 'i' is the ratio of the component's energy density ρ_i to the critical density,

$$\Omega_i = \rho_i/\rho_{\rm crit}. \tag{1.11}$$

1.2.3 Dark matter

All things we can see and touch on the Earth are made of normal baryonic matter: protons and neutrons. We would be forgiven for thinking that the entire Universe is made up of such matter, but this is almost certainly false. Zwicky (1933) observed the rotation speeds of galaxies in the Coma Cluster. The velocities of these galaxies were too high to be explainable by the presence of only the luminous matter in stars. Zwicky suggested the presence of an alternate unseen massive component

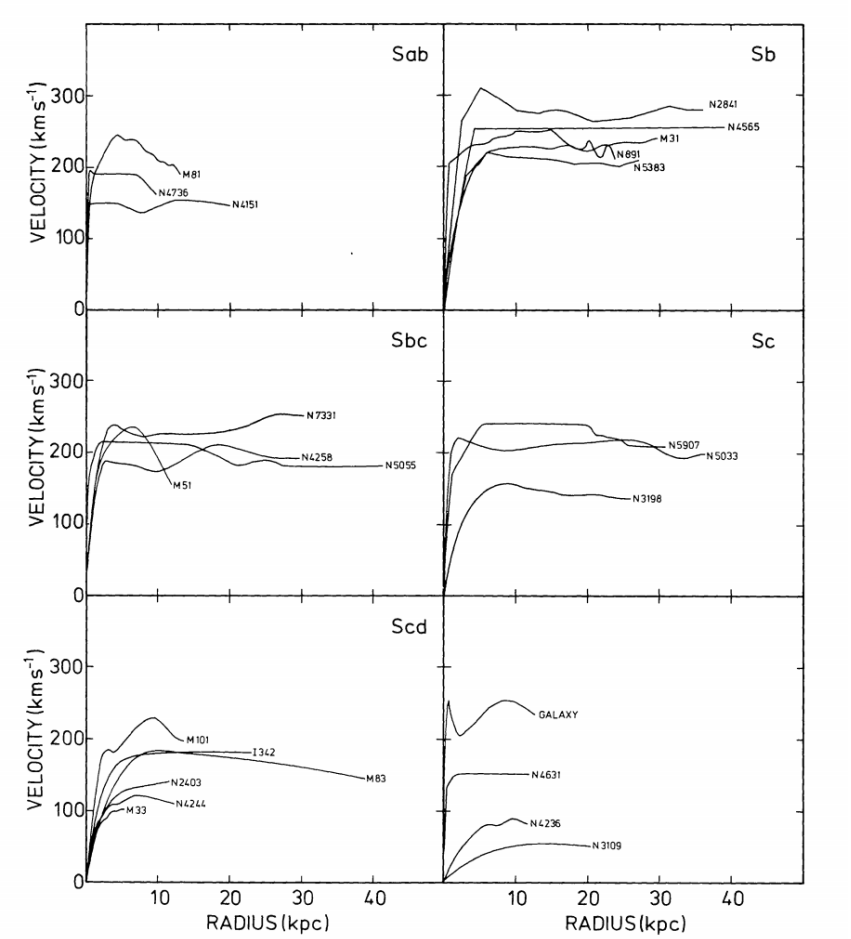

Figure 1.4: Rotation curves of 25 galaxies, taken from Faber and Gallagher (1979). The flattening of these curves at higher radius can most easily be explained by the presence of extra non-luminous dark matter.

– a dark form of matter. Since then, additional evidence from the movement of stars with galaxies has led to the widespread acceptance of this alternate form of matter. Figure 1.4 shows the rotation curves for 25 galaxies taken from Faber and Gallagher (1979). The flattening of these curves at higher radius is most easily explained by the presence of extra non-luminous matter. By observing the large-scale structure of the Universe, we can determine the preferred clustering scale of this dark matter. The clustering scale gives an indication of the temperature of dark matter: hotter matter has higher pressure and gives rise to more widely spaced-out structures, whereas colder matter is able to clump more effectively and form smaller structures. The observed clustering scales of dark matter tell us that it is likely cold, in that it moves at speeds much less than the speed of light.

Rotation curve observations betrayed the existence of extra non-luminous matter but gave no indication as to its form. For this we turn to Big Bang nucleosynthesis (BBN) analysis of the conditions in the early Universe. BBN gives an indication of the abundances of the light elements formed during the Hadron Epoch (see Ghosh et al. 2015 and Section 1.1 earlier). The results of this analysis indicate that baryonic matter can account for less than a quarter of the total matter in the Universe. Whatever the form of dark matter, it is almost certainly not baryonic. There are many other possible candidates including Weakly Interacting Massive Particles (WIMPS), slow-moving particles known as axions, and low-interacting sterile neutrinos.

1.2.4 Dark energy

In the late 1990s observations of supernovae recession speeds suggested that matter was not the only constituent of the Universe (Riess et al. 1998 and Perlmutter et al. 1998). Einstein himself included the possibility of such an extra energy component in the energy tensor $T_{\mu\nu}$. This cosmological constant term did not change with the expansion of the Universe, unlike the densities of radiation and matter which decrease as the Universe expands. Figure 1.5 shows a measured relationship from Riess et al. (1998) for Type Ia supernovae, giving the observed intensity magnitude of supernovae as a function of their observed redshift. These types of figures are known as Hubble diagrams. In addition to the data, Figure 1.5 shows the predicted relationship for three models of the energy components in the Universe, given in terms of the density parameters for matter (Ω_{M}) and the cosmological constant (Ω_{Λ}). The expansion history was seen to be most consistent with a non-zero cosmological constant. The concept of dark energy is an extension of the cosmological constant, allowing for the energy density to change with time or location. Another argument in favour of dark energy is from the total energy of the Universe. The Universe is known to be almost perfectly flat from CMB anisotropies (for example De Bernardis et al. 2000). The total energy density of a flat Universe should match the critical density, but summing the contributions from baryonic and dark matter only accounts for around one third of this amount.

Parameter	Symbol	Fitted Value
Baryonic matter density	$\Omega_{\rm b}h^2$	0.02237 ± 0.00015
Total matter density	$\Omega_{\rm M}h^2$	0.1430 ± 0.0011
Dark energy density	$\Omega_{\Lambda}h^2$	0.3107 ± 0.0011
Current-day Hubble	H_0	67.36 ± 0.54
Matter fluctuation amplitude	σ_8	0.8111 ± 0.0060
Redshift of reionization	$z_{\rm re}$	7.67 ± 0.73
Optical depth to reionization	τ	0.0544 ± 0.0073

Table 1.1: Cosmological parameters as fitted to observations of anisotropies in the CMB from Planck Collaboration (2018). Observations of the anisotropies in temperature (TT), polarisation (EE) and the cross-correlation between temperature and polarisation (TE) were used for these best-fit values.

Equation (1.9) can be written in terms of the redshift-dependent density parameters $\Omega_i(z)$ for the different constituents of the Universe,

$$\begin{aligned}\left(\frac{H(z)}{H_0}\right)^2 &= \sum \Omega_i(z) \\ &= \Omega_{\rm R}\,(1+z)^4 + \Omega_{\rm M}\,(1+z)^3 + \Omega_{\rm K}\,(1+z)^2 + \Omega_{\Lambda}\end{aligned} \tag{1.12}$$

where the values $\Omega_{\rm R}$ and $\Omega_{\rm K}$ are the current-day density parameters for radiation and spatial curvature.

1.2.5 Cosmological parameters

ΛCDM is a physical Big Bang model which can make quantitative predictions about the history and fate of the Universe. Many aspects of the model are not derived from fundamental principles of physics, but are left as free parameters to be matched with observations. Several of these parameters have been mentioned already: the density parameters for baryonic matter $\Omega_{\rm b}$, total matter $\Omega_{\rm M}$ and dark energy Ω_{Λ}; and the current-day value of the Hubble parameter H_0. Another important parameter is σ_8 which quantifies the amplitude of density fluctuations on a specific scale ($8\,{\rm h/Mpc}$). Table 1.1 shows the best values for some of the ΛCDM parameters from observations of the CMB (Planck Collaboration, 2018). Many parameters are more naturally quoted including the current-day Hubble parameter h.

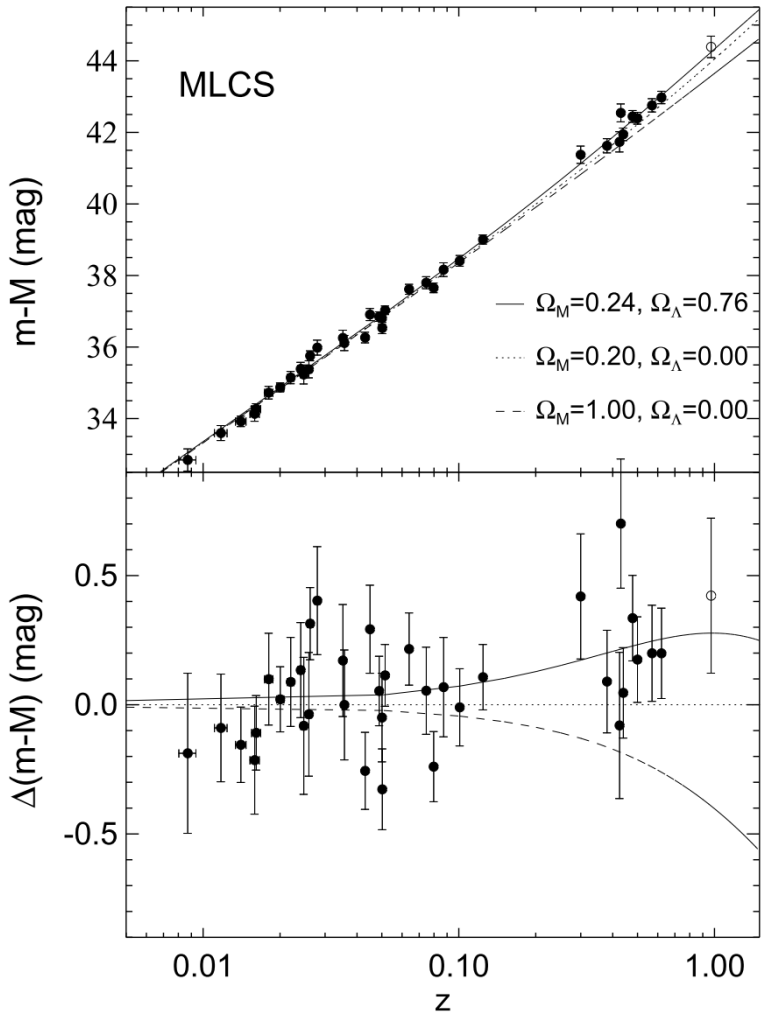

Figure 1.5: Hubble diagram using magnitudes of Type Ia supernovae from Riess et al. (1998). The top panel shows Type Ia supernova magnitudes against redshift with three possible cosmological models: two with zero cosmological constant ($\Omega_\Lambda = 0$) and low- and high-mass universes $\Omega_M = 0.3$ and 1.0; and the best fit model with $\Omega_\Lambda = 0.76$ and $\Omega_M = 0.24$. The bottom panel shows residuals relative to the lower-mass cosmology model, emphasising that the model with non-zero cosmological constant fits the high-redshift measurements more closely.

1.3 Epoch of Reionization physics

This thesis is focused on the Epoch of Reionization. The Universe after recombination was filled with neutral hydrogen. This fully-neutral Universe can be observed in the CMB around 400,000 years after the Big Bang. However observations of quasars (see for example Becker et al. 2007 and Gunn and Peterson 1965) show that the Universe was almost fully ionised at around 1 billion years after the Big Bang. At some point in the intervening time the hydrogen in the Universe transitioned from fully neutral to fully ionised. The most likely culprits of this dramatic change are the earliest luminous sources. As soon as the first stars and galaxies began to form, they emitted radiation and started to ionise the surrounding regions of neutral hydrogen. At first these ionised regions remained small and isolated around each source, but they soon began to overlap. The overlapping regions allowed free passage to any new ionising photons. These excess photons could travel further and ionise ever more distant neutral regions. The growth of ionised regions thus accelerated and eventually filled the entire Universe. Actually observing this process requires distinguishing between ionised regions and neutral regions of space. One of the most promising probes for this is the 21cm transition of hydrogen, which is emitted exclusively by neutral hydrogen during the proton-electron spin interaction. In this section I introduce the framework for analysing 21cm radiation.

1.3.1 21cm measurements

Transition from parallel to anti-parallel spin alignment between the proton and electron in neutral hydrogen causes an overall decrease in energy. This happens spontaneously — albeit with an extremely low probability — and leads to the emission of a photon with a wavelength of 21cm. The low probability of emission means that the probability of re-absorption is also small. These photons are able to travel for billions of years and can eventually be observed by radio telescopes on Earth. Measurements of the 21cm signal on the sky thus give us an image of the neutral hydrogen in the Universe. By tracing this signal back through redshifts, we can

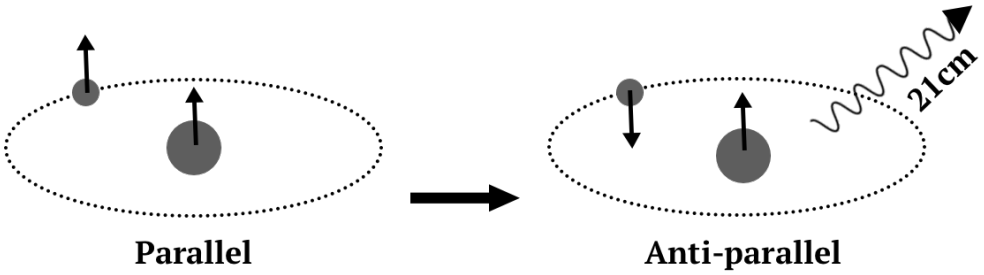

Figure 1.6: The spontaneous spin-flip transition in neutral hydrogen leads to the emission of a photon at 21cm.

extend these two-dimensional images into three-dimensional maps.

The observed 21cm intensity from a region in space depends on three factors: the emitted intensity due to 21cm transitions of hydrogen; the background illumination intensity from other sources at the same wavelength; and absorption of radiation during the journey to Earth. Consider the radiation from a distant background source with specific intensity I_ν^0. This radiation travels along the line of sight to our detectors, passing through regions of space with varying absorption coefficients α_ν and emission coefficients j_ν. The radiative transfer equation can be used to find a solution for the observed intensity of radiation received at our detectors,

$$I_\nu^{\text{obs}} = I_\nu^0 e^{-\tau_\nu} + \frac{j_\nu}{\alpha_\nu}\left(1 - e^{-\tau_\nu}\right), \tag{1.13}$$

where $\frac{j_\nu}{\alpha_\nu}$ is the net emission intensity along the path from the background source to our detectors. The optical depth $\tau_\nu = \int_S \alpha_\nu(s)ds$ quantifies the total absorption along the ray path. The Rayleigh-Jeans law

$$I_\nu = \frac{2\nu^2 k_B T_\text{b}}{c^2} \tag{1.14}$$

relates the specific intensity I_ν at frequency ν to a brightness temperature T_b. The value $k_B = 1.38 \times 10^{-23}\text{JK}^{-1}$ is the Boltzmann constant. The intensities in Equation (1.13) can thus be converted to temperatures

$$T_\text{b}^{\text{obs}} = T_\text{b}^0 e^{-\tau_\nu} + T_\text{b}^{\text{ex}}(1 - e^{-\tau_\nu}), \tag{1.15}$$

where $T_{\rm b}^{\rm obs}$ is the observed temperature at our detectors, $T_{\rm b}^{0}$ is the background illumination temperature, and $T_{\rm b}^{\rm ex}$ is the net emission temperature along the path. If the background illumination is the CMB temperature (T_Γ) and the excitation temperature is from 21cm emissions, then we write

$$T_{\rm b}^{\rm obs} = T_\Gamma e^{-\tau_\nu} + T_{\rm S}(1-e^{-\tau_\nu}), \tag{1.16}$$

where the excitation temperature $T_{\rm S}$ is known as the spin temperature, defined by

$$\frac{n_1}{n_0} = 3\ \exp\left(-\frac{\Delta E_{21\rm cm}}{k_{\rm B}T_{\rm S}}\right). \tag{1.17}$$

The spin temperature is controlled by the relative number densities of exited (n_1) and de-excited (n_0) hydrogen atoms. Here $\Delta E_{21\rm cm} = 5\times10^{-6}$eV is the energy change for the 21cm transition.

The 21cm differential brightness temperature $\delta T_{\rm b}$ is then the difference between the observed temperature $T_{\rm b}^{\rm obs}$ and the background CMB temperature $T_\Gamma(z)$:

$$\begin{aligned}
\delta T_{\rm b} &= T_{\rm b}^{\rm obs} - T_\Gamma \\
&= T_\Gamma(e^{-\tau_\nu}-1) + T_{\rm S}\left(1-e^{-\tau_\nu}\right) \\
&= (T_{\rm S}-T_\Gamma)\left(1-e^{-\tau_\nu}\right) \\
&= \frac{T_{\rm S}-T_\Gamma}{1+z}\left(1-e^{-\tau_\nu}\right)
\end{aligned} \tag{1.18}$$

with the last step introducing an explicit redshift dependence. The magnitude of $\delta T_{\rm b}$ specifies the extent of 21cm emission ($\delta T_{\rm b}>0$) or absorption ($\delta T_{\rm b}<0$) relative to the CMB. The optical depth τ_ν on the right-hand side of this equation is given by the solution to

$$\tau_\nu = \int_{\rm S}\sigma_{01}\left[1-exp\left(-\frac{\Delta E_{21\rm cm}}{k_{\rm B}T_{\rm S}}\right)\right]\phi(\nu)\ n_0\ ds. \tag{1.19}$$

where $\sigma_{01} = \frac{3c^2A_{10}}{8\pi\nu^2}$ is the 21cm absorption cross section; A_{10} is the 21cm spontaneous emission coefficient with value $2.85\times10^{-15}{\rm s}^{-1}$; and $\phi(\nu)$ is the 21cm line

profile including the effects of thermal broadening, pressure broadening and bulk motion of hydrogen. The value n_0 is again the number density of de-excited hydrogen atoms. A full calculation (Furlanetto et al., 2006) yields the optical depth as

$$\begin{aligned}\tau_\nu &= \frac{3}{32\pi}\frac{hc^3A_{10}}{k_{\rm B}T_{\rm S}\nu^2}\frac{x_{\rm HI}n_{\rm H}}{(1+z)\,\delta_r v_r} \\ &\approx 0.0092(1+\delta)\frac{x_{\rm HI}}{T_{\rm S}}(1+z)^{3/2}\left[\frac{H(z)/(1+z)}{\delta_r v_r}\right]\end{aligned} \tag{1.20}$$

with the total hydrogen number density $n_{\rm H}$, and the radial velocity gradient $\delta_r v_r = dv_\parallel/dr_\parallel$. The optical depth is small at all relevant redshifts, so that $e^{-\tau_\nu} \approx 1-\tau_\nu$. Using this approximation and combining Equations 1.18 and 1.20 gives the approximate relationship

$$\begin{aligned}\delta T_{\rm b}(\mathbf{r}) = 27 x_{\rm HI}(\mathbf{r})\,\left[1+\delta(\mathbf{r})\right]\left(\frac{\Omega_{\rm b}h^2}{0.023}\right)\left(\frac{0.15}{\Omega_{\rm M}h^2}\right)^{1/2} \\ \left(1-\frac{T_\Gamma}{T_{\rm S}}\right)\left(\frac{1+z}{10}\right)^{1/2}\left(\frac{H(z)}{H(z)+\delta_r v_r(\mathbf{r})}\right)\ {\rm mK}.\end{aligned} \tag{1.21}$$

This approximation includes the effects of neutral hydrogen fraction fluctuations $x_{\rm HI}(\mathbf{r})$; total matter density contrast $\delta(\mathbf{r})$; cosmological parameters for the densities of baryonic matter $\Omega_{\rm b}$ and total matter $\Omega_{\rm M}$; the CMB temperature T_Γ; the spin temperature $T_{\rm S}$; the Hubble parameter $H(z)$; and $\delta_r v_r(\mathbf{r})$, the radial velocity gradient.

1.3.2 Ionisation fraction history

The progress of the EoR can be tracked by observing the redshift history of the global ionisation fraction $\langle x_{\rm HII}\rangle$ or global neutral fraction $\langle x_{\rm HI}\rangle$. Over the course of the EoR, the Universe transitions from completely neutral ($\langle x_{\rm HII}\rangle = 0$) to fully ionised ($\langle x_{\rm HII}\rangle = 1.0$). The precise speed and duration of this process depends on the underlying physics. I summarise the main factors affecting the ionisation fraction history, using the review in Furlanetto et al. (2006).

The rate at which ionising photons are produced by galaxies is strongly affected by three main properties. First, the star formation efficiency f_*, since higher star formation rates will result in more ionising photons. Values of f_* in the local universe are on the order of 10% but typically models with $0.5\% < f_* < 50.0\%$ are considered (see for example Cohen et al. 2016). Secondly, the fraction $f_{\rm esc}$ of ionizing photons that can escape their source galaxy and reach the inter-galactic medium (IGM). Values of $f_{\rm esc}$ for EoR redshifts from numerical simulations can vary between $f_{\rm esc} < 10\%$ (Razoumov and Sommer-Larsen, 2006) and $f_{\rm esc} < 80\%$ (Wise and Cen, 2008). Finally, the number $N_{\rm ion}$ of photons that are actually produced by baryons in stars clearly affects the efficiency. The value of $N_{\rm ion}$ is different for Population II stars (Loeb and Furlanetto, 2013) or early metal-poor Population III stars (Bromm et al., 2001), with typical values of 4000 and 10^4 respectively. The ionising efficiency is then written the product of these properties, namely

$$\zeta_{\rm ion} = A_{\rm He} f_* f_{\rm esc} N_{\rm ion}, \tag{1.22}$$

where $A_{\rm He}$ is a correction factor that accounts for the presence of stellar helium, given by $A_{\rm He} = 4/(4-3Y_{\rm P})$ with $Y_{\rm P} \approx 0.25$ being the mass fraction of helium. If all galaxies have the same ionising efficiency, then the global ionisation fraction can be written

$$\langle x_{\rm HII} \rangle = \zeta_{\rm ion} f_{\rm coll}. \tag{1.23}$$

After hydrogen atoms have been ionised, it is possible that they will recombine with free electrons. A simple prescription to account for recombinations assumes an average number of recombinations $\bar{n}_{\rm rec}$ for each ionised hydrogen atom. In this case, the global ionisation fraction is given by

$$\langle x_{\rm HII} \rangle = \zeta_{\rm ion} f_{\rm coll}/(1+\bar{n}_{\rm rec}). \tag{1.24}$$

If these values are allowed to evolve over time, then the rate of ionisation can be written in terms of both sources and sinks of ionising photons,

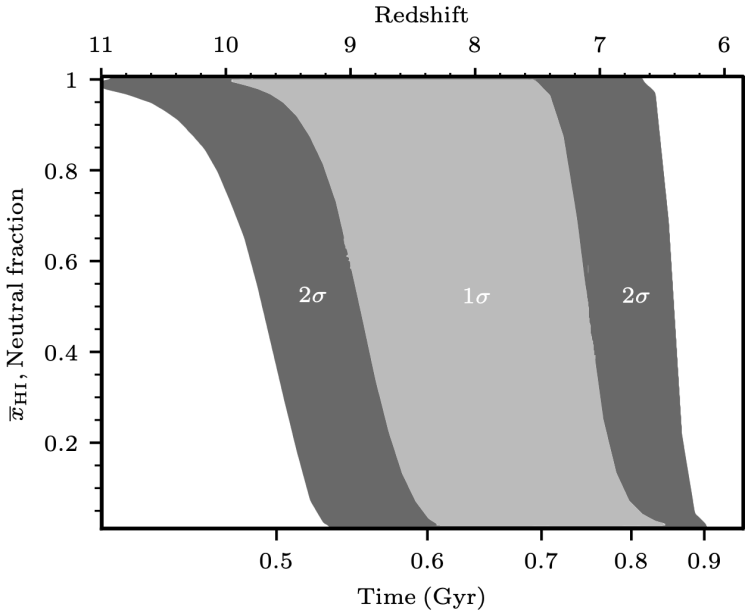

Figure 1.7: Evolution of the global neutral fraction $\langle x_{\rm HI}\rangle = 1 - \langle x_{\rm HII}\rangle$ adapted from Banados et al. (2018). The filled regions show the 1σ and 2σ constraints imposed by several dataset in Planck Collaboration (2016). See Section 1.3.4 for a discussion of these constraints.

$$\frac{d\langle x_{\rm HII}\rangle}{dt} = \zeta_{\rm ion}(z)\frac{df_{\rm coll}}{dt} - \alpha\, C(z, x_{\rm HII})x_{\rm HII}(z)n_{\rm e}(z). \tag{1.25}$$

In this equation, α is the coefficient of recombination controlling the sink term. The value $n_{\rm e}$ is the average number density of electrons, and the clumping factor $C(z, \langle x_{\rm HII}\rangle) = \langle n_{\rm e}^2\rangle / \langle n_{\rm e}\rangle^2$ is a measure of the extent to which electron density varies compared to its mean value. When the clumping factor is large, recombinations are more likely and the ionising efficiency is lower. Efficient modelling of the clumping factor is extremely difficult, since it depends on both small and large scales (see for instance Mellema et al. 2006b, Gnedin 2014 and Kaurov and Gnedin 2018). Figure 1.7 shows a typical range of neutral fraction histories, adapted from Banados et al. (2018). A discussion of constraints on the the midpoint and duration of the EoR is given later in Section 1.3.4.

1.3.3 Spin temperature history

The spin temperature defined in Equation (1.17) can strongly affect the 21cm signal. If the spin temperature is the same as the background CMB temperature then no signal will be observed ($\delta T_{\rm b} = 0$). If the spin temperature is greater than the CMB temperature then an emission signal will be observed ($\delta T_{\rm b} > 0$). If the spin temperature is less than the CMB temperature then an absorption signal will be observed ($\delta T_{\rm b} < 0$). In this subsection I review the evolution of the spin temperature over the history of the Universe, summarising the descriptions in Pritchard and Loeb (2012) and Watkinson and Pritchard (2015). Three main processes affect the evolution of the spin temperature. First, photon emission and absorption from the CMB. Secondly, collisions between neutral hydrogen atoms and electrons which can excite or de-excite the hydrogen to different energy levels. Finally, the Wouthuysen-Field (WF) effect (named after Wouthuysen 1952; Field 1958) in which photons of a different wavelength can trigger a transition in the neutral hydrogen. The overall spin temperature is a combination of these effects,

$$T_{\rm S}^{-1} = \frac{T_\Gamma^{-1} + x_\alpha T_\alpha^{-1} + x_{\rm c} T_{\rm K}^{-1}}{1 + x_\alpha + x_{\rm c}} \tag{1.26}$$

with T_Γ the CMB temperature; T_α the Lyman-α colour temperature for the WF effect; $T_{\rm K}$ the kinetic gas temperature for collisions; and coupling coefficients for scattering of Lyman-α photons (x_α) and atomic collisions ($x_{\rm c}$). The Lyman alpha temperature T_α is closely coupled to the kinetic gas temperature $T_\alpha \approx T_{\rm K}$ by repeated scattering (Pritchard and Loeb, 2008). The evolution of the global spin temperature depends on which effect dominates at each epoch of the Universe. Local fluctuations in these effects can also affect the local spin temperature.

Figure 1.8 shows how the redshift evolution of the spin temperature is related to the evolutions of the gas temperature and background CMB temperature. For $z > 200$, the neutral hydrogen gas remains thermally coupled to the CMB so that $T_{\rm K} = T_\Gamma$. Collisional coupling dominates over the other effects at these times due to the high gas density, so that $T_{\rm S} = T_{\rm K} = T_\Gamma$ and no 21cm signal is observed ($\delta T_{\rm b} = 0$). Between $200 > z > 40$, the gas cools more quickly than it can be heated by the CMB

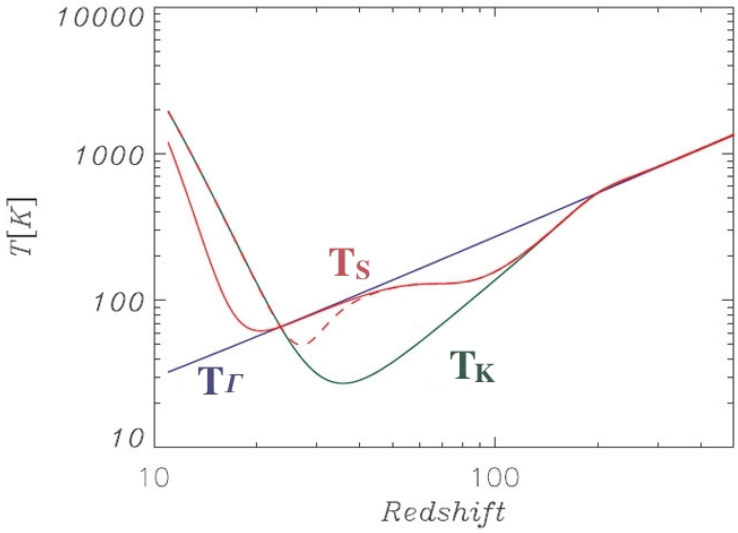

Figure 1.8: Evolution of the spin temperature (red line) over redshift taken from Zaroubi (2019) after changing the text labels to match the convention in this thesis. The kinetic gas temperature $T_{\rm K}$ (green line) and the background CMB temperature T_Γ (blue line) are also shown. The shape of the spin temperature line is described in the text.

($T_{\rm K} < T_\Gamma$). Collisional coupling is still efficient and the spin temperature remains coupled to the gas temperature, so that the spin temperature becomes cooler than the background CMB temperature ($T_{\rm S} < T_\Gamma$) and an absorption 21cm signal would be observed ($\delta T_{\rm b} < 0$). Soon after, the expansion of the Universe causes a much lower gas density. Collisional coupling becomes less efficient, so that the spin temperature again couples to the CMB ($T_{\rm S} = T_\Gamma$) and we again observe $\delta T_{\rm b} = 0$.

Figure 1.9 from Cohen et al. (2016) shows simulated redshift-histories of the 21cm signal for a standard range of reionization scenarios, starting at around $z = 40$. These scenarios have varying parameters for the star formation efficiency (f_*); the efficiency of X-ray sources (f_X); the X-ray spectral energy distribution (SED); and the integrated optical depth to the CMB (τ); and various cooling mechanisms. The efficiency of X-ray sources at high redshift is normally quoted around $f_X \approx 1$, although Fialkov et al. (2015) use the unresolved X-ray background to show that the favoured values of f_X are one to two orders of magnitude higher than the standard literature value. The solid line shows a standard case with atomic cooling of hydrogen, $f_* = 0.05$, $f_X = 1.0$, and a 'hard' X-ray SED (with more high-energy photons

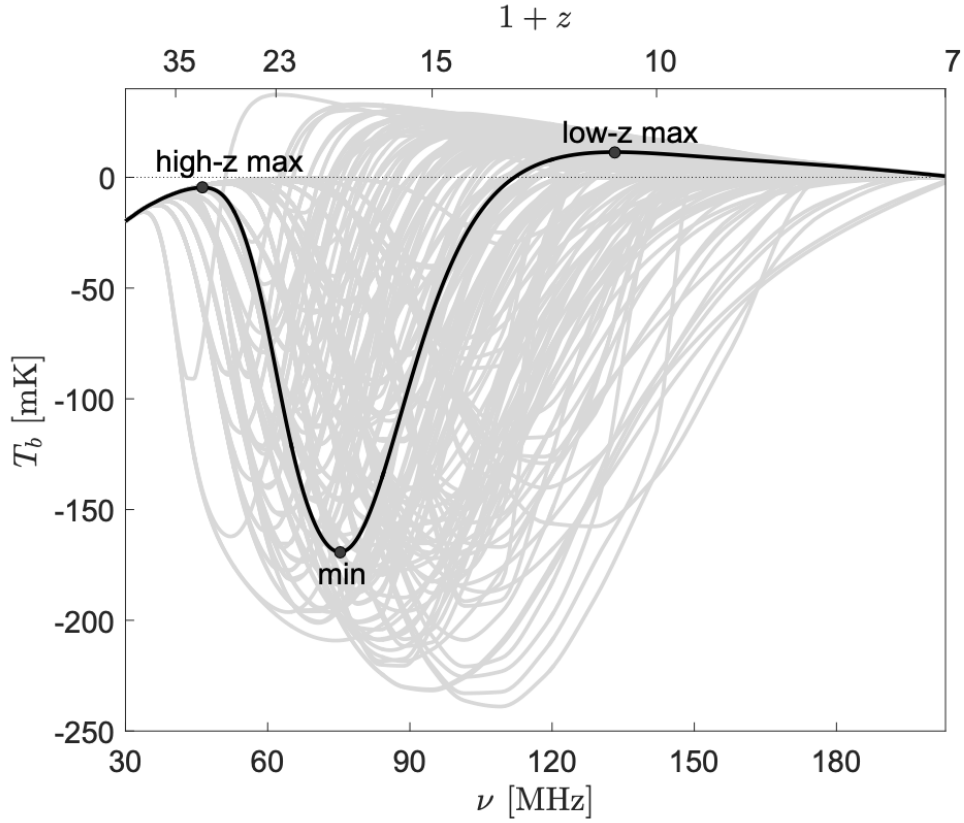

Figure 1.9: Cohen et al. (2016) range of reionization histories for common parameter space restrictions. The use of frequency as x-axis means that more recent times are to the right-hand side of this plot. The solid black line shows the history for a standard scenario with standard atomic cooling of hydrogen, star formation efficiency $f_* = 0.05$, X-ray efficiency $f_X = 1.0$, and the hard X-ray SED from Fialkov et al. (2014).

and fewer low-energy photons) from Fialkov et al. (2014). The 'high-z max' point occurs as just described, with increasingly inefficient collisional coupling giving a small absorption signal rising to no signal ($\delta T_\mathrm{b} = 0$). This remains true until the first stars appear and emit Lyman-α radiation. The spin temperature then couples to the cold gas temperature ($T_\mathrm{K} < T_\Gamma$), giving an absorption signal $\delta T_\mathrm{b} < 0$. This can be seen in Figure 1.9 as the signal decreasing towards its minimum value at the point labelled 'min'. The highest density regions have the most sources and emit the most Lyman-α photons, so local density fluctuations have a strong effect on the spin temperature fluctuations during this time. Eventually, the widespread formation of stars causes the Lyman-α flux to saturate. The gas temperature be-

gins to heat slowly, eventually surpassing the background CMB temperature and giving a 21cm emission signal, $\delta T_b > 0$. This change from large absorption signal to an emission signal can be seen in Figure 1.9, between the minimum point and the low-z-max point. As the gas temperature continues to rise, the emission signal in Figure 1.9 also grows. Eventually the gas temperature greatly exceeds the background CMB temperature ($T_K \gg T_\Gamma$) and so does the coupled spin temperature ($T_S \gg T_\Gamma$). At these times, the spin temperature can be ignored in Equation (1.21), since $1 - \frac{T_\Gamma}{T_S} \approx 1$. For the redshifts in this thesis, the spin temperature is generally larger than the background CMB temperatures. Other than Chapter 4, I ignore the effect of local fluctuations in the spin temperature.

1.3.4 Current observations

Picking out individual local stars from the interstellar medium is easy: the radiation from stars is so much brighter than emissions from the interstellar medium that stars are clearly visible on the sky, even to the naked eye. Observational difficulties have so far prevented us from doing the same with 21cm radiation. The intensity of emitted 21cm radiation from neutral regions is much weaker than other foreground sources at similar frequencies. It is difficult to extract the actual 21cm signal from these foregrounds. Past and ongoing experiments have begun to place limits on the overall intensity of the signal.

The Giant Metrewave Radio Telescope[1] (GMRT) is a set of 30 steerable 45m-diameter dishes located near Khodad, India. The GMRT reported (Paciga et al., 2013) an upper limit on the 21cm power spectrum of $(248\ \mathrm{mK})^2$ at $z = 8.6$ on sacles $k = 0.5\ \mathrm{h/Mpc}$. The Precision Array for Probing the Epoch of Reionization[2] (PAPER, Parsons et al. 2009) is a set of antennae separated into West Virginia and South Africa, trying to mitigate the effects of foregrounds by using instrument calibration and improved antenna design. By removing galactic synchrotron radiation, continuum point-sources, and galactic/extra-galactic radio sources, Ali et al. (2015) placed an upper limit on the power spectrum of $(22.4\ \mathrm{mK})^2$ at $z = 8.4$ in

[1] http://www.gmrt.ncra.tifr.res.in/
[2] http://eor.berkeley.edu/

the range $0.15 < k < 0.5$ h/Mpc. The Murchison Widefield Array[3] (MWA, Tingay et al. 2013) is a set of dipole antennae in Australia, built to target radio frequencies between 80 and 300 MHz with arcminute angular resolution. The MWA made an upper limit detection on the total power in the 21cm line on scale $k = 0.27$ h/Mpc at $z = 7.1$, giving the limit in Beardsley et al. (2016) as $(164 \text{ mK})^2$. The Large-aperture Experiment to detect the Dark Ages[4] (LEDA) is a sub-instrument of the Long Wavelength Array. In Bernardi et al. (2016), LEDA data were used to find an upper limit on the 21cm signal amplitude of $(890 \text{ mK})^2$ in the range $13.2 < z < 27.4$. The LOw Frequency ARray [5] (LOFAR, van Haarlem et al. 2013) is another radio-telescope array located in the Netherlands. LOFAR grappled with the effect of foregrounds in the 21cm signal, caused by relatively-nearby radio sources emitting orders of magnitude brighter than the EoR 21cm signal. The LOFAR upper limit on the 21cm power quoted in Patil et al. (2017) is $(79.6 \text{ mK})^2$ at $k = 0.053$ h/Mpc in the range $z = 9.6 - 10.6$. The Shaped Antenna measurement of the background RAdio Spectum (SARAS, with capitals here intended to indicate the acronym) experiment measured a sky-averaged 21cm signal in Singh et al. (2018). Without giving an explicit limit on the amplitude of the 21cm signal, they considered 264 reionization scenarios and were able to reject 20 of them, most notably excluding scenarios with rapid reionization. The SARAS2 experiment is ongoing.

The Experiment to Detect the Global EoR Signature[6] (EDGES) measured the first detection of a 21cm signal, published in the scientific journal Nature (Bowman et al., 2018). After an initial upper-limit detection using the EDGES High Band data in Monsalve et al. (2017), Figure 1.10 shows the observed 21cm absorption profile in Bowman et al. (2018), centred at 78 MHz with a full-width at half-maximum of 19 MHz and an amplitude of 500mK. The corresponding redshifts for this range of 21cm frequencies are between $15 < z < 20$. This groundbreaking observation has a number of implications for our understanding of 21cm physics. Note the similarity between this observed frequency trough and the predicted profile from Figure 1.9.

[3] http://www.mwatelescope.org/telescope
[4] http://www.tauceti.caltech.edu/leda/
[5] http://www.lofar.org/
[6] https://www.haystack.mit.edu/ast/arrays/Edges/

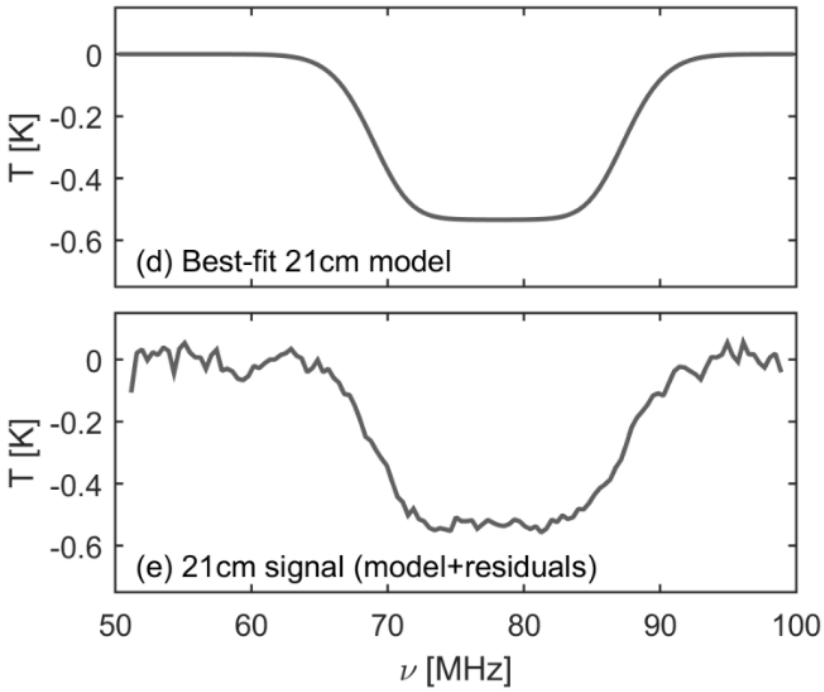

Figure 1.10: Bowman et al. (2018) observation of the sky-averaged 21cm spectrum giving the first detection of 21cm signal. The amplitude and width of the profile provide constraints on the midpoint and duration of the EoR, indicating that the primordial gas was colder than anticipated.

The minimum amplitude of the observed profile (500mK) is at least twice as large as anticipated by any theoretical model shown in Figure 1.9. This indicates that the gas was observed to be colder than anticipated: the spin temperature is coupled to the gas temperature during these times, so a colder gas temperature leads to a larger absorption signal. One possible resolution of this discrepancy is to allow for greater cooling interactions between the gas and dark matter. The low-frequency edge of the observed profile constrains the likely abundance of stars at early times, and the high-frequency edge provides information about when reionization finished.

Upcoming experiments will be able to provide more detailed measurements and should allow us to make better parameter constraints on our reionization models. The Hydrogen Epoch of Reionization Array[7] (HERA, et al DeBoer 2017) is a radiotelescope dedicated to measuring the signal from the EoR. HERA builds on the technology from MWA and PAPER, with a hexagonal grid of 14 meter dishes.

[7] http://reionization.org/

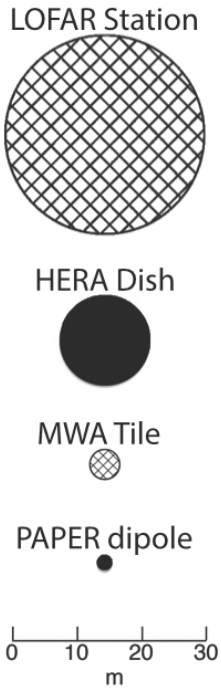

Figure 1.11: Relative sizes of the HERA elements compared to previous experiments. A single HERA element has forty times the collecting area of a PAPER element, and seven times the area of an MWA element. One HERA dish is also a significant fraction of the total collecting area of LOFAR.

Figure 1.11[8] shows the size of the HERA dishes relative to the last generation of experiments. The data from all dishes are combined and correlated, giving a total data output between 4TB and 8TB per night. The current commissioning array has 19 dishes, but the full array will eventually contain 350 dishes. The commissioning array has been used to analyse the polarization response (Kohn et al., 2018) of the array. Construction of more elements is ongoing.

The Square Kilometre Array[9] (SKA, Mellema et al. 2012) is an international experiment situated across South Africa and Australia. With thousands of dishes and millions of radio antennae, the total collecting area of the SKA will be over a million square meters, equivalent to an area of land one kilometre by one kilometre, hence the name 'Square Kilometre' Array. The SKA will start making observations in the mid-2020s. Many recent publications quote the SKA as a benchmark of future experiments – using the projected sensitivity and resolution to determine what

[8] http://reionization.org/science/technical-design/
[9] https://www.skatelescope.org/

future constraints the SKA will place on current theories.

Observations of extremely luminous distant quasars give some idea as to when reionization likely ended. The spectra of these quasars show a distinctive Gunn-Peterson trough (see Gunn and Peterson 1965) due to absorption by neutral hydrogen along the line of sight between us and the source object. The frequency of this trough indicates the redshift at which intergalactic neutral hydrogen was present in significant abundances. Using this method, Becker et al. (2001) and Bolton et al. (2011) state that the Universe was 90% ionised by $z = 7.1$ and that reionization had almost certainly finished by $z \approx 6$.

Thomson scattering of CMB photons can also be used to constrain the duration (Δz) and midpoint ($z_{\rm re}$) of the EoR. Free electrons in space cause CMB photons to scatter. The strength of this scattering depends on the photon polarisation, so that the presence of free electrons imprints a signal onto the polarisation maps of the observed CMB. Measurements of the total integrated optical depth of Thomson scattering from Planck Collaboration (2018) are best fit by the midpoint of reionization occurring at $z_{\rm re} \approx 7.7$. This model assumes a *tanh* reionization history with a full-width of two redshifts.

A more detailed method for constraining Δz and $z_{\rm re}$ involves observing the effect of the bulk movement of the Universe. Bulk movement of electrons causes additional secondary anisotropies, known as the kinetic Sunyaev-Zeldovich (kSZ) effect. Measurements of the kSZ effect using the South Pole Telescope (SPT) indicated that reionization lasted less than $\Delta z < 3$ with $1 - \sigma$ confidence, as reported in George et al. (2014). The measured amplitude of the kSZ effect can be used to determine the most likely values of $z_{\rm re}$ and Δz. For a longer reionization (larger Δz) the movement of electrons has a longer time over which to affect the CMB. A higher amplitude of kSZ effect thus indicates that reionization occurred over a longer range of redshifts. Figure 1.12 from Planck Collaboration (2016) shows the resulting constraints on $z_{\rm re}$ and Δz. The blue contours show the 1σ and 2σ regions of parameter space imposed by using only the kSZ amplitude. The green contours show the constraints by additionally requiring that $z_{\rm re} > 6$ from the Gunn-Peterson

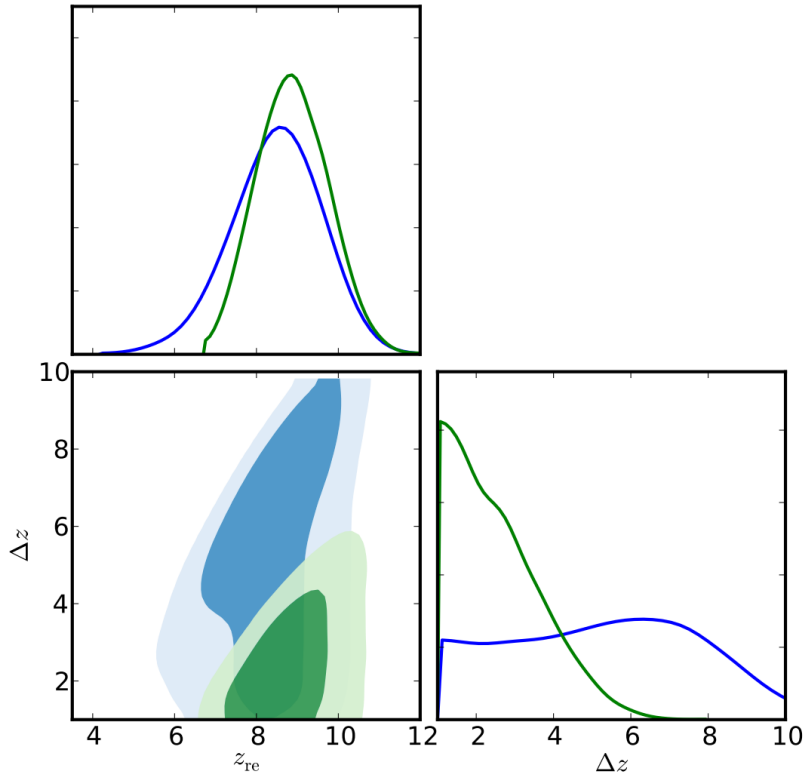

Figure 1.12: Planck Collaboration (2016) constraints on the midpoint of reionization ($z_{\rm re}$) and on its duration (Δz). The blue contours show the constraints including only the kinetic Sunyaev-Zeldovich effect amplitude. The green contours show the constraints by requiring that $z_{\rm re} > 6$ from the Gunn-Peterson troughs off distant quasars.

troughs of distant quasars. The combined results indicate that $7 < z_{\rm re} < 10$ and $\Delta z < 4$.

Sources of Lyman-alpha radiation can be used to measure the neutral fraction at different redshifts. Dijkstra et al. (2011) find that the Universe at $z = 8.6$ is still highly neutral, by modelling the Lyman-alpha emission from a galaxies in dark matter halos at $z = 8.6$. Kakiichi et al. (2015) use the visibility of Lyman-α emitters during the EoR to constrain the neutral fraction at $z = 7$. They find that the neutral fraction to be of order tens of percent at these times. Many galaxies show a distinctive Lyman-alpha break, since radiation with shorter wavelengths than the Lyman limit (9.12×10^{-8} m) are absorbed by neutral hydrogen. Observations of the location of this break in the spectra the indicate the galaxy's redshift, and can high-

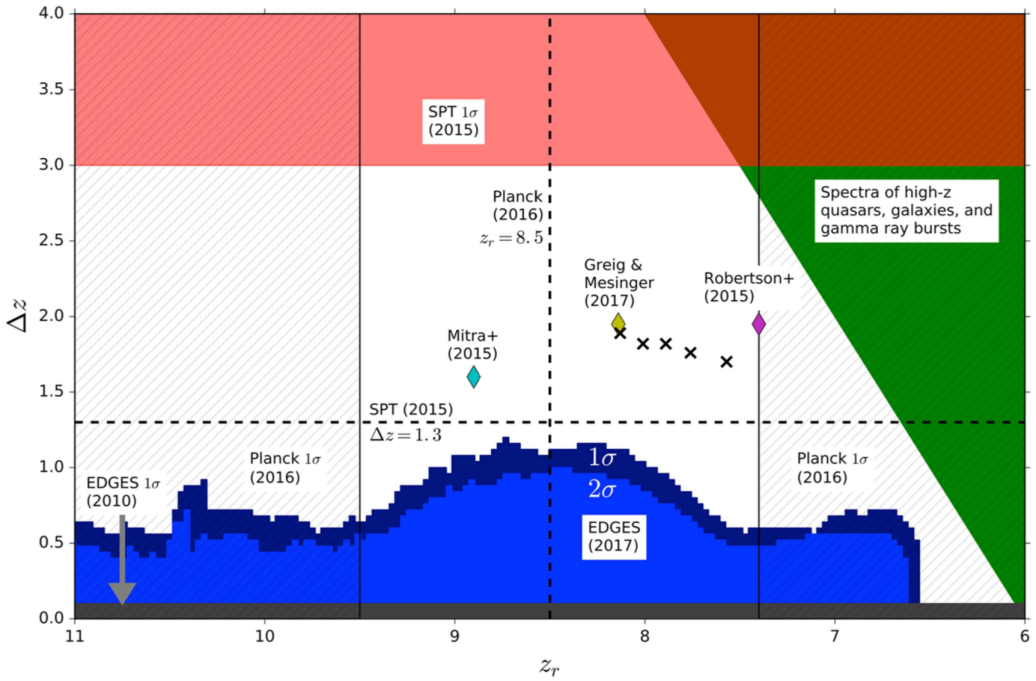

Figure 1.13: Monsalve et al. (2017) summary of reionization constraints on the midpoint of reionization (z_r) and on its duration (Δz). The specific constraints labelled here are described in the text.

light the times at which galaxies were building up. Oesch et al. (2016) find a bright galaxy at $z = 11.09$ using the observed Lyman break at 1.47×10^{-6} m, indicating that galaxies were already forming early in the EoR.

Figure 1.13 from Monsalve et al. (2019) shows a summary of constraints on the duration and midpoint of reionization. The constraints from EDGES and Planck are labelled, along with constraints from the South Pole Telescope (SPT), and those from high-z quasars as discussed earlier in this section. The specific estimates for reionization parameters indicated by markers are given by Mitra et al. (2015) using Planck data, and by Robertson et al. (2015) using joint Planck and Hubble Space Telescope data. The gold marker indicates the results of Greig and Mesinger (2017b) again using data from multiple sources: the 'dark fraction' in the Lyman alpha forest from quasars Mesinger (2018); the optical depth to the CMB using Planck 2016 data.

Chapter 2

Introductory material: models and methods

Modelling the 21cm signal involves accounting for complex non-linear reionization processes and solving radiative transfer equations. Three methods are currently used to generate predictions of the 21cm signal. Analytic models (Furlanetto et al., 2004) make predictions of the 21cm power spectrum by estimating the abundance and clustering of the ionising sources which give rise to ionised bubbles. Numerical simulations (Mellema et al., 2006a) make predictions using full N-body simulations including radiative transfer models to simulate the emission and absorption of ionising photons. Semi-numerical simulations such as SIMFAST21 (Santos et al., 2010) and 21CMFAST (Mesinger et al., 2011) use the approximation in Equation (1.21) to generate faster but less accurate predictions than full numerical simulations. In this chapter I describe the methods and models used throughout the other chapters. I first describe the semi-numerical simulations SIMFAST21 and 21CMFAST in detail in Section 2.1. Much of the recent literature has focused on these models, because they provide more detailed predictions than the analytic schemes and are more efficient than full numerical models. The simulations have been combined with sampling techniques such as Markov chain Monte Carlo (MCMC) methods to perform parameter estimation on mock interferometer data (Greig and Mesinger, 2015). Comparison between observations and these simulations allows us to extract the most likely reionization scenarios (see for example, Hassan et al. 2017,

Liu et al. 2016, Pober et al. 2016, Greig et al. 2016, Greig and Mesinger 2017a). In Section 2.2, I review the use of statistics in Cosmology, in particular the use of summary statistics for efficient comparison between data and theory. In Section 2.3, I review the Halo Model for the density field, which forms a basis for my halo model in Chapter 5. I then give a brief description of machine learning techniques in Section 2.4, as these are used in Chapters 4 and 6. Finally, in Section 2.5, I review some remaining recent literature that is relevant to the work in this thesis.

2.1 Modelling the Epoch of Reionization

2.1.1 Analytic model: Furlanetto-Zaldarriaga-Hernquist

Here I review the work of Furlanetto et al. (2004) which describes an excursion set formalism for the growth of ionised hydrogen regions during reionization. Determining the morphology of ionised regions is complicated because it requires locating and counting the sources of ionisation. Traditional cosmologists might have thought that the extend of ionisation near a galaxy depends only on the number of photons emitted by that galaxy. This is no longer thought to be true. The old-fashioned picture of many small ionised bubbles around each galaxy was quickly seen to be unrealistic, with the observed morphology from simulations showing a smaller number of large ionised regions around *clusters* of galaxies.

In the Furlanetto-Zaldarriaga-Hernquist (FZH) model, the intergalactic medium begins fully neutral. Discrete bubbles of fully ionised hydrogen grow around distinct sources. At first the sources are indeed assumed to be galaxies: the size distribution of sources is the same as the distribution of ionising galaxies, i.e. the galaxy mass function. The size distribution of ionised bubbles is found by assuming that the mass of an ionised region $M_{\rm ion}$ is directly proportional to the mass of the underlying galaxy $M_{\rm gal}$, connected by a constant efficiency parameter $\zeta_{\rm ion}$. This relation is written

$$M_{\rm ion} = \zeta_{\rm ion} M_{\rm gal}. \tag{2.1}$$

In order for an isolated region to become fully ionised, the enclosed luminous matter

must emit at least as many photons as there are neutral hydrogen atoms in the region. This condition is the same as requiring that the fraction of collapsed (luminous) matter $f_{\rm coll}$ is large enough to ionise the whole region, namely

$$f_{\rm coll} \geq \zeta_{\rm ion}^{-1}. \tag{2.2}$$

Using the Press-Schechter model for the fraction of collapsed matter with a smoothing scale M, the collapse fraction can be written

$$f_{\rm coll}(z,M) = 1 - {\rm erf}\left[\frac{\delta_c(z) - \delta}{\sqrt{2\left[\sigma^2(M_{\rm min}) - \sigma^2(M)\right]}}\right]. \tag{2.3}$$

The mass value $M_{\rm min}$ can be chosen as a minimum size of allowed ionised regions. The value δ_c is the critical overdensity for collapse of massive regions (often $\delta_c = 1.68$ is used, see Seljak 2000). The density threshold condition in Equation (2.2) for a region to fully self-ionise is then

$$1 - {\rm erf}\left[\frac{\delta_c(z) - \delta}{\sqrt{2\left[\sigma^2(M_{\rm min}) - \sigma^2(M)\right]}}\right] \geq \zeta_{\rm ion}^{-1},$$

which simplifies to $\delta \geq \delta_x$ with

$$\delta_x \equiv \delta_c(z) - \sqrt{2\left[\sigma^2(M_{\rm min}) - \sigma^2(M)\right]}\ {\rm erf}^{-1}(1 - \zeta_{\rm ion}^{-1}). \tag{2.4}$$

Measuring the abundances of differently-sized ionised bubbles would involve trying every possible smoothing scale and checking which regions of space pass the threshold condition. Note that the smoothing scale M must be chosen before testing the threshold. Also if one naively tested all smoothing scales from smallest to largest then one would over-count small ionised regions, as follows. Consider two nearby small regions which each have enough photons to self-ionise: both regions pass the threshold condition. But they would also pass the threshold condition for a larger region which encapsulates both smaller regions. Should the small regions be counted as two separate regions, or as one large region? This is the exact problem of the relative abundances of large and small regions as described at the start

of this subsection. The solution is equivalent to the excursion set formalism for determining the abundances of differently-sized halos. When searching for ionised bubbles, start by smoothing on the largest scales to see whether any regions pass the threshold condition. All points within the smoothing horizon of such regions are thereafter assigned to a single ionised bubble. This process is repeated for regions of decreasing size, until the minimum mass of ionising regions $M_{\rm min}$ is reached.

Theoretically the distribution of abundances for such ionised regions can be determined in the same way as the excursion set formalism, as the distribution of the first up-crossings of the density field above the threshold barrier $\delta_x(M,z)$ defined in Equation (2.4). It is simpler to find the distribution of first up-crossings around a linear approximation to this barrier by using the tangent to the curve at $\sigma^2(M)=0$. Defining B_0 as the limit of $\delta_x(M,z)$ as $\sigma^2(M)\to 0$, we find

$$B_0 \equiv \delta_c(z) - \sqrt{2}\mathrm{erf}^{-1}(1-\zeta_{\rm ion}^{-1})\sigma(M_{\rm min}). \tag{2.5}$$

Similarly if B_1 is the gradient of $\delta_x(M,z)$ as $\sigma^2(M)\to 0$, then

$$B_1 \equiv \frac{\mathrm{erf}^{-1}(1-\zeta_{\rm ion}^{-1})}{\sqrt{2\sigma^2(M_{\rm min})}}. \tag{2.6}$$

The resulting bubble size distribution is given by Furlanetto et al. (2004) as

$$M\frac{dn}{dM} = \sqrt{\frac{2}{\pi}}\frac{\bar{\rho}}{M}\left|\frac{d\ln\sigma}{d\ln M}\right|\frac{B_0}{\sigma(M)}\exp\left[\frac{B^2(M,z)}{2\sigma^2(M)}\right] \tag{2.7}$$

where $B(M,z)=B_0+B_1\sigma^2(M)$, and B_0 and B_1 are the limits defined in Equations 2.5 and 2.6. The function $\frac{dn}{dM}$ then gives number of ionised bubbles with masses between M and $M+dM$.

2.1.2 SIMFAST21 simulation

The formalism in the previous subsection describes how ionised bubbles can be located, by smoothing the collapse fraction $f_{\rm coll}$ on decreasing scales and assigning bubbles to any regions with high enough $f_{\rm coll}$. The bubble size distribution in Equation (2.7) was found by solving the first-up crossings above the density thresh-

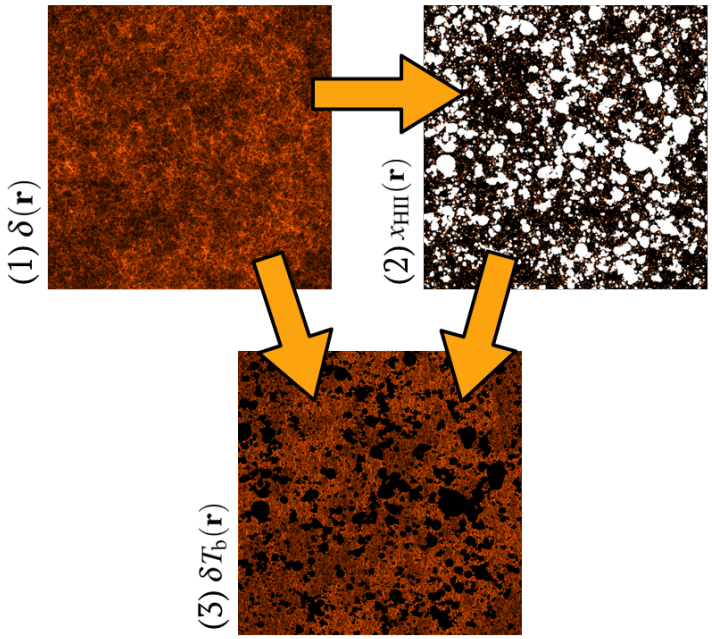

Figure 2.1: Images sliced through three of the main SIMFAST21 simulation outputs. The non-linear density field $\delta(\mathbf{r})$ is used to calculate the ionised fraction field $x_{\rm HII}(\mathbf{r})$. Both fields are then used to calculate the brightness temperature $\delta T_{\rm b}(\mathbf{r})$.

old analytically. Another method for locating and counting ionised regions is to simulate the same process, explicitly resolving the density field into pixels and determining where the smoothed collapse fraction exceeds the threshold barrier. After locating the ionised regions, Equation (1.21) can be used to make predictions of the 21cm brightness temperature. This is the basis for semi-numerical simulations such as SIMFAST21 and 21CMFAST. Both simulations start from a randomly-seeded density field and output predicted three-dimensional realisations of the 21cm brightness temperature field. The following are the main steps in these simulations, example outputs from which are shown in Figure 2.1.

1. Seed the initial density field onto a three-dimensional grid at high redshift;

2. Evolve the initial density field into a non-linear density field $\delta(\mathbf{r})$, using first-order perturbation theory from Zeldovich (1970). The output from this step is shown in panel (1) of Figure 2.1.

3. Determine the regions of the non-linear density field where the enclosed col-

lapsed matter contributes enough photons to self-ionise the enclosed neutral hydrogen. Starting with the largest and most massive regions, SIMFAST21 explicitly resolves individual dark matter halos using an excursion-set formalism (Furlanetto et al., 2004). This process is repeated for decreasing smoothing scales until the ionised regions are too small to be resolved by a single pixel. For regions which are smaller than the pixel size, SIMFAST21 uses an approximate ellipsoidal collapse method (Sheth et al., 2001). The smoothing scale is decreased until the minimum halo mass $M_{\rm min}$ in reached.

4. Generate the ionisation fraction field $x_{\rm HII}(\mathbf{r})$ from the results of the previous step. Regions with sufficient collapsed matter ($f_{\rm coll} \geq \zeta_{\rm ion}^{-1}$) to self-ionise are painted as fully-ionised with $x_{\rm HII} = 1$. Regions without sufficient collapsed matter are set to partially ionised based on the extent to which they fall short of the required threshold, giving $x_{\rm HII} = \zeta_{\rm ion} f_{\rm coll}$. The output from this step is shown in panel (2) of Figure 2.1.

5. Use Equation (1.21) to determine the 21cm brightness temperature field $\delta T_{\rm b}(\mathbf{r})$ from both the non-linear density field $\delta(\mathbf{r})$ and the neutral fraction field $x_{\rm HI}(\mathbf{r}) = 1 - x_{\rm HII}(\mathbf{r})$. The output from this final step is shown in panel (3) of Figure 2.1. SIMFAST21 also has the option to account for local fluctuations in the spin temperature at the expense of considerably more computation time, which is used only in Chapter 4.

2.1.3 21CMFAST simulation

21CMFAST follows a similar procedure to SIMFAST21 with a few key differences. The first difference is the method for calculating collapse fractions from the non-linear density field. By default 21CMFAST does not resolve individual halos, but rather calculates the collapse fraction directly from the non-linear density field following the model of spherical collapse (Press and Schechter, 1974). In order to match the more accurate ellipsoidal collapse model (Sheth et al., 2001), 21CMFAST afterwards normalises the spherical collapse fractions so that their average value matches that expected from ellipsoidal collapse.

The second difference is the method for calculating the ionisation fraction from the collapsed matter. Both simulations determine whether the collapsed matter in a region emits enough photons to ionise the surrounding matter. However, if there are enough photons then SIMFAST21 paints the entire spherical region as ionised using the fully overlapping-spheres method (Mesinger and Furlanetto, 2007), whereas the default 21CMFAST algorithm is to paint only the central pixel of the region (Zahn et al., 2006). The latter method is much faster but the algorithms give a considerably different reionization history for the same inputs (Hutter, 2018). 21CMFAST has an option to match the method of SIMFAST21.

The final difference is in the evolution of the parameter $M_{\rm min}$. The default 21CMFAST implementation allows the minimum halo mass $M_{\rm min}$ to evolve with redshift by setting a minimum virial temperature $T_{\rm vir}$ for ionising photons.

2.1.4 Three-parameter model

Three semi-numerical simulation parameters stand out as the most powerful ways to constrain reionization scenarios from data:

- $\zeta_{\rm ion}$ – the ionisation efficiency. This is common to both simulations and controls how many ionising photons are generated per unit of collapsed matter.

- $R_{\rm max}$ (SIMFAST21) or $R_{\rm mfp}$ (21CMFAST) – an upper limit specifying the maximum travel distance for ionising photons.

- $M_{\rm min}$ (SIMFAST21) or $T_{\rm vir}$ (21CMFAST) – a lower mass limit, specifying the minimum mass of collapsed matter which is needed to produce ionising photons. The default 21CMFAST implementation allows the minimum halo mass to evolve with redshift, by fixing a minimum virial temperature $T_{\rm vir}$ for ionising photons.

In Chapter 4, I additionally use the X-ray minimum energy value E_0, which is described in the subsections of that chapter.

2.1.5 Numerical simulations

The 21cm brightness temperature can also be modelled using more expensive numerical simulations. BEARS (Bubble Expansion Around Radiative Sources, Thomas et al. 2009) uses one-dimensional radiative transfer simulations to model the likely radiation profile around different sources. These profiles are then stamped onto sources to generate the predicted 21cm maps, identifying sources by their number of emitted photons and spectra. The C^2-RAY simulation (Mellema et al., 2006a) models the actual processes of emission and conservation of photons. Rays of ionising radiation are traced from sources and give rise to the ionisation bubble morphology discussed in Section 2.1.1. The Cosmic Reionization On Computers (CROC, Gnedin 2014) is another program of numerical simulations, accounting for cooling of hydrogen and helium with an Adaptive Refinement Tree code (Kravtsov et al., 1997) to allow for improved modelling of both large and small scale structures. The LICORICE code (Semelin et al. 2007, Semelin 2015) also uses a tree-based approach, with Monte-Carlo radiative transfer steps to compute the effect of reionization. Ross et al. (2016) present a similar numerical simulation also accounting for multi-frequency heating, both with and without X-ray sources. Finally, the Cosmic Dawn simulation (CoDa, Ocvirk et al. 2015) use numerical prescription to simulate the effect of reionization in the local universe. No numerical codes are used in this thesis and these descriptions are included for the sake of completeness.

2.2 Statistics in cosmology

Statistics is fundamental to data analysis in cosmology. We can only make observations of our single Universe — and only a small part of that Universe. An anomalous observation could be a random fluctuation or it could be a hint of new physics. Comparisons between theory and observations must be able to account for such fluctuations.

2.2.1 Bayes' theorem

Bayes' theorem,

$$P(\text{theory}|\text{observation}) = \frac{P(\text{observation}|\text{theory})P(\text{theory})}{P(\text{observation})}, \tag{2.8}$$

is used to determine the probability that a theory is correct given some observations. This allows us to compare theory and data in a quantitative and logically consistent way. The prior probability $P(\text{theory})$ encompasses all our prior knowledge about the Universe and how likely a new theory is given that prior experience. The model evidence $P(\text{observation})$ specifies the probability that the observation would have been made, by marginalising over all possible theories. The likelihood function $P(\text{observation}|\text{theory})$ specifies the probability of observing the data given predictions from a theory. Theories are often in the form of parameterised models such as those from ΛCDM. Given a theory which makes different predictions for these parameters, Bayes' theorem can be used to determine the most likely value of a particular parameter, given some new observations and some prior function. Contour plots such as Figure 1.12 earlier indicate the most likely regions of the parameter space. In that example, the ΛCDM model made predictions for the amplitude of the kSZ effect and the total integrated optical depth to reionization, allowing the midpoint $z_{\rm re}$ and duration Δz of reionization to remain as non-fixed parameters. Using the new observed data and Equation (2.8), the posterior probability $P(\text{theory}|\text{observation})$ is then found as function of the parameter values $z_{\rm re}$ and Δz. The parameter values most favoured by the data are those with the highest posterior probability. Posterior plots such as Figure 1.12 show the regions of parameter space with highest posterior probabilities. Contour panels such as the one at the bottom left show the posterior probabilities for each pair of parameters indicated by the relevant x-axis and y-axis labels, by integrating out all but two of the parameters. The one-dimensional line-plot panels along the top diagonal also show the posterior probability for each parameter individually by marginalising over all other parameters.

2.2.2 Summary statistics

In order to use Bayes' theorem one must be able to make quantitative predictions from a theory of what the observations *should be*. It is often not possible to make predictions for every individual observation, either due to time and memory constraints or due to the statistical nature of our theories. Summary statistics can be used both to characterise data that is too large to handle otherwise, and also to make predictions from statistical theories which do not make direct predictions for the observable data. For instance, our model for the growth of structure in the Universe does not make predictions for the locations and masses of every individual galaxy. Instead it makes predictions for the average separation and clustering of galaxies, if we were somehow able to 'run the experiment' many times and look at a large ensemble of universes with the same underlying physics.

Common summary statistics for 21cm data involve the size, shape and clustering properties of the ionised bubbles throughout reionization (see Shimabukuro et al. 2017a, Watkinson et al. 2017, Majumdar et al. 2018 and Watkinson et al. 2019). These features contain valuable information about the evolution of the Universe and about the reionization process. This subsection contains a review of mathematical descriptions for the statistical clustering properties of a continuous field $\delta(\mathbf{r})$. The clustering of a field is a function of scale and measures the clumpiness of the data. The two-point correlation function $\xi^{(2)}(\mathbf{r})$ is defined as the ensemble average over pairs of points in real space,

$$\xi^{(2)}(\mathbf{r_1}-\mathbf{r_2}) = \langle \delta(\mathbf{r_1})\delta(\mathbf{r_2})\rangle \tag{2.9}$$

The angular brackets denote an ensemble average over a large region of space (or over a large number of universe realisations) to mitigate the effect of statistical fluctuations. The spherically-averaged two-point correlation function is found by averaging over points which are separated by a given scale R, namely

$$\xi^{(2)}(R) = \left\langle \xi^{(2)}(\mathbf{r_1}-\mathbf{r_2})\right\rangle_{|\mathbf{r_2}-\mathbf{r_1}|=R}. \tag{2.10}$$

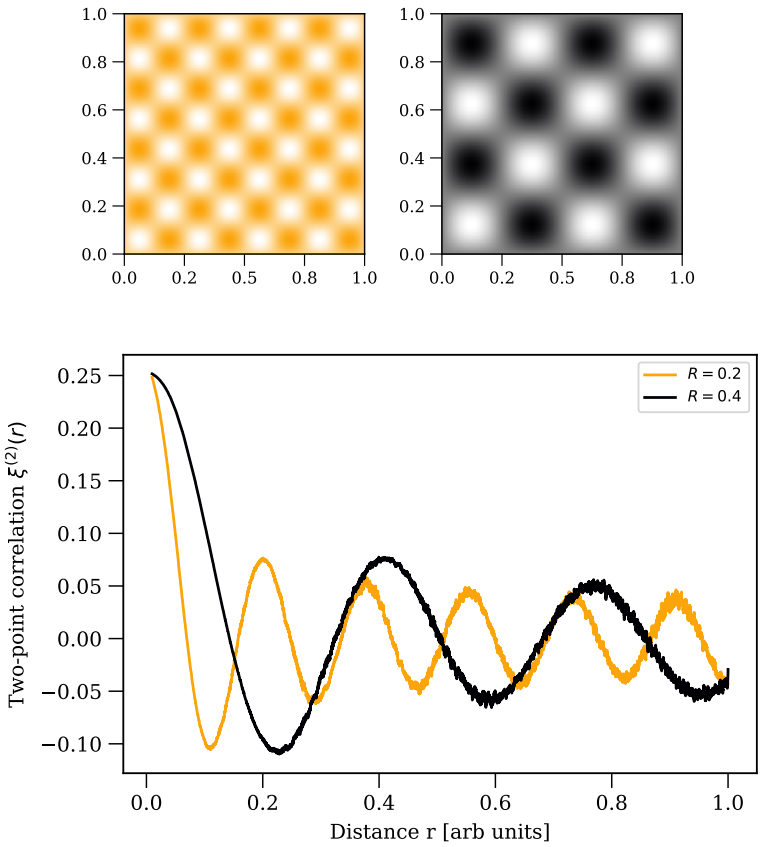

Figure 2.2: Example of clustering in two sinusoidal data fields. The top left panel has period of 0.2 and the top right has period 0.4, and the full box size is 1.0 in both cases. The two-point correlation function of these data are shown in the bottom panel. The correlation function has peaks near multiples of the clustering periodicity in each case, with the peaks of the larger-period data (black line) more spaced than the shorter-period data (orange line).

The spherically averaged two-point correlation function $\xi^{(2)}(R)$ describes the excess probability over a random field of finding similar-intensity regions separated by a distance R. If $\xi^{(2)}(R)$ is large for a scale R then any two locations with separation R in the measured field are more likely to have similar intensities than two locations in a random field. Figure 2.2 shows the two-point correlation functions for two example datasets. The datasets are sinusoidal with different periodicity. The data fields are shown along with their corresponding measured two-point correlation functions. The data with larger clustering scale has correlation function peaks that are spaced further apart.

Many of the theoretical expressions in this thesis simplify considerably in

Fourier space. I use the convention

$$\widetilde{\delta}(\mathbf{k}) = \frac{1}{(2\pi)^3} \int d^3\mathbf{k}\ \delta(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}} \tag{2.11}$$

for the Fourier transform of a continuous field $\delta(\mathbf{r})$. The two-point clustering in Fourier space is quantified as the power spectrum,

$$P(k)\ \delta_D^3(\mathbf{k_1} - \mathbf{k_2}) = \frac{1}{(2\pi)^3} \left\langle \widetilde{\delta}(\mathbf{k_1})\ \widetilde{\delta}(\mathbf{k_2}) \right\rangle, \tag{2.12}$$

where the Dirac-delta function $\delta_D^3(\mathbf{k_1} - \mathbf{k_2})$ enforces $\mathbf{k_1} = \mathbf{k_2}$. The power spectrum can be seen as the variance of the data field $\delta(\mathbf{r})$ on different scales k. If the power spectrum is high for a particular scale k then pairs of locations separated by that scale show a large variance in the data field values. If the power spectrum is low for a scale k then parts of the data field separated by that scale tend to be similar and have low variance.

The power spectrum $P(\mathbf{k})$ and the two-point correlation function $\xi^{(2)}(\mathbf{r})$ are Fourier-space pairs, so that

$$\begin{aligned} P(\mathbf{k}) &= \int d^3\mathbf{r}\ \xi^{(2)}(\mathbf{r})\ e^{-i\mathbf{k}\cdot\mathbf{r}} \\ \text{and } \xi^{(2)}(\mathbf{r}) &= \int \frac{d^3\mathbf{k}}{(2\pi)^3} P(\mathbf{k})\ e^{+i\mathbf{k}\cdot\mathbf{r}}. \end{aligned} \tag{2.13}$$

The three-point correlation function (3PCF) is similarly defined as the ensemble average of triplets of points in real space,

$$\xi^{(3)}(\mathbf{r_1}, \mathbf{r_2}, \mathbf{r_3}) = \langle \delta(\mathbf{r_1}) \delta(\mathbf{r_2}) \delta(\mathbf{r_3}) \rangle\,. \tag{2.14}$$

Note that the three vectors $\mathbf{r_1}, \mathbf{r_2}, \mathbf{r_3}$ connect three points in real space and so have a vector sum of $\mathbf{0}$, i.e. they form a closed triangle. In practice, $\xi^{(3)}$ measurements are actually made over *configurations* of triangles, i.e. over sets of unique triangle side-lengths. The 3PCF for a single triangle configuration is then an average over all triangles with side lengths (r_1, r_2, r_3), namely

$$\xi^{(3)}(r_1, r_2, r_3) = \left\langle \delta(\mathbf{r_1})\delta(\mathbf{r_2})\delta(\mathbf{r_3}) \right\rangle_{(|\mathbf{r_1}|,|\mathbf{r_2}|,|\mathbf{r_3}|) = (r_1,r_2,r_3)}. \tag{2.15}$$

The Fourier-space equivalent of the three-point correlation function is the Bispectrum,

$$B(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3)\ \delta_D^3(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3) = \frac{1}{(2\pi)^3} \left\langle \widetilde{\delta}(\mathbf{k}_1)\ \widetilde{\delta}(\mathbf{k}_2)\ \widetilde{\delta}(\mathbf{k}_3) \right\rangle. \tag{2.16}$$

2.3 Halo model for density field

In this section I review the work of Sheth et al. (2001) and Seljak (2000) which form a basis for my halo model in Chapter 5. The halo model has been successfully used to predict power spectra for the clustering of matter in the Universe (see for example Smith et al. 2002, Cooray and Sheth 2002). The halo model is an analytic model for the growth and evolution of structure in the Universe. The next subsections describe each one of the following required ingredients for a halo model:

- Halo profile function giving the density of matter around individual halos;
- Halo mass function number densities of halos as a function of mass;
- Clustering of halo centres.

2.3.1 Profile function

The density profile $\rho(\mathbf{r})$ of a halo specifies the density at all points $\mathbf{r}$ from the halo centre. In a very simple model dark matter halos might be thought of as a single point mass. In this case the profile is given by the Dirac-delta function $\rho(\mathbf{r}) = \delta_D(\mathbf{r})$, so that it is zero everywhere except $\mathbf{r} = \mathbf{0}$. In reality, halos have a non-zero size and shape. The most commonly-used halo profiles (Navarro et al., 1996) decrease with the radial distance from the centre according to

$$\rho(r) = \frac{\rho_s}{(r/r_s)\,(1 + r/r_s)^2}, \tag{2.17}$$

where r_s gives a characteristic scale radius and $\rho_s \propto \rho(r_s)$ is related to the density at the scale radius. Both parameters depend on the redshift and total mass of the halo. The total mass of the profile is usually defined by finding the total enclosed dark matter mass within the virial radius r_v,

$$M_{\rm halo} = \int_0^{r_v} \rho(r) 4\pi r^2 dr \tag{2.18}$$

$$= 4\pi\rho_s r_s^3 \left[\ln\left(1+\frac{r_v}{r_s}\right) - \frac{r_v}{r_s+r_v} \right]. \tag{2.19}$$

It is common to define a concentration parameter $c = r_v/r_s$ for these profiles, so that the profile can be parameterised by any pair of parameters from $\{\rho_s, r_s, c\}$. N-body simulations (Bullock et al., 2001) have shown that the concentration parameter is well approximated by the form

$$c(M,z) = \frac{c_0}{1+z}\left(\frac{M}{M_*(z)}\right)^{\beta}, \tag{2.20}$$

with best fit values $c_0 = 9$ and $\beta = -0.13$. The value $M_*(z)$ is the non-linear mass-turnover scale, with halos above this mass being more clustered than matter in general. In general the virial radius r_v is a function of the redshift (Cooray and Sheth 2002 Equation 141). However, for all redshifts of interest in this thesis the virial radius is given by the radius for which the profile density is $\Delta_{\rm vir} \approx 18\pi^2$ times the density of the background matter (see Cooray and Sheth 2002, Equation 52).

In my halo model for reionization, two spherically symmetric profiles are required. The density profile $\rho(r,M)$ specifies the total hydrogen density around halo centres. The ionisation fraction profile $\rho_x(r,M)$ specifies the ionisation fraction around halo centres.

2.3.2 Mass function

The abundance of dark matter halos can be modelled by the halo mass function $\frac{dn(M)}{dM}$, which gives the number density of halos with masses between M and $M+dM$. The number density of halos generally decreases with mass. The halo

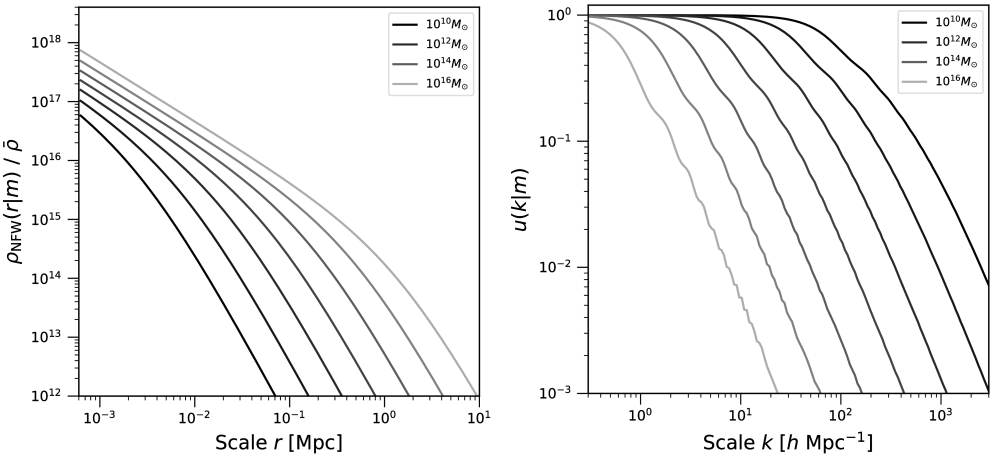

Figure 2.3: Navarro-Frenk-White profiles profiles for a range of halo masses for $z = 0$, shown in real-space (left panel) and Fourier-space (right panel). The functions $u(k|m) = \widetilde{\rho}(k|M)/M$ in the right panel are normalised to be unity as $k \to 0$.

mass function can be used to weight the relative clustering contributions from halos with different masses. Integrating over all possible halo masses with their relative abundances gives the total mass in the Universe. This gives a normalisation condition for the halo mass function as

$$\frac{M_{\rm tot}}{\text{Volume}} = \overline{\rho}$$
$$\int M \frac{dn(M)}{dM}\, dM = \overline{\rho}. \tag{2.21}$$

2.3.3 Two-point correlation function

The two-point correlation function of the density field is a sum of two distinct parts: contributions when the two points lie within the same halo, and contributions when the two points lie in different halos. These terms are written $\xi^{\rm 1h}(r)$ and $\xi^{\rm 2h}(r)$ respectively and are called the 1-halo and 2-halo terms. These terms can be calculated from the ingredients in the previous subsections, according to

$$\xi^{\rm 1h}(\mathbf{r}-\mathbf{r}') = \int dM\, \frac{1}{\bar{\rho}^2} \frac{dn(M)}{dM} \int d^3\mathbf{y}\, \rho(\mathbf{y};M)\, \rho(\mathbf{y}+\mathbf{r}-\mathbf{r}';M), \tag{2.22}$$

and

$$\xi^{\rm 2h}(\mathbf{r}-\mathbf{r}') = \int dM_1 \, \frac{1}{\bar{\rho}} \frac{dn(M_1)}{dM} \int dM_2 \, \frac{1}{\bar{\rho}} \frac{dn(M_2)}{dM} \int d^3\mathbf{y_1} \, \rho(\mathbf{r}-\mathbf{y_1};M_1) \int d^3\mathbf{y_2} \, \rho(\mathbf{r}'-\mathbf{y_2};M_2) \, \xi^{hh}(|\mathbf{y_2}-\mathbf{y_1}|;M_1,M_2). \tag{2.23}$$

Note that the 2-halo term additionally depends on the halo-halo correlation function $\xi^{hh}(r;M_1,M_2)$, which gives the correlation between the centres of halos.

2.3.4 Power spectrum

The 1-halo and 2-halo terms in Equations 2.22 and 2.23 can be seen as real-space convolutions of pairs of profiles $\rho(r,M)$. The convolution theorem allows us to write these integrals more simply in Fourier-space,

$$P^{\rm 1h}(k) = \int dM \, \frac{dn(M)}{dM} \frac{1}{\bar{\rho}^2} |\widetilde{\rho}(k;M)|^2, \tag{2.24}$$

and

$$P^{\rm 2h}(k) = \int dM_1 \, \frac{1}{\bar{\rho}} \frac{dn(M_1)}{dM} \int dM_2 \, \frac{1}{\bar{\rho}} \frac{dn(M_2)}{dM} \, \widetilde{\rho}(k;M_1) \, \widetilde{\rho}(k;M_2) P^{\rm hh}(k;M_1,M_2). \tag{2.25}$$

The Fourier-transformed profiles $\tilde{\rho}(k,M)$ are given by

$$\tilde{\rho}(k;M) = \int_0^{r_v} 4\pi r^2 \frac{\sin(kr)}{kr} \rho(r,M) dr, \tag{2.26}$$

also known as the Hankel transform (Weisstein, 2019). The correlation functions $\xi(r)$ have become power spectra $P(k)$ in Fourier-space, according to the convention in Equation (2.13). The total power spectrum is a sum of the two terms,

$$P(k) = P^{1h}(k) + P^{2h}(k). \tag{2.27}$$

2.3.5 Halo-mass bias relationship

The halo-halo power spectrum $P^{hh}(k;M_1,M_2)$ can be written (Seljak, 2000) as a separable function of the masses M_1 and M_2,

$$P^{hh}(k,M_1,M_2) = P^{lin}(k)b(M_1)b(M_2). \tag{2.28}$$

The bias function $b(M)$ accounts for the fact that halo centres preferentially pick out positions of high density contrast compared to the linear power spectrum $P^{lin}(k)$. For regions of high density contrast, this systematically biases the power spectrum to be of higher magnitude. Conservation of mass in the Universe previously given by Equation (2.21) is then rewritten with the bias function,

$$\frac{1}{\overline{\rho}} \int b(M) \frac{dn(M)}{dM} MdM = 1. \tag{2.29}$$

The 2-halo term can also be written more simply,

$$P^{2h}(k) = P^{lin}(k) \left[\int dM \, \frac{b(M)}{\bar{\rho}} \frac{dn(M)}{dM} \widetilde{\rho}(k;M) \right]^2. \tag{2.30}$$

2.3.6 Peak strength $\nu(M)$ formalism

In the Press and Schechter (1974) formalism, halo masses are often parameterised by introducing the peak strength

$$\nu(M) = \frac{\left[\delta_c(z)\right]^2}{\sigma^2(M)}. \tag{2.31}$$

The function $\delta_c(z)$ gives the critical density for collapse at the redshift of interest, and $\sigma^2(M)$ gives the current-day fluctuations in the matter field as a function of mass scale. The halo mass function $\frac{dn(M)}{dM}$ is then parameterised by the function $f(\nu)$ according to

$$\frac{dn(M)}{dM}dM = \frac{\overline{\rho}}{M(\nu)}f(\nu)d\nu. \tag{2.32}$$

Using this parameterisation, Sheth et al. (2001) give a fitted form for the halo mass function by simulating random density field fluctuations and determining the distribution for collapsed masses. The halo mass function is then written

$$\nu f(\nu) = A(1+\nu'^{-p})\nu'^{1/2}e^{\nu'/2} \tag{2.33}$$

with $\nu' = a\nu$ and best fit values $a = 0.707$ and $p = 0.3$, with $A = 0.1285$ to preserve the overall density of the Universe in Equation (2.29). Equation (2.33) is often known as the Sheth-Tormen (ST) mass function. The bias function also has a fitted form

$$b(\nu) = 1 + \frac{\nu' - 1}{\delta_c} + \frac{2p}{\delta_c(1+\nu'^{p})}. \tag{2.34}$$

I use the HMFCalc module (Murray et al., 2013) for generating values for these halo mass functions and bias functions at any redshifts.

2.4 Machine learning techniques

The machine learning techniques in this thesis are methods of multi-dimensional regression: learning the behaviour of some function $f(\mathbf{x})$ from noisy example training data $y_n = f(\mathbf{x_n}) + \text{Noise}$. After fitting, the models can make predicted evaluations $f(\mathbf{x}^*)$ at new input values $\mathbf{x}^*$. This section describes the different machine learning techniques in this thesis along with theoretical descriptions of their specific training methodologies. Each method learns the behaviour of some unknown function, for instance the SIMFAST21 power spectrum for any reionization scenario as specified by a continuous range of SIMFAST21 input parameters. The trained models can then make fast predictions for new function inputs, provided the new inputs do not lie far outside the range of the representative training data.

2.4.1 'Black boxes' and interpretability

Questions are often raised about the problem of interpretability of machine learning models. Whereas analytic models can immediately lead to increased understanding of the underlying physics, the process by which an AI model makes predictions can be convoluted and difficult to understand. Such models are often described as a 'black box', since the processes that occur to generate predictions are so obscure as to be effectively unintelligible. In this thesis, the issue of interpretability is generally less important: I use machine learning methods to replace existing simulations, by training them to emulate the exact behaviour of the simulations themselves. Provided that the final trained model accurately mimics the simulation behaviour, then we can interpret its predictions as if they came directly from the true simulations. However when the model fails to emulate the simulation behaviour perfectly, it can in general be extremely difficult to determine how to resolve the mismatch. For instance, in Section 6.3.6 I discuss the fact that for some scenarios the trained SIMFAST21 emulator fails to perfectly reproduce the simulated power spectrum predictions. In such situations, analytic models are often superior since the developer of the model can investigate and improve inaccuracies.

2.4.2 Interpolation

The simplest method for prediction is to interpolate the outputs within the training data. I use two interpolation methods, linear interpolation and nearest-neighbour interpolation, implemented using the classes LINEARNDINTERPOLATOR and NEARESTNDINTERPOLATOR from the SCIPY module (Jones et al., 2007). These methods involve no model choices and ignore the effect of sample variance noise in the training data. I include them as a naive benchmark to compare the accuracy and speed performance with the other models. The scipy LINEARNDINTERPOLATOR class uses QHULL (Barber et al., 1996) to triangulate the input data, computing five-dimensional surfaces in the input space and then performing linear interpolation on these triangles. This process takes a long time, both for training and prediction. The scipy NEARESTNDINTERPOLATOR class makes predictions by returning the output value from the nearest training data point. This process is very fast but generally

results in poorer predictions.

2.4.3 Multilayer perceptron

An artificial neural network (ANN) represents the function $f(\mathbf{x_i})$ by manipulating its input values $\mathbf{x_i}$ through a series of weighted summations and simple function evaluations. This series of repeated operations can be thought of as occurring in a series of layers. The values in the first layer $\mathbf{h^{(0)}}$ are simply the input values $\mathbf{x_i}$. The network manipulates the values from one layer $h_j^{(l-1)}$ to the values in the next layer $h_j^{(l)}$ using

$$\mathbf{h^{(l)}} = h_j^{(l)} = \phi_\theta \left(\sum_{i=1}^{N_i} W_{ij}^{(l)} h_j^{(l-1)} \right) . \tag{2.35}$$

The values in the l-th layer are a weighted sum over the values in the previous layer, using trainable weight values $W_{ij}^{(l)}$, and are then passed through an activation function $\phi_\theta(x)$. The final layer contains the network's fitted evaluations of the function, $f(\mathbf{x_i})$. Training the network requires finding the weight values $W_{ij}^{(l)}$ which most closely mimic the function's behaviour. Neural networks have been used previously for learning a variety of non-linear complex relationships in astrophysics. Use-cases include learning to extract redshifts from photometric measurements (for example Collister and Lahav 2003 and Sadeh et al. 2016); mimicking density field power spectra (Agarwal et al., 2014); classification of supernovae from light curves (Lochner et al., 2016); and classification of galaxy morphologies from images (Lahav, 1995).

Multilayer perceptrons (MLPs) are ANNs which contain at least one hidden layer and have a non-linear activation function. Figure 2.4 shows a schematic of a typical MLP's layer structure. Lines represent the weighted connections between values. Circles represent the neurons which schematically hold the values $h_j^{(l)}$ and pass the weighted inputs through the activation function. I use the SCIKIT-LEARN package (Pedregosa et al., 2011) for all MLPs.

MLP training involves finding the weight values $W_{ij}^{(l)}$ which minimize the objective function,

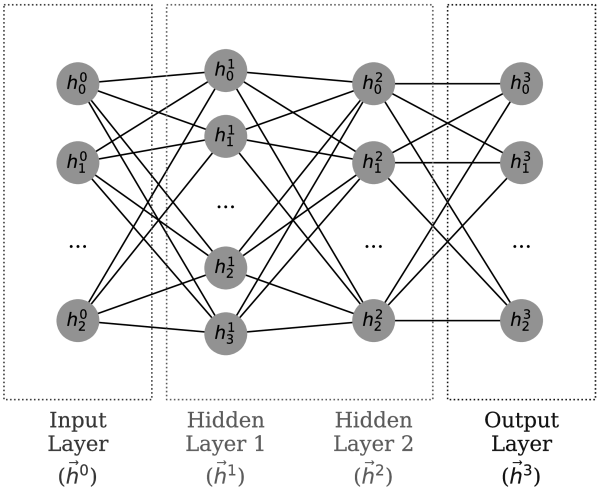

Figure 2.4: Visualisation of a multilayer perceptron with two hidden layers (Jennings et al., 2018). Lines are weighted connections directed from left to right. Circles are the neurons which schematically hold the values, pass the weighted sum of inputs through the activation function, and send this final value to the next layer.

$$\text{MLP Objective} = \frac{1}{2N}\sum_{n=1}^{N}\left(f(\mathbf{x_n}) - y_n\right)^2 - \frac{\alpha}{2}\sum_{i,j,l}\left(W_{ij}^{(l)}\right)^2 \tag{2.36}$$

for training data $(\mathbf{x_n}, y_n)$. The weights are initialised using a different random seed for each model. The function evaluation $f(\mathbf{x_n})$ in Equation (2.36) follows the procedure given in the previous subsection: passing the input values $\mathbf{x_n}$ through multiple layers of weighted sums and activation function evaluations. Before training, one must fix the number of hidden layers and the number of neurons in each hidden layer. The L2 regularization parameter can be used to reduce the effect of overfitting. The SCIKIT-LEARN class for MLP uses the backpropagation algorithm (Werbos, 1974) to update the weights towards the 'best' values that optimise the objective function. This involves efficient calculation of the gradient of the objective function, see Rumelhart et al. (1986) for a more detailed description of this algorithm. I use the 'adam' optimization method (Kingma and Ba, 2014) which terminates when the objective function falls below a tolerance of 10^{-10} for at least two consecutive iterations. The coarseness with which the weights are updated is controlled by a parameter known as the learning rate. A high learning rate means that the weights

are changed with a large magnitude at each step. The learning rate can be set to a constant value for all epochs, but it can also adapt to the current speed of learning. An adaptive learning rate decreases if the objective function plateaus (i.e. begins to fall slowly between epochs). It is common to set an upper limit for the number of epochs allowed for training. In my MLP models the weight values are initialised using the Xavier initialiation strategy (Glorot and Bengio, 2010). This method sets the weights in the i'th layer by sampling uniform values in the range $[-U_i, U_i]$. The normalising value $U_i = \sqrt{6}/\sqrt{n_i + n_{i+1}}$ is different for each layer, using values for the total number of input weight connections (n_i, also known as 'fan in') and output weight connections (n_{i+1}, also known as 'fan out').

Convolutional neural networks (CNNs) are another special case of neural networks, designed specifically to facilitate better modelling of images. CNNs take images as inputs and, rather than learning to understand the effect of each pixel individually, instead apply a series of 'filters' to each part of the image using a sliding window. The trained parameters are then the contents of these series of filters, which learn to pick out specific types of features from the training images. Figure 2.5 shows a schematic of this process. Convolutions allow CNN models to pick out features with translational and rotational invariance if the main object depicted in images often appears in different locations or at the different angles. For instance, the series of pictures depicting handwritten digits in the often-used MNSIT[1] dataset shown in Figure 2.6. The digits in these images are rotated and have varying degrees of thickness, making MNIST a common benchmarking dataset for analysing the quality of new methods. No convolutional neural networks are used in this thesis.

2.4.4 Gaussian processes regression

Gaussian process regression (GPR) is a fitting process for a function whose values are drawn from a Gaussian process. A Gaussian process is a set of random variables, any subset of which follow a jointly multi-variate Gaussian. For a finite set of D random variables stored in a vector $\mathbf{f} = [f_1, \ldots, f_D]$, the probability density function

[1] http://yann.lecun.com/exdb/mnist/

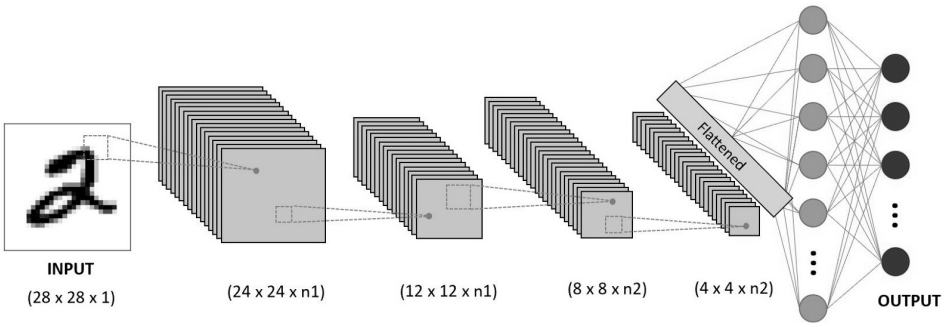

Figure 2.5: Schematic of a convolutional Neural Network from Sumit Saha (2018). The pixels in the image on the left are inputs to the model. The pixels are manipulated by placing a sliding window (indicated by the dotted squares) at all possible location in the image and a weighted sum of all pixels inside the window lead to a new (smaller) image. The window filter is defined by a large number of trainable weight parameters, and multiple filters can be used (see here as the first stack of blue images). This process is repeated iteratively, leading to continually smaller stacked images. Near the end of the network, the pixels in all images are flattened and used as the inputs to a standard ANN. Training involves finding the best weights values for all window filters and the final ANN connections.

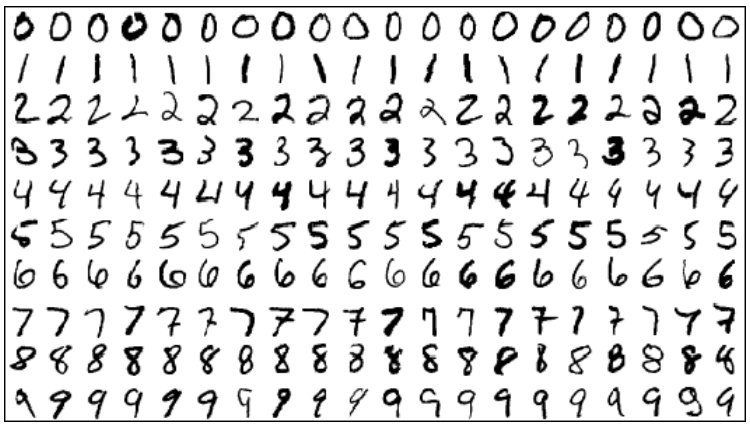

Figure 2.6: MNIST dataset examples of handwritten digits. The digits can appear shifted and at different angles. Convolutional Neural Networks such as the one in Figure 2.5 can be used to capture the translational and rotational invariance.

$P(\mathbf{f})$ of a multi-variate Gaussian has the form

$$\log P(\mathbf{f}) = -\frac{1}{2}\sum_{i,j=1}^{D}\left(f_i-\mu_i\right)K_{ij}\left(f_j-\mu_j\right) + \text{constant.} \tag{2.37}$$

Fitting this finite distribution involves finding the elements $\mu = [\mu_1,\ldots,\mu_D]$ of the mean vector, and the elements K_{ij} of the covariance matrix. A Gaussian process extends the concept of a multi-variate Gaussian to infinite dimensions, by replacing the finite-dimensional forms [$\mathbf{f}$, μ, K_{ij}] with functional forms [$f(\mathbf{x})$, $m(\mathbf{x})$, $k(\mathbf{x_i},\mathbf{x_j})$]. A Gaussian process can then be thought of as a distribution over functions, and training involves finding the optimal forms for the mean function $m(\mathbf{x})$ and a covariance kernel $k(\mathbf{x_i},\mathbf{x_j})$. Predictions are made by finding the function values which maximize the joint posterior of the training data and the new input values, all of which are assumed to be drawn from the same Gaussian processes. The choice of covariance kernel reflects the expected properties of the underlying process, such as smoothness or periodicity. Figure 2.7 shows an example of fitting a Gaussian process, where both the fitted mean function and covariance kernel have been shown.

Gaussian process regression involves finding the likelihood distributions of the mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x_i},\mathbf{x_j})$ which result analytically from the noisy training data. These likelihood distributions are combined with input prior distributions, to give the final posterior distributions from which predictions can be made. The prior for the mean function is often assumed to be linear,

$$m(\mathbf{x}) = \mathbf{A} + b\mathbf{x} \tag{2.38}$$

with trainable parameters $\mathbf{A}$ and b (initialised to zeros) specifying a linear relationship to each of the five input dimensions. The prior for the covariance function throughout this thesis is the MATERN32 kernel,

$$k_{\text{M32}}(\mathbf{x_i},\mathbf{x_j}) = \sigma^2\left(1+\frac{\sqrt{3}|\mathbf{x_i}-\mathbf{x_j}|}{\rho}\right)\exp\left(-\frac{\sqrt{3}|\mathbf{x_i}-\mathbf{x_j}|}{\rho}\right) \tag{2.39}$$

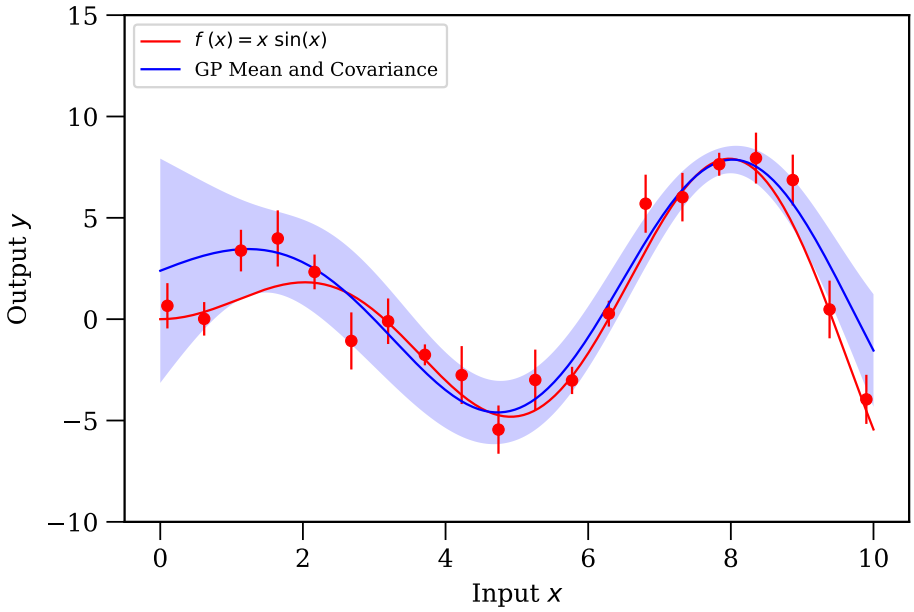

Figure 2.7: Example of Gaussian process regression on noisy data $y_n = x_n \sin(x_n) + \mathcal{N}(0, \varepsilon_n)$, with the noise amplitude on each data point ε_n being randomly drawn randomly from the interval $[0.5, 1.5]$. The mean function (solid blue line) and covariance kernel (shaded blue region) are found which best match the training data (red points).

with trainable parameters for the kernel variance σ^2 and kernel length-scale ρ (both initialised to unity). The MATERN32 is used to represent data with a moderate level of smoothing. Both of these kernel parameters control over-fitting. For instance, a smaller value of ρ allows the mean function to change more rapidly as a function of the inputs, which can cause the model to overfit the training data.

Training this model involves finding the matrix elements $K_{ij} = k(\mathbf{x_i}, \mathbf{x_j})$ of the training data. The expected mean and variance for a new prediction test location $\mathbf{x}^*$ are then given by

$$f(\mathbf{x}^*) = \sum_{i,j=1}^{N} k(\mathbf{x_i}, \mathbf{x}^*)\left(K_{ij} + \sigma^2 \delta_{ij}\right)^{-1} y_j \,, \tag{2.40}$$

$$\mathrm{Var}(f(\mathbf{x}^*)) = k(\mathbf{x}^*, \mathbf{x}^*) - \sum_{i,j=1}^{N} k(\mathbf{x_i}, \mathbf{x}^*)\left(K_{ij} + \sigma^2 \delta_{ij}\right)^{-1} k(\mathbf{x}^*, \mathbf{x_j}) \,, \tag{2.41}$$

(Rasmussen and Williams, 2006). Note that these equations involve inverting the large matrix $(K_{ij} + \sigma^2 \delta_{ij})$. In the case of Chapter 6, the matrix has 91000^2 elements. Using python 8-byte FLOAT64 values, simply storing a single object instance of this matrix takes 60GB of RAM. The computer architecture with 128GB of RAM is not large enough to invert such a matrix, since inversion requires much more RAM than a single matrix instance. Sparse Gaussian process regression (SGPR) is an approximation of GPR for huge data sets. SGPR approximates the matrix inversion by using only a subset of m observed data points and inverting this smaller matrix instead. These inducing points are effectively an additional set of fitting parameters. The SGPR models use the GPFLOW package[2] (Titsias, 2009) using TensorFlow (Abadi et al., 2016)). The GPFLOW package uses the SCIPY.OPTIMIZE.MINIMIZE function with the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method to find the best set of inducing points. The minimisation uses the default termination method, i.e. when the maximum component of the objective function's gradient falls below a tolerance of 10^{-5}.

2.4.5 Support vector machine

Support vector machine (SVM) models are often used for classification, but can also be used for regression. In SVM classification, training involves finding a set of hyperplanes which separate the training data into their labelled classes while at the same time maximising the distance between the hyperplanes and the nearest training data points. SVM regression extends this concept to functional forms, so that training the model involves finding a function $f(\mathbf{x})$ whose evaluations at the training points $\mathbf{x_n}$ are most similar to the observed training values y_n, while at the same time ensuring that the function is as simple as possible. I use the SCIKIT-LEARN package (Pedregosa et al., 2011) for the support vector machine models.

SVM training involves finding the functional form $f(\mathbf{x})$ such that the residual errors between the training data $(\mathbf{x_n}, y_n)$ and the function evaluations $f(\mathbf{x_n})$ all lie within some tolerance $-\varepsilon \leq f(\mathbf{x_n}) - y_n \leq \varepsilon$. This stringent constraint usually makes it impossible to find any such form $f(\mathbf{x})$. To weaken the condition and allow a

[2] http://gpflow.readthedocs.io/en/latest/intro.html

solution, the slack variables (ξ_n, ξ_n^*) are introduced so that the residual fitting error $f(\mathbf{x_n}) - y_n$ for the training point $(\mathbf{x_n}, y_n)$ obeys $-\varepsilon - \xi_n^* \leq f(\mathbf{x_n}) - y_n \leq \varepsilon + \xi_n$. This optimization problem is more easily solved in the dual form, with objective function

$$\begin{aligned} \text{SVR Objective} = & \sum_{i,j=1}^{N} \left(\alpha_i - \alpha_i^*\right) k(\mathbf{x_i}, \mathbf{x_j}) \left(\alpha_j - \alpha_j^*\right) \\ & + \varepsilon \sum_{i=1}^{N} \left(\alpha_i + \alpha_i^*\right) - \sum_{i=1}^{N} \left(y_i(\alpha_i - \alpha_i^*)\right). \end{aligned} \tag{2.42}$$

Training involves finding the values (α_i, α_i^*) which minimize this objective function, subject to margin constraints

$$\sum_{i=1}^{N} \left(\alpha_i - \alpha_i^*\right) = 0 \quad \text{and } 0 \leq \alpha_i, \alpha_i^* \leq C. \tag{2.43}$$

The kernel function $k(\mathbf{x_i}, \mathbf{x_j})$ in Equation (2.42) controls the functional form $f(\mathbf{x})$. I try three different kernel functions: radial basis function (RBF), polynomial, and sigmoid. The RBF kernel,

$$k_{\text{RBF}}(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\gamma \left|\mathbf{x_i} - \mathbf{x_j}\right|^2\right). \tag{2.44}$$

is infinitely differentiable, hence is often used to model data from smooth distributions. Before training, one must set the penalty term C, the kernel influence range γ (hereafter written GAMMA to match the python class parameter), and the margin tolerance ε (written EPSILON). Overfitting for SVR models is discouraged by C the penalty term, since higher C values give rise to overfitting and lower C values give rise to underfitting.

2.4.6 General training methodology

In this subsection I describe the general methodology for machine learning projects. Standard practice is to split the available data into two distinct parts. The first (usually larger) part is used to train the model, and the second part to test the model's

performance on unseen data. These two data sets are known as the testing and training datasets respectively.

Many machine learning algorithms have a set of 'hyperparameter' choices that must be made before starting to train the model. Such choices can have a strong effect on the final accuracy of predictions, but it is rarely obvious at the start of a project what hyperparameter values will result in the most accurate model. Instead, one normally tries a range of models with many different hyperparameter values and selects the model with highest performance. I mention two common search strategies for comparing models with different hyperparameters: an exhaustive grid search; and a random search.

The most basic strategy is to try an exhaustive list of hyperparameters combinations. Many models have tens or hundreds of possible hyperparameter choices, and each parameter can take one of many different values. Instead of comparing all possible combinations of all hyperparameters, one normally chooses a subset of the hyperparameters that are expected to have the largest effect on the prediction outcome. By training models on all different combinations of choices, the best model can be selected and retained for future use. Clearly this method becomes infeasible for more than a handful of hyperparameter choices, as the required number of models can quickly become computationally impossible.

A more efficient method is to choose the hyperparameters randomly within defined ranges. For this strategy, one needs to choose the allowed range of values for each hyperparameter that will be varied. For instance, one might select the range of hidden layer sizes in a neural network, or the allowed kernels for a Gaussian processes regression model. By choosing random hyperparameters, one can choose exactly how many models to run. The choice of how many models to run is then a balance between the available resources and the quality of the final model: if too few models are trained. then the hyperparameter space will be poorly sampled; but training too many models is much more expensive. This strategy generally allows for a much wider range of models to be compared than the exhaustive search.

By trying a range of different hyperparameters, one can usually find a model

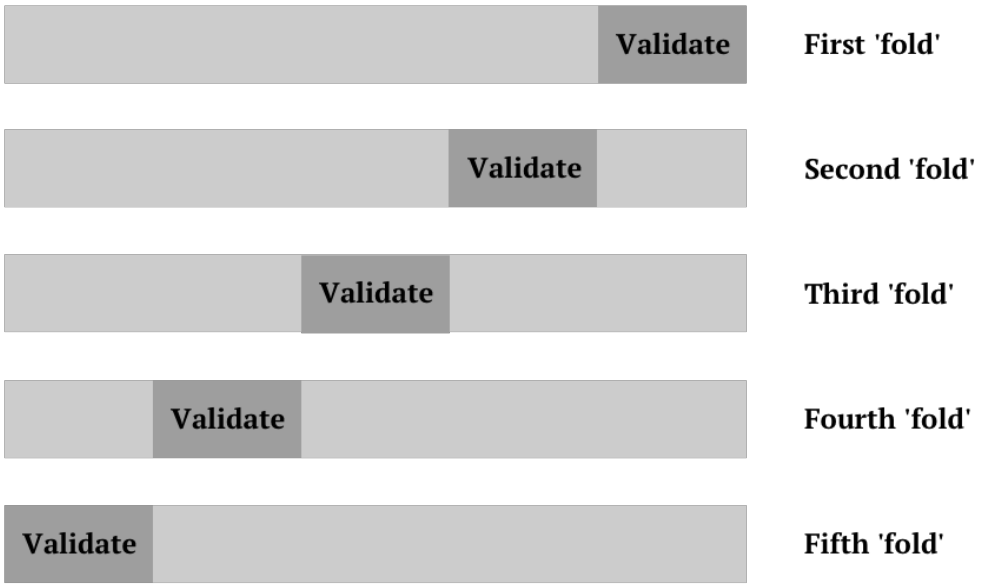

Figure 2.8: Schematic of five-fold cross-validation over a single dataset. The model is trained once using four fifths of the data, excluding the remaining one fifth. This excluded segment is used to calculate the model performance on unseen data. The model is trained four more times, each time excluding a different one fifth of the training data. The average performance over the five folds gives a measure of the model's overall performance, including its ability to extend to new unseen data.

with better prediction accuracy. However this process is sensitive to overfitting: one might find a model which performs very well on the training data, but makes poor predictions on new unseen data. In order to determine which model has the highest accuracy while reducing the chances of overfitting, a cross-validation approach can be used. Cross-validation splits the training data into a number of segments or 'folds', as shown in Figure 2.8. In this figure, five separate models are trained with the same fixed hyperparameters, where each model is provided with data from only four of the five folds. The fifth excluded fold is used to calculate the prediction performance each model. Measuring the performances on unseen data means that the best model is one which extends well to new data. The overall accuracy score is usually taken as the the mean of the validation scores.

2.4.7 Input and output scaling

Data scaling can be used to improve the efficiency of artificial neural networks during training, and also to improve the quality of the final predictions. The weight values in many neural networks are initialised at small values, often usinng the Xavier method described in Section 2.4.3. In general, different input features into a model have different scales and magnitudes. Ideally all inputs into the network would have similar orders of magnitude and simple distribution such as normal or uniform. This can easily be achieved by separately normalising or standardising each input feature. Normalising an input feature forces all values to lie in the range $[0,1]$ by using the minimum and maximum feature values. The normalised features x' are then related to the original features x according to $x' = (x - \min(x))/(\max(x) - \min(x))$. Standardising an input feature scales the feature to have a mean of zero and a standard deviation of 1.0, i.e. $x' = (x - \text{mean}(x))/\text{std}(x)$.

Scaling the model output value(s) also has a beneficial effect on the final prediction accuracy. Training a model involves minimising some objective function to match the predicted outputs $f(\mathbf{x_n})$ with the true output values in the training data y_n. Data points with large output values y_n will contribute disproportionately to this objective function compared to data points with small output values. This may be desirable for some applications, but in general scaling the output values using normalisation or standardisation can help mitigate the relative importance of output values with different magnitudes.

2.4.8 Analysing model performance

After training a model, we almost certainly wish to know how accurately it makes predictions for new data. Standard practice is to hold back a random representative sample of the available data for testing the final model. This sample is known as the 'testing dataset' or 'holdout set'. For each measurement in the testing dataset, the model can be used to predict what the relevant outputs *should be*. These predictions can then be compared the the *actual* observed values. Such comparisons are often made using summary metrics or illustrative plots. I mention a few standard methods here.

The accuracy of a model's predictions can be most easily quantified by using a single numerical accuracy metric. This number should quantify the similarity of the model's predicted outputs (often written $\hat{y}_i$) to the true outputs (y_i) observed in the testing data. Common metrics are the mean squared error (MSE),

$$\mathrm{MSE}\big[\mathbf{y},\hat{\mathbf{y}}\big] = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2 \tag{2.45}$$

and the mean absolute error (MAE),

$$\mathrm{MAE}\big[\mathbf{y},\hat{\mathbf{y}}\big] = \frac{1}{N}\sum_{i}^{N}|y_i - \hat{y}_i| \tag{2.46}$$

along with the derived value for the root mean squared error, $\mathrm{RMSE} = \sqrt{MSE}$. Both use the absolute difference $|y_i - \hat{y}_i|$ between the observed and predicted output values.

These single-number metrics are useful guidelines and are often quoted as the final measure of a model's performance. However, they provide no further insight in how the model can be improved. Often the performance of a model is strongly correlated with the input or output values: the model might make poorer predictions for large values of one of the outputs, or might struggle when the inputs are outside a certain range. In this case, prediction plots can provide valuable insight into the model's behaviour. In the following subsections I describe two common prediction plots used in this thesis: predicted vs true plots and error histograms.

Predicted vs true plots

Plot the predicted output values as a function of the true output values in the testing data can illustrate which output regimes the model struggles with the most. Figure 2.9 shows an example of one such plot, copied from Figure 4.4 later. For a perfect model, all predicted values would exactly equal the true values. The predicted vs true plot for such a model would have all points lying on the main diagonal. Any deviations from this diagonal indicates poorer predictions and, in particular, the scatter of points around the diagonal can highlight which output regions are least accurately predicted. In Figure 2.9 for instance, the model performance depends

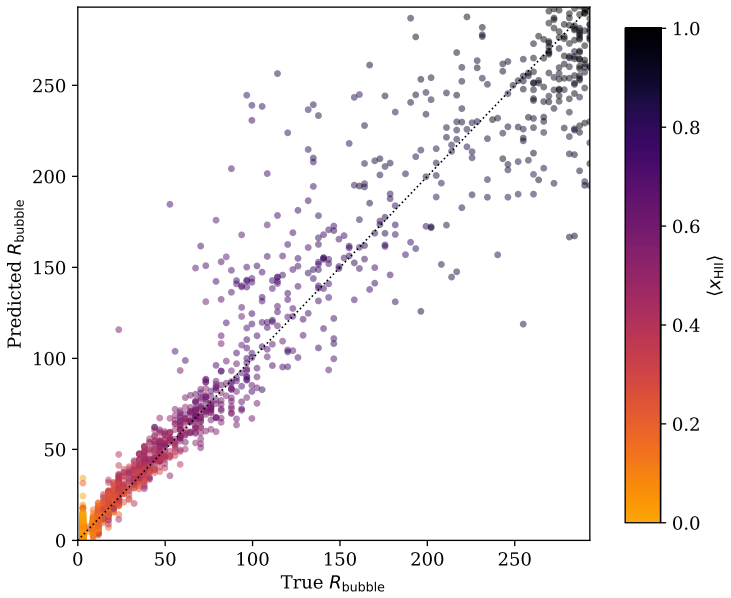

Figure 2.9: Example plot of predicted vs true values for a model predicting the typical bubble size of ioinised regions. The model accuracy is strongly dependent on the output value: low values of R_{bubble} have accurate predictions that that lie along the main diagonal (shown as a dotted black line); high values of $R_{\text{bubble}} >$ 100Mpc have much worse predictions accuracies.

strongly on the typical bubble size: for larger typical bubbles ($R_{\text{bubble}} > 100$Mpc) the predicted values are on average significantly further from the diagonal, whereas the higher accuracy predictions for smaller bubbles ($R_{\text{bubble}} < 100$Mpc) lie closely along the main diagonal. Knowing which output regimes are harder to model allows one to attempt to refine the model, either by retraining the model with different choices or by gathering more data for those situations.

Histograms of errors

The single-value metrics in Equations 2.45 and 2.46 above summarise the overall error of the model's predictions. Instead, one can plot a histogram of the full distribution of individual errors for each measurement in the testing data. Plotting the full distribution provides a more in-depth illustration of the performance. The top panel of Figure 2.10 shows example typical error histograms for two models, one of which is more accurate than the other. The x-axis shows a range of RMSE values, and the y-axis demonstrates the frequency of predictions in bins of RMSE values. For the more accurate model in this figure (black line), almost all predictions errors

are below 20%. For the less accurate model (purple line), the prediction errors are spread to much higher values of 60% RMSE. Error histograms can also highlight whether the model has outlier predictions. The bottom panel of Figure 2.10 shows a typical histogram for a model with a significant number of outlier predictions. The outliers are seen as a secondary peak at RMSE of 60%. Investigating the origin of these outlier points could potentially reveal the reason for their poorer performance, and indicate what new data or modelling choices could be used to improve the model accuracy. All histograms in this subsection were generated for illustrative purposes only, using fake prediction data.

2.5 Review of recent literature

In this section I review recent publications that are relevant to the remaining chapters in this thesis, highlighting how my work fits into the existing literature. Section 2.3 already gives a detailed review of the halo model which is the starting point for the analytic model in Chapter 5. The following subsections review current literature for the other chapters.

2.5.1 21cm tomography

Simulated maps of the 21cm signal show a wealth of complex structure. The simulated $\delta T_{\rm b}(\mathbf{r})$ signal shown in Figure 2.1 is the result of complex interactions between many non-linear physical processes. Observations from upcoming experiments such as the SKA will for the first time be able to generate similar maps for the actual Universe (Bacon et al., 2018). Ionised regions grow continually during the EoR, so that larger regions are more frequent at later times than at earlier times. Measurements of the bubble size distribution over a range of redshifts can indicate the speed of ionised bubble growth. Bubble size distributions can thus constrain any model parameters which affect the duration of the EoR, such as the ionising efficiency $\zeta_{\rm ion}$ and the minimum halo mass $M_{\rm min}$ with high enough star formation rate to produce ionising photons.

Many recent publications have considered how to analyse the complex structure within 21cm data. In particular, four different techniques are commonly used

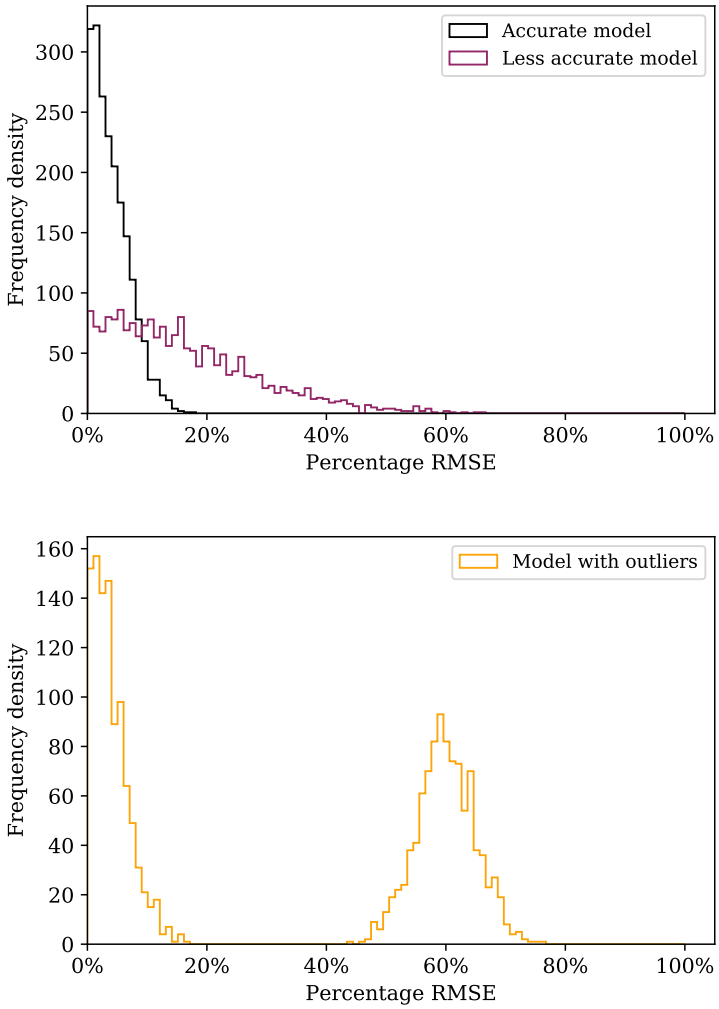

Figure 2.10: Top panel: typical error histograms for more accurate (black line) and less accurate (purple line) models. For the accurate model, almost all predictions errors are below 20%. For the less accurate model, the prediction errors are spread over a wider range of values, with fewer low-error predictions and more high-error predictions. Bottom panel: typical error histogram for a model with outlier predictions. Many of the predictions have high accuracy, but a significant fraction of the predictions have error near 60%.

to measure the size distribution of ionised regions: the mean free path method (Mesinger and Furlanetto, 2007); the spherical averaging method (Zahn et al. 2006, Mcquinn et al. 2006); the friends-of-friends method (Iliev et al. 2005); and the watershed algorithm (Lin et al., 2015). The mean free path method simulates the emission of ionised photons from within ionised regions, allowing these rays to propagate until they are absorbed by a neutral region. The distribution of the distances travelled by these photons measures the distribution of ionised bubble sizes. The spherical average method finds spherical regions whose average ionisation fraction exceeds some fixed threshold. Spheres of decreasing size are used and the bubble distribution calculated by counting the fraction of pixels lying within bubbles of each size. Lin et al. (2015) analyse these first two methods and find that they are biased: they give inconsistent results for different datasets with the same underlying bubble distribution. Instead, Lin et al. (2015) propose an adapted form of the 'watershed' algorithm used in 2D image analysis. In this algorithm, pixels with similar values are connected and effectively treated as the contours of a 3D tomographic terrain map. A rising threshold (akin to a rising water-level in the terrain metaphor, hence 'watershed') is used to segment the data into isolated regions, and the distribution of resulting regions used as a measure of the bubble size distribution. The friends-of-friends method is similar, linking together neighbouring pixels if they both exceed an ionised fraction threshold (often $x_{\rm HII} = 0.5$). The bubbles size distribution is then found by the distribution of groups of these connected pixels (or 'friends'). Kakiichi et al. (2017) present a new granulometry technique for measuring the full size distribution of ionised regions from these data, using a mathematical concept similar to 'sieving' the data (Matheron, 1974). Chapters 3 and 4 represent a new way of probing similar information to these metrics. I find that the 3PCF of the 21cm signal encodes information about the bubble size distribution, although it would be interesting to extend my models and attempt to fully recreate the bubble size distributions resulting from the techniques mentioned here.

Classifying ionised regions from $\delta T_{\rm b}(\mathbf{r})$ maps is a difficult process. Other properties such as the density field and spin temperature can cause local fluctua-

tions that masquerade as ionised regions. Normally a fixed threshold of $\delta T_{\rm b} = 0$ is used, so that regions with $\delta T_{\rm b} < 0$ are assumed to be ionised. Giri et al. (2018b) investigate a new technique for choosing this $\delta T_{\rm b}$ threshold. The distribution of $\delta T_{\rm b}$ pixel values in simulated maps are typically observed to be bimodal. Giri et al. (2018b) find that choosing a $\delta T_{\rm b}$ threshold to separate the two peaks gives recovers the bubble size distribution more better than using a simple $\delta T_{\rm b} = 0$ threshold. They use a K-means clustering algorithm to find the two peaks and choose a threshold halfway between. In Chapter 4, I use the simplest mean-free path method to calculate the bubble size distributions. Other methods could potentially provide more robust measurements for the typical bubble size. I use this method only on ionisation fraction data, avoiding the issues with choosing a $\delta T_{\rm b}$ threshold.

2.5.2 Ionisation fraction history

The global ionisation fraction $\langle x_{\rm HII} \rangle$ as a function of redshift parameterises the evolution history of the Epoch of Reionization. The exact timeline of $\langle x_{\rm HII} \rangle (z)$ remains poorly constrained by observations, as shown by the wide range of redshift values in Figure 1.7. In Figure 6.18, I present results that the ionisation fraction history from semi-numerical simulations depends strongly on the choice of minimum halo mass parameter. In fact, a wide range of cosmological and reionization model choices can affect the ionisation fraction history. In Chapter 4, I show that the morphology of ionisation fraction and 21cm differential brightness temperature fields encodes information about the global ionisation fraction. Observations of distant quasars indicate the redshifts at which the Universe still contains neutral hydrogen. Current observational results indicate that the EoR was completed by around $z \approx 6$ (Becker et al. 2001, Bolton et al. 2011). CMB anisotropies (Planck Collaboration, 2018) can also put a limit of the total duration of the EoR, since free electrons released during the EoR can interact with CMB photons and imprint a signal. Current observations put an upper limit on the duration $\Delta z < 4$ of the EoR.

2.5.3 Clustering statistics

One of the most powerful statistical tools for analysing the 21cm signal is the power spectrum. The power spectrum encapsulates clustering information on both small and large scales. The power spectrum of the 21cm differential brightness temperature (often referred to as 'the 21cm power spectrum') will be measured from upcoming interferometer experiments. A wide variety of reionization model choices affect the clustering of the 21cm signal. Cosmological parameters can be constrained using the high-redshift 21cm signal between $z \approx 500 - 1000$ (Fialkov and Loeb, 2013). The total mass of neutrinos (Pritchard and Pierpaoli 2008, Mao et al. 2008) and Dark Energy (Wyithe et al., 2007) can also be constrained. The 21cm signal in the range $z = 7 - 40$ can provide clues as to the character of heating sources at these times. Reionization model effects such as the extent of X-ray heating of gas in the IGM (Shimabukuro et al., 2015), and baryon-dark matter scattering (Fialkov et al., 2018) can also be measured. The position-dependent power spectrum (Giri et al., 2018a) can also be used to measure non-Gaussianities in the 21cm signal. Gravitational instabilities, for instance, cause non-Gaussian clustering in the density field from which ionising sources are seeded, and could imprint a signal onto the 21cm signal.

A more direct probe of non-Gaussianities in the 21cm signal is the bispectrum. Many recent publications investigate the potential benefits of the 21cm bispectrum, including efficient calculation algorithms (Watkinson et al., 2017) from observed data. Although the signal-to-noise requirements are much higher (Yoshiura et al., 2015) for the bispectrum than for the power spectrum, it has the potential to constrain more strongly several astrophysical effects such as the Wouthysen-Field (WF) effect (Shimabukuro et al., 2017a) and X-ray heating processes (Watkinson et al., 2019). Majumdar et al. (2018) determine that the sign of the bispectrum measures the relative importance of non-Gaussian fluctuations in the neutral fraction field to fluctuations in the matter density field. A positive bispectrum implies that the matter field is dominant, and a negative sign indicates that the neutral fraction (or ionisation fraction) field is dominant. In Chapter 4, I measure the 3PCF of the 21cm

differential brightness temperature and find that high-order clustering encodes similar information to the bubble size statistics. It would be interesting to investigate further whether the sign of these 3PCF measurements also contains information about the sources of non-Gaussian behaviour.

Visualising and understanding the bispectrum is difficult. See Figures 1 and 2 of Watkinson et al. (2019) for excellent visualisations of how the bispectrum should be interpreted. Instead, the three-point correlation function can be used to measure the equivalent clustering in real-space. This is the basis for Chapter 3. Measuring in real-space distances (instead of Fourier-space modes) gives a more intuitive understanding of the relationship between clustering amplitude and physical effects, although the bispectrum is more immediately available from interferometer experiments. Several recent papers have investigated using the 3PCF with 21cm data. Hoffmann et al. (2018) investigate whether the 21cm 3PCF can be modelled using a local bias model, eventually making predictions with around 20% accuracy for large ionised regions at early times. Their model breaks down for other scenarios. This accuracy is comparable to that of my machine learning model in Chapter 4, although my model has more consistent accuracy across a wider range of scenarios. Gorce and Pritchard (2019) use a derived statistic from the 3PCF to concentrate on phase information, finding that their statistic can be used to model the size of isolated neutral regions near the end of the EoR. Their tool can also be used to determine the length of the overlap stages of the EoR, since the signal remains constant during this time. The earliest stages of the EoR are more difficult to model, and they present only an upper limit for the size of ionised regions. My model in Chapter 4 has similar issues with modelling the earliest stages of the EoR.

2.5.4 Applications of machine learning to EoR data

Machine learning techniques are becoming an increasingly popular method of astrophysical data analysis. With the possibility of huge datasets from international collaborations such as Planck and the Square Kilometre Array, understanding and interpreting the results of these experiments is more and more challenging. Within the community of EoR research, machine learning has been suggested for a number

of uses, discussed here.

A major benefit of machine learning is the speed of the final models. Numerical and semi-numerical simulations such as SIMFAST21, 21CMFAST and C^2-RAY require large computational resources (such as time, memory, and CPU). Machine learning techniques can dramatically reduce the need for further computations in an application known as emulation. Emulators are trained to learn the simulation behaviour and replace the need for any further simulations. The models must be fitted to a representative training set of pre-computed simulations. The ultimate aim of such an approach would be to train models to mimic the more accurate numerical simulations, allowing for more accurate parameter estimation with Markov chain Monte Carlo (MCMC) analysis.

To date, several papers have attempted to emulate semi-numerical reionization simulations with moderate success. Schmit and Pritchard (2018) use neural networks to learn the power spectrum outputs of 21CMFAST simulations, training their model on 100 simulations and finding accurate predictions for the power spectrum. These models differ from my models in Chapter 6 in that they only make predictions for fixed redshifts and fixed k-scales. My models are more flexible and can make predictions for any new redshift or k-scale but, despite the fact that my models are trained using SIMFAST21 data with a much larger training set of 1000 simulations, the added complexity of including z and k as inputs causes my models to have an accuracy that is somewhat lower than those of Schmit and Pritchard (2018). The algorithm for SIMFAST21 differs slightly from 21CMFAST, see Sections 2.1.2 and 2.1.3 for a review of these differences. In a similar application, Kern et al. (2017) use a Gaussian Processes regression model to emulate the 21cm power spectrum from 21CMFAST. They use a data compression technique known as Principle Component Analysis (PCA) and find a model which can accurately infer the reionization parameters for unseen testing data, by using MCMC analysis.

Instead of training emulators and running MCMC analysis to derive the most likely regions of parameter space, Shimabukuro and Semelin (2017) train an inference model to map from power spectrum outputs directly to the model parameters.

Note that this is the reverse of my emulators presented in Chapter 6, which map from the model parameters to the power spectrum outputs. The model in Shimabukuro and Semelin (2017) recovers the EoR parameters for most situations, although it struggles to predict the values of the mean free path of ionising photons at early redshifts, whereas my models for the power spectrum can be used in any situation to replace the original simulation. Additionally, the model is only useful at making point estimates of the best parameters. It is not possible, for instance, to get uncertainty estimates on parameters from these models by using MCMC analysis.

Finally, machine learning has been used to derive the reionization parameters directly from 21cm images. For this application, Gillet et al. (2018) use a convolutional neural network like the one described in Section 2.4.3 earlier. This model uses images of the 21cm signal to perform parameter inference in the same way as Shimabukuro and Semelin (2017): by mapping directly from the simulation outputs to the most likely parameters. However, the model in Gillet et al. (2018) uses 2D images of the 21cm signal instead of the power spectrum. This model also makes successful predictions for the model's reionization parameters. None of my models make use of 21cm images, instead compressing the 21cm maps using summary statistics such as the power spectrum and 3PCF.

2.6 Thesis structure

The rest of this thesis is separated into five chapters. Chapter 3 describes my code for calculating higher-order clustering statistics from observed or simulated data. I test the code by using it on mock data with known three-point clustering properties. I then use the code to measure the three-point correlation function of the ionisation fraction data from semi-numerical simulations. In Chapter 4, I use machine learning techniques to extract useful information about the Epoch of Reionization from measurements of the three-point correlation function. In Chapter 5, I fit an analytic clustering model to simulated ionisation maps from SIMFAST21. Starting with the simplest model possible, I slowly add more complexity and fit the required extra properties directly from SIMFAST21 simulations. In Chapter 6, I use machine learning techniques to replace semi-numerical simulations with surrogate models. I present my overall conclusions and outlook for how the work in this thesis will affect the field in Chapter 7.

Chapter 3

High-order clustering calculations

The clustering of ionised hydrogen bubbles encodes much information about the reionization process. Clustering is a measure of clumpiness: if some data are highly clustered on a specific scale then locations separated by that scale tend to have similar value. A simple measure of clustering is the two-point correlation function $\xi^{(2)}(r)$, which gives the excess probability of finding two similar-valued spots at a separation r compared to a random data field. I investigate the spatial correlations between triplets of points, quantified by the three-point correlation function $\xi^{(3)}(r)$ (hereafter written 3PCF). See Section 2.5 for a review of relevant recent literature for this chapter. In this chapter I investigate correlations in real-space to see whether any information can be extracted in real-space that does not exist in the Fourier-space equivalents. Section 2.5.3 presents a review of the recent high-order clustering literature. My code is available publicly on GitHub[1].

The rest of this chapter is structured as follows. In Section 3.1, I describe my code implementation. Section 3.2 contains detailed testing of the accuracy and validity of my code by measuring the equilateral three-point correlation function of realisations from a distribution with known analytic form. In Section 3.3, I use the code on the outputs from semi-numerical reioinzation simulations, to see how the three-point correlation function for equilateral triangles is affected by the simulation parameters. I conclude this chapter in Section 3.4 with a summary and some potential further uses for the code.

[1] https://github.com/wdjennings/3PCF-Fast

3.1 Implementation

Calculating the three-point correlation function involves placing differently-sized triangles into the data field. The product of the data values at each of the three triangle vertices is summed over a large number of similarly-sized triangles, and an estimate of the correlation function is built up. The final output of the algorithm is the correlation function estimates $\xi^{(3)}(r_i)$ at a discrete set of radius values r_i, corresponding to a discrete set of radius bins. The correlation function estimate for each radius bin is calculated by using a set of many triangles with similar (but not identical) side lengths. In this section I first describe how to find these sets of triangles, by matching triangles whose side-lengths lie within a given binned range of radii $R_{\rm min} \leq r < R_{\rm max}$. Then I give pseudocode for my C++ algorithm to place these triangles onto the data and sum the data at the resulting vertices. Finally I discuss how I use the output statistics from the code to estimate a correlation function value.

3.1.1 Matching equilateral triangles

Efficiently finding sets of similarly-sized triangles is a key preparation stage of the algorithm. The data in this section are represented as a pixelised set of scalar values in three-dimensions. For each radius bin I find all the triangles whose edge lengths r_1, r_2, r_3 lie within a fixed range of side-lengths $R_{\rm min} \leq r_i < R_{\rm max}$. There are a finite number of such triangles because the three vertices are constrained to lie on the centres of pixels in the data. To find explicit matching triangles I place the first vertex at the origin. I then find all possible second vertices ($\mathbf{r_2}$) which lie within the spherical shell $R_{\rm min} \leq |\mathbf{r_2}| < R_{\rm max}$ of the origin. From each of the matching second vertex points, I find the third vertex points ($\mathbf{r_3}$) which are a valid distance both from $\mathbf{r_2}$ and from the origin. This last step is effectively finding pixels which lie in the overlap of two spherical shells. Figure 3.1 shows an example in two dimensions: with the first triangle vertex at the origin, the darker annulus indicates the allowed region for the second vertex between $R_{\rm min}$ and $R_{\rm max}$. The lighter region then shows the allowed region of third vertices from one of the possible second vertices. The final matching triangles (of which there are two) are outlined in black in the figure.

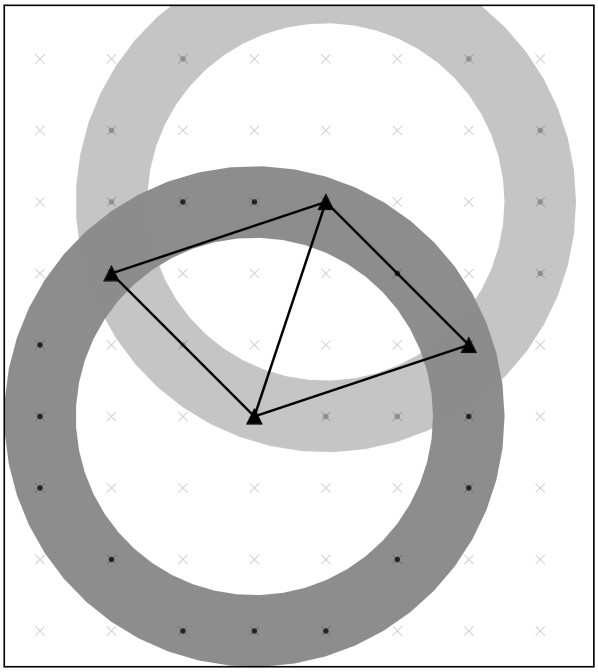

Figure 3.1: Triangles matching the radius bin condition 2.5 pixels $\leq r <$ 3.5 pixels. The two regions shown are the radius conditions around the first and second points. The allowed third point(s) then lie in the overlap of these annuli.

To prevent repeated calculations I use a PYTHON script to search for these matching triangles and store the resulting pairs of vectors $(\mathbf{r_2}, \mathbf{r_3})$ in a binary file. This binary file can be loaded by the main C++ algorithm many times. I refer to these binary files as VERTICES files.

3.1.2 Cross-correlation statistics

The correlation function of a data field is usually calculated in comparison to a random field. The correlation function then quantifies the extent to which the data field is more clustered than a random field. For a scalar data field, the random field should be uniform with mean equal to the data mean. The outputs from my code are the auto- and cross-correlation statistics between the data (D) and random (R) fields. For the 3PCF these statistics are the data-data-data statistic (DDD), data-data-random (DDR), data-random-random (DRR) and random-random-random (RRR). DDD is the auto-correlation found by multiplying the data field at all three vertices. DDR is the cross-correlation found by multiplying the data at two vertices and the

random field at the final vertex; and so on. These statistics will later be combined to give an estimate of the correlation function. It is practically simpler to normalise the data field to have a mean unity, so that the random field is everywhere unity and can be more efficiently summed as integer counts.

To understand why the random field should be uniform in our case, consider the equivalent method for galaxy clustering calculations. The random field for a galaxy clustering calculation is generated by placing a large number of galaxies randomly into the same survey volume as the data field. Ideally one would use as many galaxies as possible for the random field. In this chapter, we are calculating the correlation of a continuous scalar field. For galaxy clustering calculation, our code code be used by binning the survey volume into pixels and counting the total number of galaxies in each pixel. Any number of galaxies can be used in the random catalog. Thus, allowing the number of galaxies to tend to infinity causes the mean number of galaxies in all pixels tends towards being completely uniform.

3.1.3 Pseudocode

The 3PCF algorithm requires two inputs: the data field (δ) and a binned VERTICES file. The data field is immediately normalised to have a mean of unity. The VERTICES file contains the pairs of $(\mathbf{r_2}, \mathbf{r_3})$ triangle vectors for each radius bin. The algorithm outputs the three-point correlation statistics (DDD, DDR, DRR, and RRR) for each radius bin.

Algorithm 1 Three-point correlation algorithm

```
1:  procedure 3PCF(δ[r], R_min, R_max)
2:      DDD, DDR, DRR, RRR ← 0                          ▷ Initialise to zero
3:      Load (r_2, r_3)                  ▷ using (R_min, R_max) VERTICES file
4:      for all r_1 do                                   ▷ over all data pixels
5:          for each r_2,r_3 pair do                     ▷ over matching triangles
6:              DDD += δ[r_1] × δ[r_1 + r_2] × δ[r_1 + r_3]
7:              DDR += δ[r_1] × δ[r_1 + r_2]
8:              DRR += δ[r_1]
9:              RRR += 1
10:         end for
11:     end for
12:     return DDD, DDR, DRR, RRR
13: end procedure
```

3.1.4 Estimators

Algorithm 1 outputs the correlation statistics (DDD, DDR, DRR, RRR). An estimate of the correlation function is found by combining these statistics. The simplest such estimator is given by ratios of the data- and random-field auto-correlations,

$$\zeta^{(3)} = \frac{DDD - RRR}{RRR}. \tag{3.1}$$

Another estimator,

$$\zeta^{(3)} = \frac{DDD - 3DDR + 3DRR - RRR}{RRR}, \tag{3.2}$$

from (Landy and Szalay, 1993) generally leads to less biased results, because it takes account of cross-correlations between the data and random fields which the simple estimator ignores.

3.2 Testing distribution: points-on-spheres

I test my code by generating three-dimensional realisations for a distribution with known 3PCF. I compare the measured 3PCF from my code to the theoretical form, to get an indication of the regimes in which the code has good accuracy and precision. The easiest way to explain the testing distribution is to describe what realisations look like. Each three-dimensional realisation is made up of a set of points in a box. First, a large set of points are uniformly placed on the surfaces of many identically-sized spheres. The data are then saved to a data file by overlaying a three-dimensional pixelised grid and counting the number of points in each grid: zero for no points, one for a single point, and so on. For all realisations in this section, the data are represented as a box with side length 100 arbitrary units pixelised into 512^3 pixels. The amplitude and shape of the theoretical 3PCF for these realisations depend on the sphere radius R and the number density of spheres $n_{\rm s}$. I describe a scenario as a particular pair of these two parameters. I also use the number of spheres $N_{\rm s} = n_{\rm s} \times 10^6$, since all realisations in this section have a fixed box size of 100 arbitrary units in each of the three dimensions.

3.2.1 Theoretical three-point correlation function

The equilateral three-point correlation function of points-on-sphere realisations has a closed analytic form, kindly provided by Lorne Whiteway (2018). For a scenario with parameters n_s and R, the three-point correlation function for equilateral triangles as a function of the triangle side length r is given by

$$\xi^{(3)}(r) = \begin{cases} \dfrac{1}{16\pi^3 R^3 n_\mathrm{s}^2 r^2 \sqrt{3R^2 - r^2}} & \text{if } r < R\sqrt{3}, \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

3.2.2 Generating realisations

Generating a realisation for a scenario involves choosing where to put the spheres and then placing points on the surfaces of those spheres. A uniformly random set of N_S points are chosen to be the centres of the spheres. Points are then placed randomly onto the surface of each sphere. Ensuring that the points are indeed uniformly distributed across the spheres surface is a solved problem. A first naive approach was to sample random pairs of angles $\{\theta, \phi\}$ uniformly in the ranges $\{0, \pi\}$ and $\{0, 2\pi\}$, placing points on the sphere using spherical coordinates,

$$x = R\sin(\theta)\cos(\phi) \tag{3.4}$$

$$y = R\sin(\theta)\sin(\phi) \tag{3.5}$$

$$z = R\cos(\theta). \tag{3.6}$$

This method gives a biased oversampling of points near the poles where $|\cos(\theta)| \approx 0$. At these poles the surface density of points increases due to the uniform sampling of ϕ. A better method (Muller and E., 1959) is to sample three random variables x, y, z from the normal distribution $\mathcal{N}(0,1)$ and normalise by the Euclidean norm of these three coordinates. The distribution of the normalised vectors

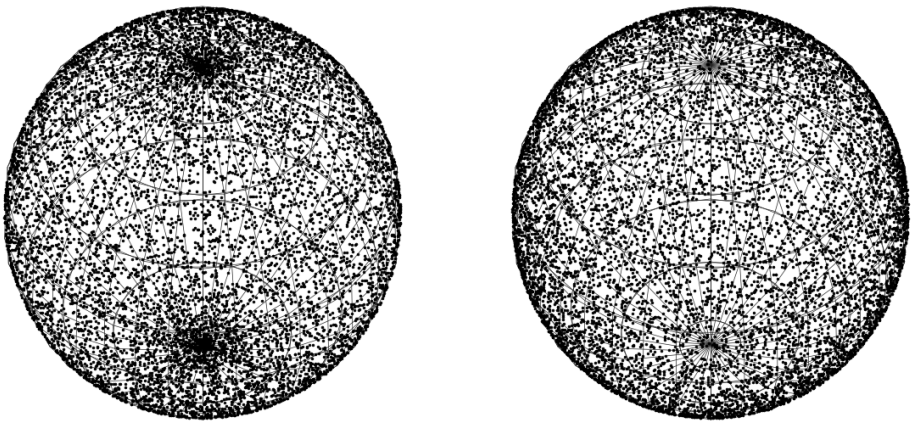

Figure 3.2: Two different sampling methods for generating points on a sphere. The left panel shows a naive method of uniformly sampling pairs of angles (θ, ϕ) which over-samples points near the vertical poles. The right panel shows a better method by sampling each coordinate from normal distributions.

$$\mathbf{r} = \frac{R}{\sqrt{x^2+y^2+z^2}} \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \tag{3.7}$$

is then uniform across the surface of a sphere with radius R. Figure 3.2 shows an example of the two different sampling methods, with the naive method visibly over-sampling points near the poles.

After storing the locations of all points on all spheres, the final pixelised realisation of the scalar field is generated by rounding the point coordinates to the nearest integer. Figure 3.3 shows a slice through an example realisation of the testing distribution. All data in this testing section are pixelised with $N = 512$ points in each dimension. For the constant box size of $L = 100$ this gives a pixel size of around 0.2 arbitrary physical units.

3.2.3 Testing the code

I test my code by generating points-on-spheres realisations for many R and N_s scenarios. I compare the outputs of my code to the true theoretical three-point correlation function in Equation (3.3). Figure 3.4 shows the theoretical and measured

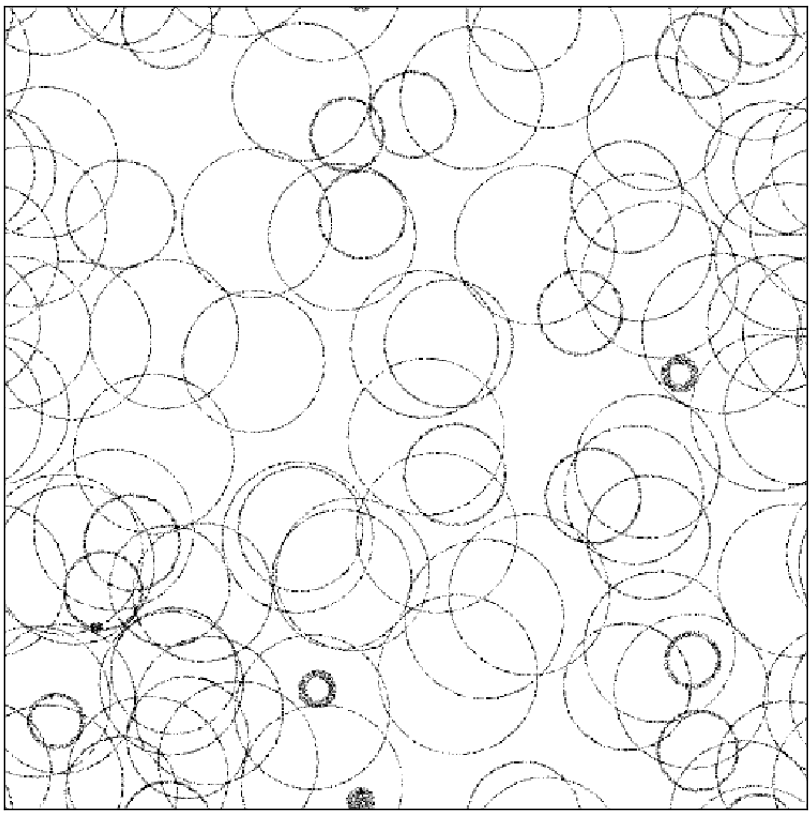

Figure 3.3: Slice through an example realisation of points-on-spheres data. This scenario uses spheres with $R = 10$ and $N_s = 200$. Each sphere appears as a circular annulus as it has been horizontally sliced for this figure. Some annuli appear thicker than others because slicing a thick spherical shell near its pole gives a wider region when viewed from above.

equilateral three-point correlation functions for seven scenarios with a range of R values and fixed $n_s = 5 \times 10^{-5}$, using the Landy-Szalay estimator. The theoretical correlation function is shown in each case as the dashed line. The measured correlation function estimates are subject to sample variance, meaning that the output from the code depends on the randomly-seeded initial conditions. I use five realisations with different random seeds to determine whether the theoretical correlation function lies inside the spread of the five measured code outputs. The shaded regions in Figure 3.4 show the standard deviation of the measured three-point correlation function across these five realisations. Figure 3.5 similarly shows the theoretical and measured correlation functions for scenarios with fixed $R = 5$ and various N_s values.

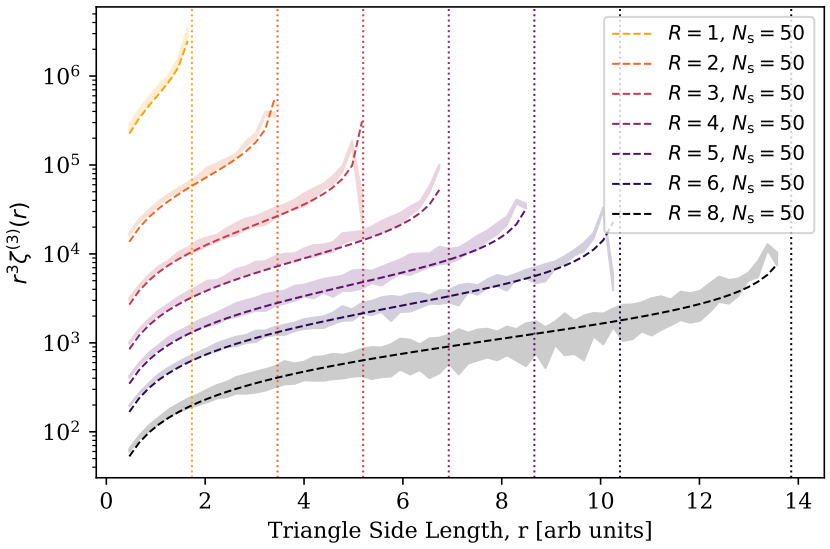

Figure 3.4: Equilateral three-point correlation functions for points-on-spheres scenarios with varying sphere radius R. The dashed lines show the theoretical correlation functions for each scenario whose parameters are shown in the legend. The shaded regions show the standard deviation of the measured correlation functions across five realisations, using my code with the Landy-Szalay estimator. The theory lines lie within the shaded regions for most triangle side lengths, except at radius values near the upper valid limits (far right hand side of each scenario) as discussed in the text. Vertical dotted lines indicate the theoretical asymptotes at $R\sqrt{3}$ for each scenario.

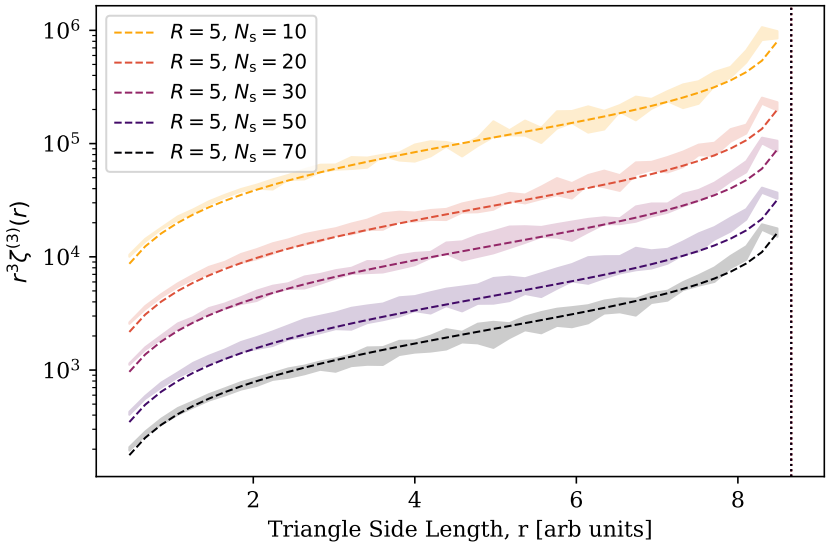

Figure 3.5: Equilateral three-point correlation functions for points-on-spheres data with varying n_s between 1×10^{-5} and 7×10^{-5}, using the LS estimator. The dashed theory lines again lie within the measured shaded regions for most triangle side lengths. The vertical dotted line indicates the theoretical asymptote at $R\sqrt{3} = 5\sqrt{3}$ for all these scenarios.

The measured and theoretical correlation functions match closely across most of the triangle side lengths. The theoretical correlation function in Equation (3.3) has a vertical asymptote at the maximum allowed radius $R\sqrt{3}$. This can be seen in Figure 3.4 as vertical dotted lines and as a slight upturn near the right-hand sides of each dashed line. My code slightly over-predicts the theory in each case near the maximum valid radius. This is due to the binning of triangles: each binned output is calculated using equilateral triangles with a range of side lengths as described in Section 3.1.1. Averaging the correlation function over these differently-sized triangles causes a discrepancy between measured and theoretical correlation functions, since some triangles are included which are larger than the valid maximum radius and thus have a theoretical 3PCF value of zero (see Equation (3.3)).

The Landy-Szalay (LS) estimator in Equation (3.2) gives better results than the simple estimator in Equation (3.1). Figure 3.6 shows the measured and predicted correlation functions for the same code outputs as in Figure 3.4, but using the simple estimator instead of the LS estimator. The results for the simple estimator are significantly biased for most of these scenarios, and I use the LS estimator for the remainder of this chapter.

3.2.4 Threading

This three-point correlation code would be a useful addition to the output statistics from numerical and semi-numerical simulations. Such simulations are often run on multiple threads. I added threading to my 3PCF code to reduce the time taken to measure $\zeta^{(3)}$ for a large simulation. Each thread is assigned a different section of the data field over which to run the $\mathbf{r_1}$-loop (for the first of the triangle vertices). The statistics from all threads are then summed at the end, since all such calculations are entirely independent.

3.2.5 Using jackknifing for sample variance

In the previous subsections I estimate the variance of the three-point correlation function by taking the spread from differently-seeded random realisations. This method is not possible when calculating $\zeta^{(3)}$ from observed data: only one realisa-

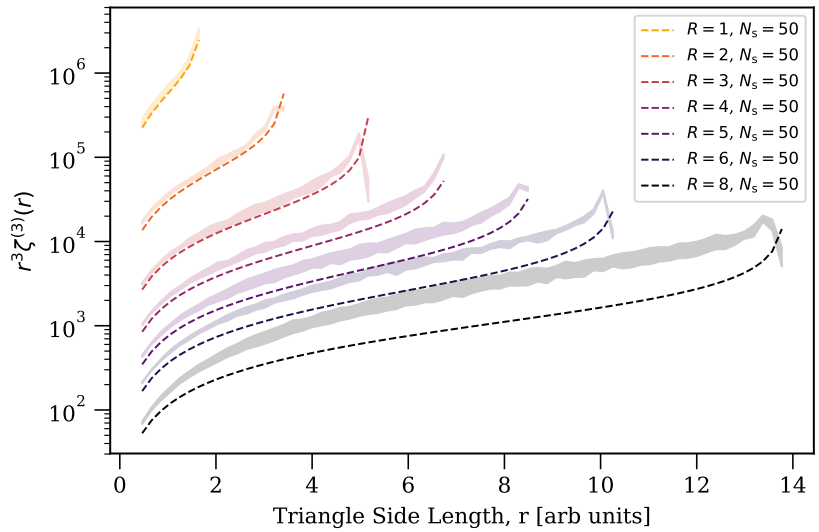

Figure 3.6: Equilateral three-point correlation functions for the same data as Figure 3.4, but using the simple estimator in Equation (3.1) instead of the LS estimator in Equation (3.2). The dashed lines lie outside the shaded regions, meaning that the simple estimator gives correlation estimates that are offset from the theoretical distribution. Vertical asymptote lines are omitted here so that the offset between theory and measured 3PCF is more easily visible.

tion of the Universe can be used. It is useful to measure the sample variance using a technique known as Jackknifing, which I implemented in the code. The three-point algorithm is run on the same data box several times, each time excluding one segment of the data. Figure 3.7 shows a schematic of jackknifing using four jackknife samples. The 3PCF is found four times, each time excluding one quarter of the displayed data field. In the top-left panel, the top-left quartile has been removed so that the statistic is calculated on the remained three-quarters of the data field. In the top-right panel, the top-right quartile has been removed; and so on. The variance in the resulting statistics gives a measure of the uncertainty in the calculated three-point correlation function. Instead of running the correlation code multiple times, it is much faster to run the full calculation over the full box, and then afterwards calculate the jackknife errors by excluding some of the measurements.

3.3 Semi-numerical simulations

The code has been well-tested on data with a known three-point distribution. I now present the 3PCF of $x_{\rm HII}(\mathbf{r}, z)$ data from SIMFAST21 simulations. First I test

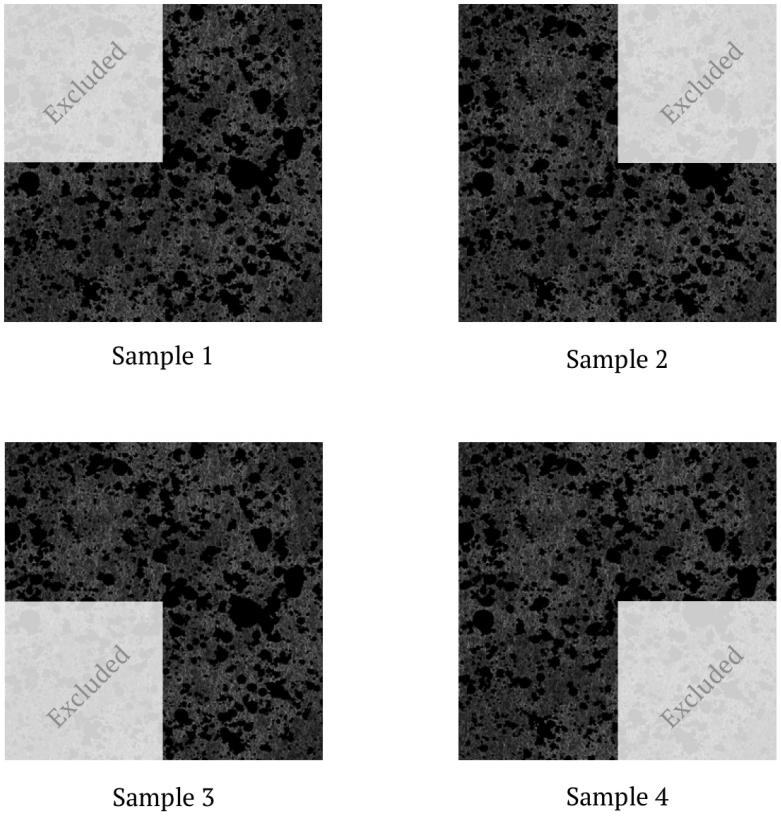

Figure 3.7: Schematic of how jackknifing can be used to estimate errors of a calculated statistic. The statistic is calculated several times on the same data field, each time excluding a subset of the data from the calculations. .

the effect of sub-sampling the triangles from the triangle-matching algorithm. I determine the optimum sampling level that balances computation time with quality of the correlation estimates. I then show how the shape and amplitude of the three-point correlation function depend on the redshift and input parameters of the simulation. All SIMFAST21 simulations in this chapter use a box size of 150Mpc/h with the halo field resolved into 768^3 pixels and the $x_{\rm HII}$ field resolved into 256^3 pixels. The cosmological parameters from Planck Collaboration (2018) were used: $\Omega_{\rm M} = 0.315$, $\Omega_{\rm b} = 0.0493$, $\Omega_\Lambda = 0.6847$, $\rm h = 0.6736$ and $\sigma_8 = 0.8111$. All data are normalised before calculating the correlation function in the same way as Section 3.1.2, dividing by the data mean so that the random field is a uniform field of value unity.

3.3.1 Effect of subsampling triangle configurations

The algorithm in Section 3.1.1 can quickly lead to hundreds of thousands of matching triangles, for side lengths larger than around ten pixels. Even running the matching algorithm itself for such side lengths can take several days and, more significantly, using such an exhaustive set of triangles in the correlation algorithm would require years of CPU time. An accurate measurement of the three-point correlation function can be obtained more efficiently by subsampling a small number of triangles from all valid matches. I test how this subsampling affects the code outputs. The three-point correlation function is calculated on $x_{\rm HII}(\mathbf{r})$ data from five randomly-seeded SIMFAST21 realisations using input parameters $M_{\rm min} = 3 \times 10^8 M_\odot$, $\zeta_{\rm ion} = 30.0$ and $R_{\rm max} = 10.0$Mpc. The variance in the correlation functions between the five realisations measures the scatter in the outputs.

Figure 3.8 shows how this variance depends on the number of triangles used in the correlation function algorithm, using the $x_{\rm HII}(\mathbf{r})$ outputs at $z = 12.5$ where $x_{\rm HII} = 0.15$. The variance is large for a small number of triangles but decreases as more triangles are used. The subsampling of triangles indeed causes scatter in the outputs from the algorithm. The scatter is smaller when more triangles are used and, for more than around 2000 triangles, the scatter plateaus. The remaining variance is most likely due to inherent sample variance in the random seeding of the five SIMFAST21 realisations. I use 5000 triangles in all the correlation function estimates from here onwards, giving calculation times of around an hour on these SIMFAST21 simulations.

3.3.2 Redshift dependence

The equilateral-triangle three-point correlation function of the ionisation fraction field $x_{\rm HII}(\mathbf{r})$ encodes information about the size of the ionised bubbles. In this subsection I show how the correlation function evolves with redshift. I use the outputs from five randomly-seeded SIMFAST21 realisations using the same canonical reionization parameters as in the previous subsection. Figure 3.9 shows the resulting correlation functions, presenting both the mean (solid lines) and standard deviation (shaded regions) of the five realisation outputs.

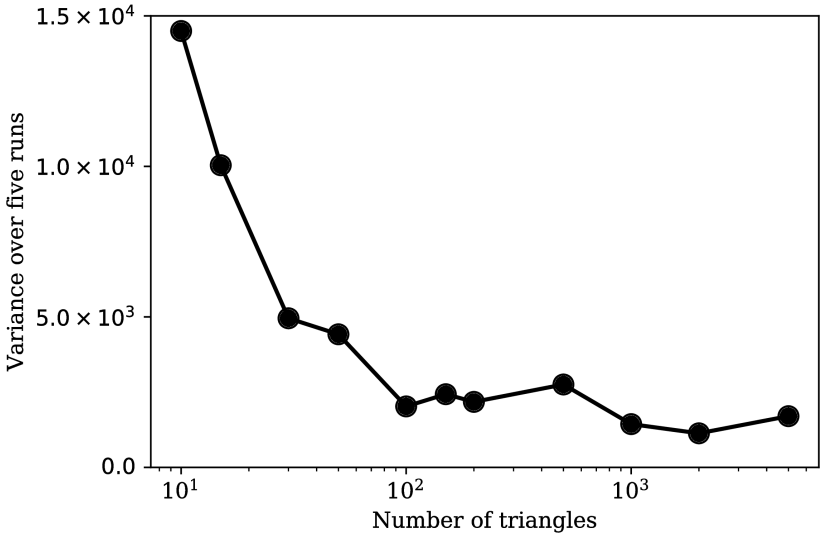

Figure 3.8: Effect of subsampling triangles on the final equilateral three-point correlation measurement. Each point shows the variance in three-point correlation function estimates from five SIMFAST21 realisations. The simulation parameters were $M_{\rm min} = 3 \times 10^8 M_\odot$, $\zeta_{\rm ion} = 30.0$ and $R_{\rm max} = 10.0$Mpc, and the variances calculated at $z = 12.5$ where $x_{\rm HII} = 0.15$. The variance decreases as the number of triangles is increased, and plateaus for more than 2000 triangles.

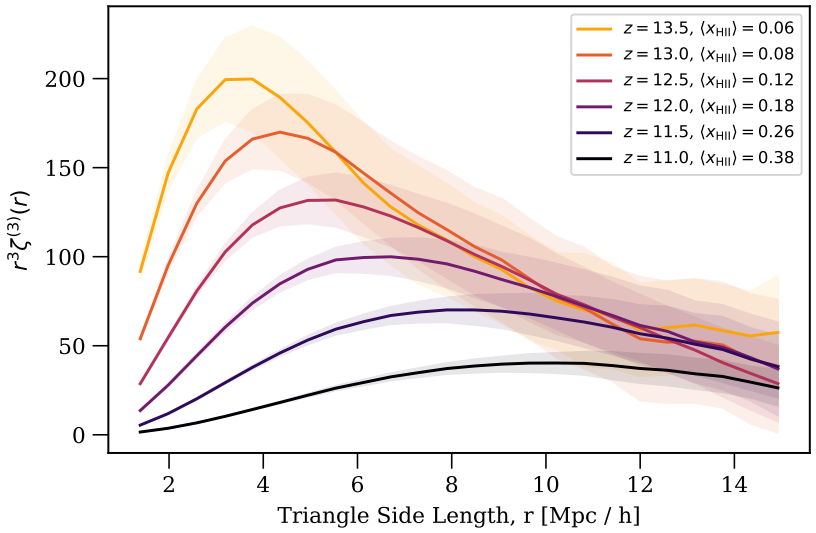

Figure 3.9: Equilateral three-point correlation function, averaged over five canonical SIMFAST21 simulations. Correlation functions for several redshifts are shown, along with the corresponding global ionised fractions $\langle x_{\rm HII} \rangle$. Solid lines show the mean of the five realisations, and shaded regions show the standard deviation. As reionization progresses the bubbles in the simulation continually grow. The peak in the correlation function traces the mean bubble size, and shifts to higher scales for later redshifts as can be seen here.

The equilateral three-point correlation function peaks at a preferential scale in each redshift. This scale increases with redshift and traces the bubble size. For small triangles there is minimal scatter between the five realisations – the shaded regions lie closely around the solid lines. This scatter increases for triangles that are larger than the bubble size, particularly for earlier redshifts. At early times the small ionised regions are still isolated and the number of resolved halos is sensitive to small fluctuations in the randomly-seeded density field. This acts as a source of shot noise and gives rise to scatter in the large-scale clustering measurements. For later times this shot noise is lessened as the bubbles become larger and less isolated. The late-time halo abundances shows less sensitivity to small fluctuations in the density field.

3.3.3 Ionisation efficiency dependence

In this subsection I show how the shape and amplitude of the equilateral three-point correlation function depend on the SIMFAST21 ionisation efficiency parameter $\zeta_{\rm ion}$. Section 2.1.1 described how $\zeta_{\rm ion}$ affects the growth of ionised bubbles, with larger values of $\zeta_{\rm ion}$ giving faster-growing bubbles. Figure 3.10 shows the measured equilateral-triangle three-point correlation function at $z = 13$ for several different ionisation efficiency scenarios, in each case using five randomly-seeded SIMFAST21 simulations with fixed $M_{\rm min} = 3 \times 10^8 M_{\odot}$ and $R_{\rm max} = 10.0$Mpc. The ionisation efficiency has a similar effect on the correlation function as redshift evolution. Larger ionising efficiencies give rise to larger ionised bubbles. Increasing the value of $\zeta_{\rm ion}$ shifts the correlation peak towards larger scales in the same way as evolving the redshift does. The global ionised fractions $\langle x_{\rm HII} \rangle$ are shown for each of these scenarios in the legend of Figure 3.10. Figure 3.11 shows how the peaks in the correlation functions of Figure 3.10 are related to the actual size of ionised bubbles. The ionised bubble sizes for this figure are measured using the mean-free-path method in Mesinger and Furlanetto (2007). The radii at which the correlation functions peak are clearly strongly correlated with the bubble size. In Chapter 4 I investigate the relationship between the typical bubble size and the 3PCF using machine learning.

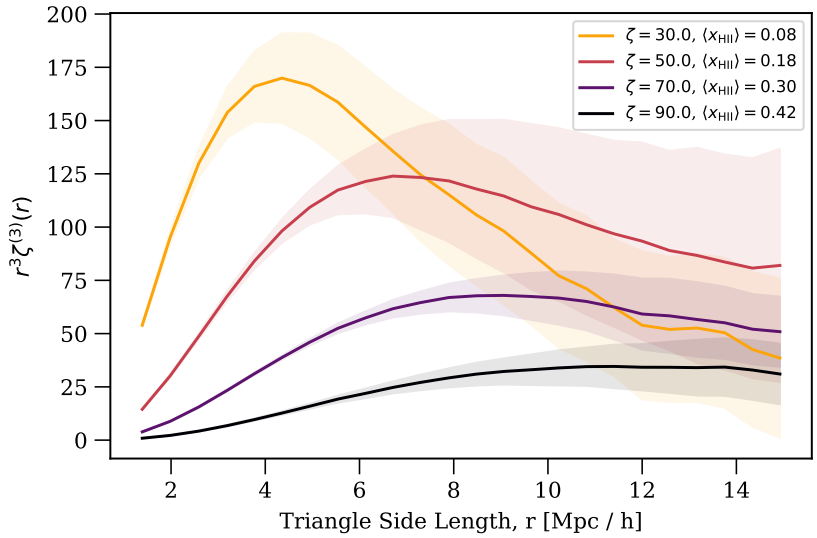

Figure 3.10: Equilateral three-point correlation function from SIMFAST21 scenarios with varying $\zeta_{\rm ion}$, in each case using the results for $z = 13$. The peak in the correlation function shifts to larger scales for increasing ionisation efficiency. The peak scale traces the mean bubble size for each scenario. A larger $\zeta_{\rm ion}$ scenario causes ionised bubbles to grow more quickly, thereby giving a peak at a higher scale.

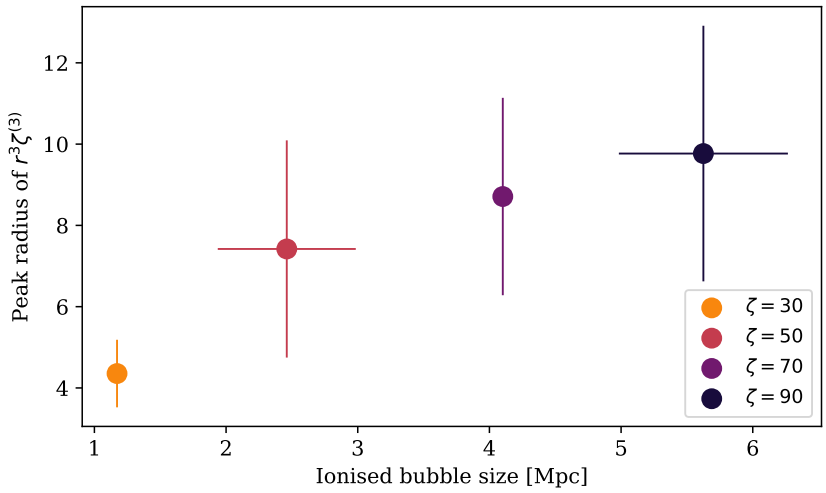

Figure 3.11: Radius at which the three-point correlation function curves in Figure 3.10 are at a maximum, against the true ionised bubble size using the mean-free-path method in Mesinger and Furlanetto (2007). The radius of the correlation function curve peak increases with the bubble size. The error bars are given as the 2σ limits for both the correlation function peak and the mean-free-path peak, using five realisations.

It is interesting to determine whether $\zeta_{\rm ion}$ has any other effect on the three-point correlation function, other than those which are degenerate with redshift evolution. Figure 3.12 shows the correlation functions of different $\zeta_{\rm ion}$ scenarios at fixed global ionisation fraction $\langle x_{\rm HII} \rangle$, instead of fixed redshift. For each scenario the correlation is plotted at a redshift whose corresponding global ionised fraction is closest to $\langle x_{\rm HII} \rangle = 0.20$. The best matching redshifts are given in the figure legend. The peak in the equilateral three-point correlation functions at fixed $\langle x_{\rm HII} \rangle$ increases only slightly with ionisation efficiency parameter, indicating that the mean bubble size is roughly similar for all scenarios despite the different redshifts. The $\zeta_{\rm ion}$ parameter does however strongly affect the amplitude of the correlation function. Higher ionising efficiencies generally lead to a larger overall amplitude in the correlation function. This effect is due to the difference in the bubble size distributions. Figure 3.13 shows slices through SIMFAST21 realisations for two of these scenarios, both of which have $\langle x_{\rm HII} \rangle \approx 0.2$. The low ionisation efficiency scenario ($\zeta_{\rm ion} = 10$) visibly has many more small and partially-ionised bubbles than the high ionisation efficiency scenario (with $\zeta_{\rm ion} = 70$). These partially-ionised bubbles appear as orange or yellow pixels in the figure. This difference is due to the speed of bubble growth affecting the redshift at which each simulation reaches the required global ionised fraction of $x_{\rm HII} = 0.20$. There are fewer small halos at earlier redshifts and so the higher ionisation efficiency scenario, which reaches the required ionising fraction at an earlier redshift, has fewer small bubbles. Similarly the bubbles in a low $\zeta_{\rm ion}$ scenario grow more slowly, reaching the required $x_{\rm HII}$ at a later redshift, and leading to the presence of more small bubbles and fewer large bubbles. The higher $\zeta_{\rm ion}$ models thus have a higher abundance bubbles that are larger than the correlation-peak scale (around 6Mpc/h), giving rise to a larger correlation amplitude.

3.3.4 Minimum halo mass dependence

In this subsection I perform the same analysis for the minimum halo mass parameter $M_{\rm min}$. Halos with masses lower than this parameter are ignored in the simulation. The motivation for this parameter is to allow for the exclusion of small dark matter

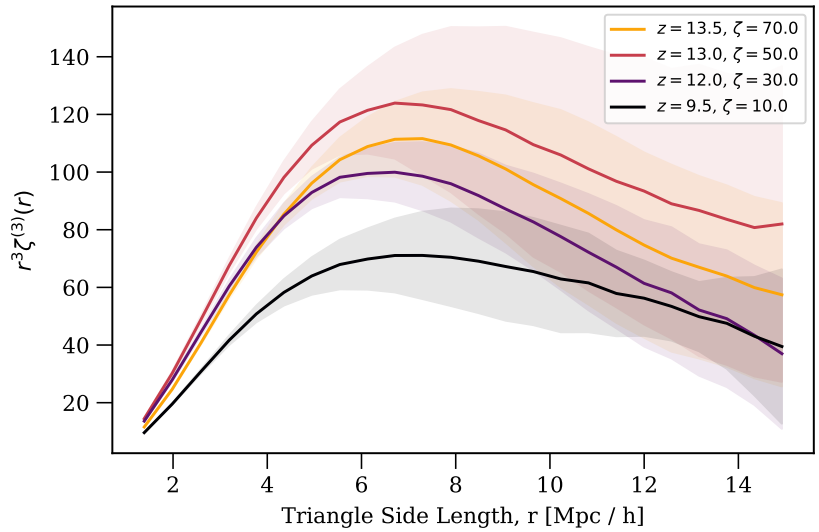

Figure 3.12: Equilateral three-point correlation function from SIMFAST21 scenarios with varying $\zeta_{\rm ion}$, in each case finding nearest redshift for which $\langle x_{\rm HII}\rangle \approx 0.20$. The nearest values are within $\langle x_{\rm HII}\rangle = 0.20 \pm 0.02$ and the corresponding redshift for each scenario is given in the legend. The ionisation efficiency at fixed $\langle x_{\rm HII}\rangle$ affects the normalisation but has minimal effect on the location of the peak.

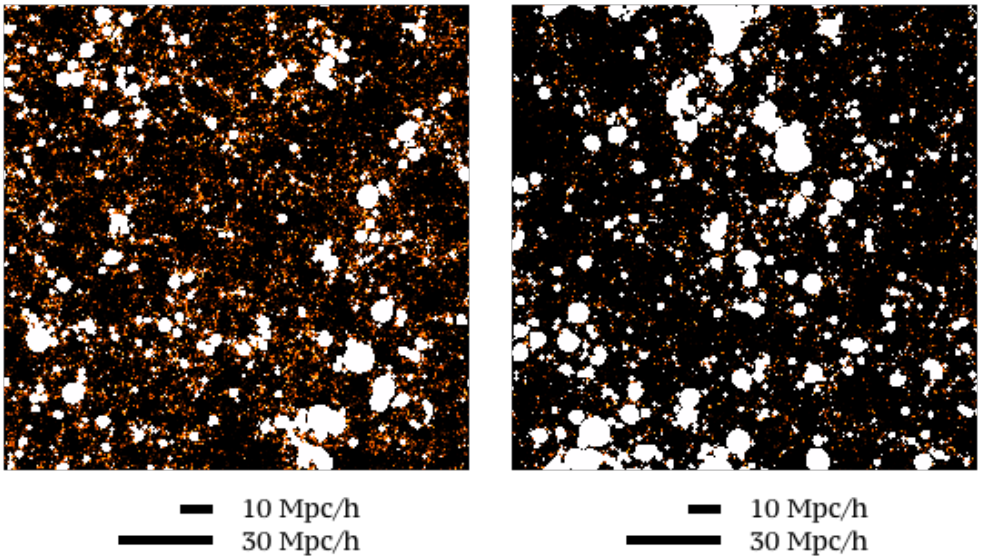

Figure 3.13: Ionisation fraction fields for fixed $x_{\rm HII} \approx 0.2$ from SIMFAST21. The left panel has low ionisation efficiency ($\zeta_{\rm ion} = 10$). The right panel has high ionisation efficiency ($\zeta_{\rm ion} = 70$). The high efficiency simulation has fewer small and partially-ionised bubbles but has the same global ionisation fraction. This difference gives rise to a different amplitude in the correlation function measurements for fixed $\langle x_{\rm HII}\rangle$.

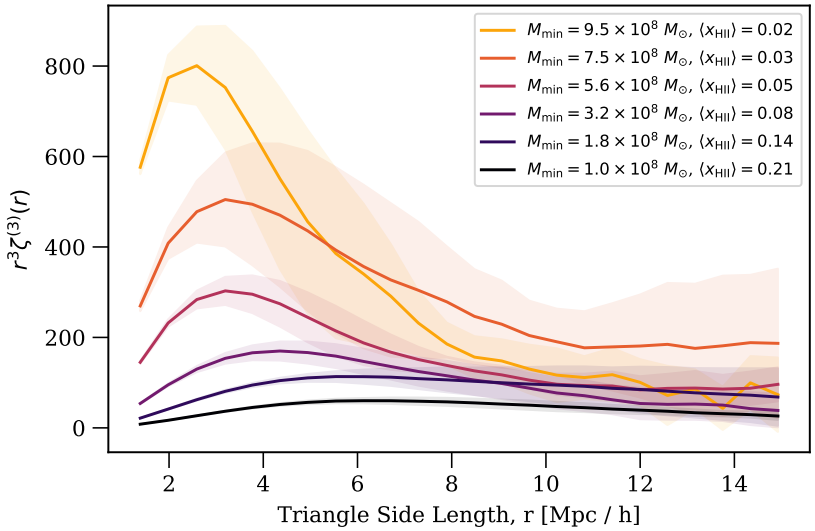

Figure 3.14: Equilateral three-point correlation function from SIMFAST21 scenarios with varying $M_{\rm min}$, in each case using the results for $z = 13$. The global ionised fractions $\langle x_{\rm HII} \rangle$ in the legend show less spread than those in Figure 3.10. The peak in the correlation function is only slightly affected by $M_{\rm min}$ but the amplitude is strongly affected. Both features are discussed in the text.

halos which generally have low star formation rates (see Barkana and Loeb 2001) and thus contribute very few photons to the ionising process. Increasing $M_{\rm min}$ reduces the number of halos and will reduce the abundance of the corresponding ionised bubbles. This process is likely to have an effect on the correlation function. Figure 3.14 shows the measured equilateral three-point correlation function at $z = 13$ for several different minimum halo mass scenarios. Each scenario again uses five randomly-seeded SIMFAST21 simulations and the other parameters are fixed at $\zeta_{\rm ion} = 30$ and $R_{\rm max} = 10.0$Mpc.

The minimum halo mass parameter has only a small effect on location of the correlation peak. The ionisation efficiency and redshift evolution both affect the speed of bubble growth, causing a shift in the peak location. The minimum halo mass parameter does not change the speed of bubble growth, but does affect the overall number of ionising sources. Increasing the minimum halo mass is equivalent to removing all small halos. A scenario with lower minimum halo mass will evolve more quickly than one with a higher minimum halo mass. This can be seen as a slight shift in the correlation function peak in Figure 3.14, still present but less

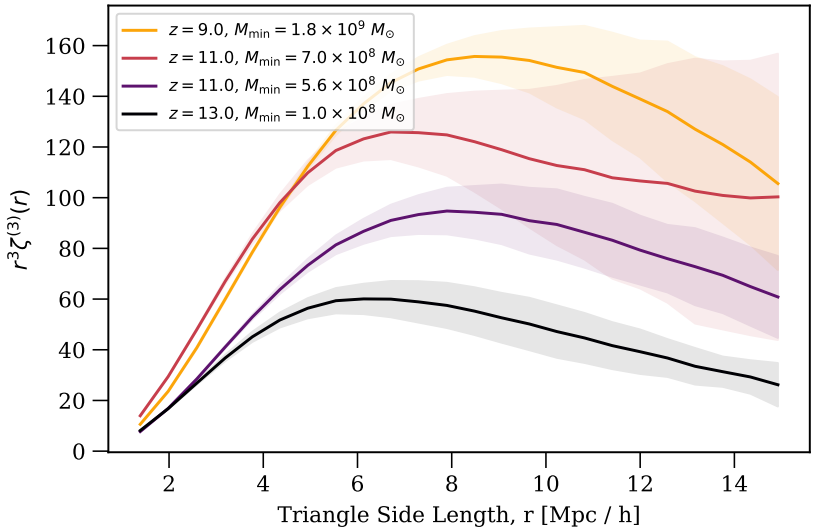

Figure 3.15: Equilateral three-point correlation function from SIMFAST21 scenarios with varying $M_{\rm min}$, in each case finding nearest redshift for which $\langle x_{\rm HII} \rangle \approx 0.20$. The nearest values are within $\langle x_{\rm HII} \rangle = 0.20 \pm 0.03$ and the corresponding redshift for each scenario is given in the legend.

significant than in the previous subsections. The amplitude relationship is again due to the difference in ionisation fractions between the scenarios. In the lower $M_{\rm min}$ scenarios there are many more small bubbles than the higher $M_{\rm min}$ ones. This leads to a higher ionised fraction in the low $M_{\rm min}$ scenarios which, when the data are normalised prior to calculating the three-point correlation function, reduces how distinct the bubbles are from the background mean value.

Figure 3.15 shows the effect of the minimum halo mass parameter excluding the redshift degeneracy. A similar effect is seen as for the $\zeta_{\rm ion}$ scenarios. The locations of the correlation function peak changes slightly with varying $M_{\rm min}$, indicating that the mean bubble size grows for the later-redshift scenarios. The bubble distributions visible in the slices through two of these scenarios (in Figure 3.16) show a similar effect, where the relative abundances of small and large bubbles affects the overall amplitude of the correlation. The effect is even more pronounced for $M_{\rm min}$, since the high $M_{\rm min}$ scenarios contain *no* small bubbles rather than a few small bubbles.

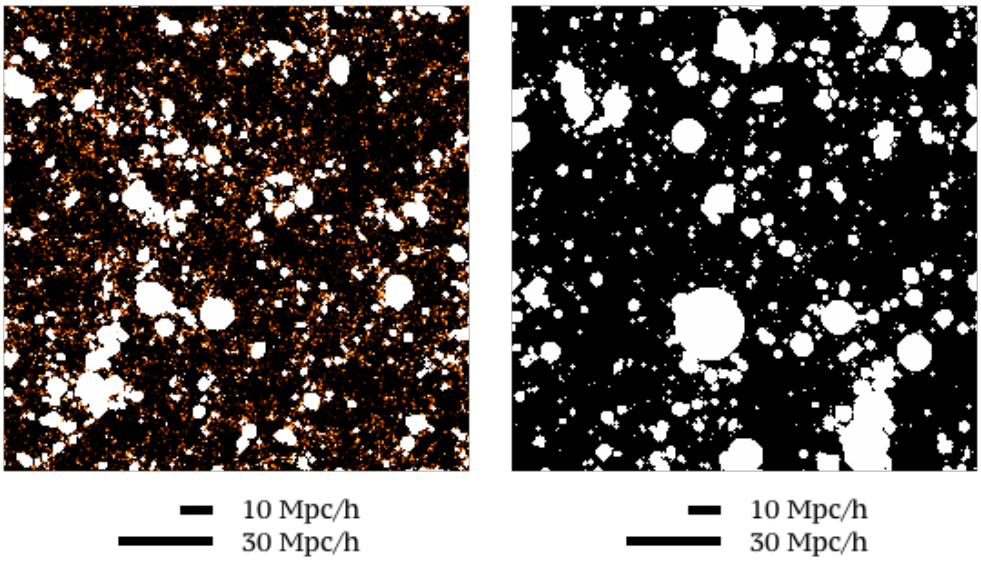

Figure 3.16: Ionisation fraction fields for fixed $x_{\rm HII} \approx 0.2$ from SIMFAST21. The left panel has low minimum halo mass ($M_{\rm min} = 10^8 M_\odot$). The right panel has high minimum halo mass ($M_{\rm min} = 1.8 \times 10^9 M_\odot$). The high halo mass scenario has no small and partially-ionised bubbles but the same global ionisation fraction. This difference gives rise to a different amplitude in the correlation function measurements for fixed $\langle x_{\rm HII} \rangle$.

3.4 Conclusions

Efficient calculation of the 3PCF is a potentially useful extra summary statistic for ionisation fraction data. My code generates accurate and precise estimates of the 3PCF, determined by using the code on data with known correlation properties. The code is far more efficient that the naive approach of looping over all possible triplets of pixels in the data. The equilateral triangle 3PCF of ionisation fraction data from SIMFAST21 shows a clear dependence on the parameters and redshift of the simulated data. Using the code in Markov-chain Monte Carlo analysis could potentially break some of the degeneracies in between the ionisation efficiency and minimum halo mass simulation parameters. It would be interesting to measure the relationship between the simulation parameters and the 3PCF of 21cm brightness temperature data. The code in its current form would also easily extent to four-point and higher order correlation functions, and could easily be used to measure the 3PCF for non-equilateral triangles.

Chapter 4

Analysing EoR data with the 3PCF

In the previous chapter I presented my optimised code for three-point correlation function (3PCF) calculations. I used the code on a small number of semi-numerical simulations to demonstrate how the 3PCF evolves with the redshift during theses simulations. I also investigated how the 3PCF depends on two of the reionization parameters, the minimum halo mass parameter $M_{\rm min}$, and the ionising efficiency $\zeta_{\rm ion}$. In this chapter I use machine learning techniques to investigate whether the 3PCF can inform us about the typical size of ionised bubbles ($R_{\rm bubble}$) and about the global ionisation fraction ($x_{\rm HII}$). See Section 2.5 for a review of the recent high-order clustering statistics literature.

The rest of this chapter is split in to the following sections. In Section 4.1, I discuss running semi-numerical simulations and how I calculate the statistics of interest from the simulation outputs. Section 4.2 details the choices I make before training my MLP models, including hyperparameters and input/output scaling types. In Section 4.3 I present my MLP models for predicting the typical bubble size, and Section 4.4 presents my MLP models for predicting the global ionisation fraction. I present my conclusions on using 3PCF data for EoR data in Section 4.5. In this chapter I use the following cosmological parameters: $\Omega_{\rm M} = 0.3153$, $\Omega_{\rm b} = 0.0493$, $\Omega_\Lambda = 0.6847$, $H_0 = 67.36 {\rm km\ s^{-1} Mpc^{-1}}$, $n_{\rm s} = 0.9649$, $\sigma_8 = 0.8111$, the latest results using the default PLIK likelihood from Planck Collaboration (2018).

4.1 Data acquisition

The data in this chapter were generated using 1000 realisations of the semi-numerical code 21CMFAST. Each realisation generates three-dimensional realisations of the $\delta T_{\rm b}(\mathbf{r})$ and $x_{\rm HII}(\mathbf{r})$ fields in cubes of size 250Mpc resolved into 256^3 pixels (smoothed from density fields resolved into 768^3 pixels). The resulting redshifts from the 21CMFAST algorithm are 5.0, 5.6, 6.3, 7.0, 7.78, 8.7, 9.6, 10.7, 11.9, 13.2, 14.6, 16.1, and 17.8. See Mesinger et al. (2011) for a description of the iterative algorithm that generates these steps. For each simulation, I calculate four summary statistics of interest: the 3PCF of the ionisation fraction field data; the 3PCF of the 21cm differential brightness temperature data; the bubble size distribution, described in this section; and the global ionisation fraction, found by trivially averaging the ionised fraction field $x_{\rm HII}(\mathbf{r})$ for each redshift.

4.1.1 Spin temperature fluctuations

A description of the 21CMFAST algorithm is given previously in Section 2.1.3. In this chapter I also include the effect of spin temperature fluctuations described briefly here. Spin temperature calculations involve modelling the kinetic gas temperature and the Lyman alpha background temperature. Allowing for spin temperature fluctuations means allowing for different local populations in the high and low energy states of the 21cm transition. The kinetic gas temperature $T_{\rm K}$ can be determined by considering the balance between a number of important heating and cooling mechanisms including X-ray emissions, Hubble expansion, adiabatic heating and cooling, and gas particle density changes due to ionization events. The dominant heating effect in 21CMFAST is from X-rays. X-ray photons are emitted with a range of wavelengths, the luminosities for which are assumed to follow a power-law relationship $L(\nu) \propto (\nu/\nu_0)^{-\alpha}$. The parameter α controls the slope of this spectral energy density function, and the parameter ν_0 controls the minimum frequency of X-rays which can escape into the Inter-Galactic Medium (IGM). This minimum frequency can also be written in terms of a minimum energy value, $E_0 = h\nu_0$, using the Planck constant $h = 4.135 \times 10^{-15}$eVs. See Mesinger et al. (2011) for a full derivation of the calculations and assumptions that 21CMFAST makes for the spin

temperature fluctuations.

4.1.2 Sampling the parameter space

Each simulation considers a different reionization scenario by changing three of the simulation parameters, namely:

1. The ionization efficiency $\zeta_{\rm ion}$, specifying how many ionising photons are sourced per unit of collapsed matter;

2. The minimum virial temperature $T_{\rm vir}$ which specifies a lower mass limit $M_{\rm min}$ of collapsed matter which produces ionising photons and X-rays;

3. The E_0 parameter which controls the minimum energy (or frequency) of X-ray photons which are able to escape into the IGM.

Fixing the other simulation parameters includes setting the efficiency of X-rays to a constant value. I use $\zeta_{\rm X} = 10^{-57} M_{\odot}^{-1}$ to match the assumption in Mesinger et al. (2011), equivalent to approximately a single X-ray photon for each stellar baryon as motivated by observations of low-redshift galaxies. The uncertain intergalactic medium X-ray properties are then parametrised by E_0.

In order sample a range of different reionization scenarios, I use a Latin Hypercube (McKay et al., 1979) approach. This method efficiently samples the input space with far fewer simulations than a naive exhaustive grid-search would require. The following ranges and scales of simulation parameters are used:

1. $\zeta_{\rm ion}$ in the linear range $[5, 100]$

2. $T_{\rm vir}$ in the logarithmic range $[10^4, 2 \times 10^5]$ K

3. E_0 in the linear range $[100, 1500]$ eV

These ranges were chosen to match those by the simulation authors (for example Greig and Mesinger 2015). The lower $T_{\rm vir}$ limit comes from a minimum temperature for the cooling of atomic hydrogen accreting onto halos. The upper limit arises from observations of high-redshift Lyman break galaxies (Greig and Mesinger, 2015). The $\zeta_{\rm ion}$ upper and lower limits correspond to escape fractions of 5% to 100% for

ionizing photons. The range for E_0 was chosen in a similar way to Park et al. (2018), motivated by hydrodynamic simulations (Das et al., 2017) and considering the energy that would allow an X-ray photon to travel a distance of roughly one Hubble length when travelling through a medium with $x_{\rm HII} = 0.5$.

4.1.3 Correlation function measurements

I use my code described in Chapter 3 to calculate the three-point correlation function of my simulated data. I calculate $\xi^{(3)}$ of both the ionisation fraction field $x_{\rm HII}(\mathbf{r})$ and of the 21cm differential brightness temperature field $\delta T_{\rm b}(\mathbf{r})$ for each simulation. The output of the 3PCF code are the correlation function amplitudes for 28 equilateral triangle bin configurations, with side lengths spaced in bins between 5 Mpc and 109 Mpc. These bins are spaced linearly for radii less than 20 Mpc, with logarithmically spaced bins for higher radii. The radius values for these 28 bins are: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 24, 29, 35, 42, 51, 62, 75, 91, and 109 Mpc.

4.1.4 Mean-free-path measurements for $x_{\rm HII}(\mathbf{r})$

In order to measure the typical bubble size I use my own implementation of the mean-free path method from Mesinger and Furlanetto (2007), summarised here. The input to the code is a single pixelised 3D ionisation fraction field data file, and the code outputs the full distribution of bubble sizes. The data field is first cleaned into 'transparent' and 'opaque' pixels by using a fixed threshold at $x_{\rm HII} = 0.5$. Pixels with $x_{\rm HII} \geq 0.5$ are transparent, and those with $x_{\rm HII} < 0.5$ are opaque. The mean-free path method simulates the emission of photons from random locations within the transparent regions. Each photon is emitted in a random direction and allowed to propagate until it reaches an opaque pixel. The distance travelled by each photon is measured and the resulting number of rays in a range of radius bins is calculated as dP/dR. I use 10^5 simulated photons in my measurements, and the resulting distances rounded to the nearest pixel size (0.98 Mpc). The distribution of bubble sizes is then directly proportional to RdP/dR (or equivalently VdP/dV). Figure 4.1 shows the RdP/dR outputs from my code for a simulation with canonical parameter

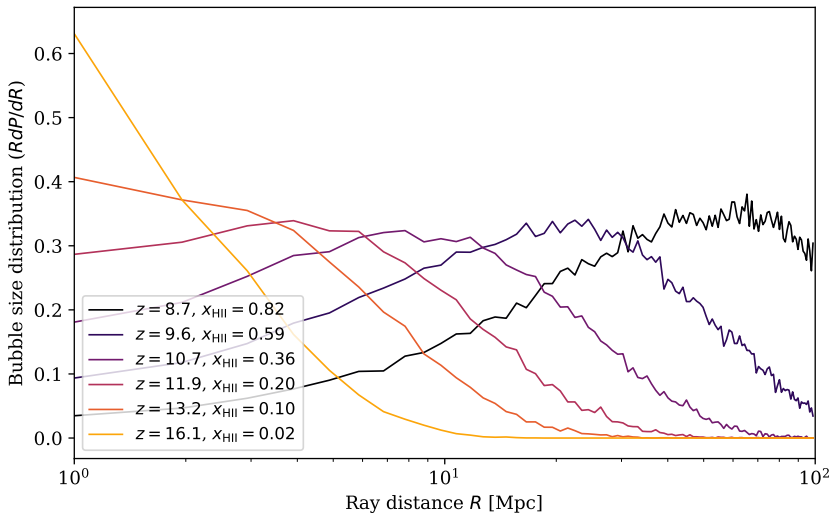

Figure 4.1: Example mean free path measurements of RdP/dR using ionisation fraction field data $x_{\mathrm{HII}}(\mathbf{r})$. Each line shows RdP/dR for a single redshift taken from a simulation with $T_{\mathrm{vir}} = 10^4$K, $\zeta_{\mathrm{ion}} = 30.0$ and $E_0 = 200$ eV. The typical bubble size (R_{bubble}) is related to the radius R_{max} at which these curves peak, according to $R_{\mathrm{bubble}} = 3R_{\mathrm{max}}$ (Giri et al., 2018b).

values $T_{\mathrm{vir}} = 10^4$K, $\zeta_{\mathrm{ion}} = 30$ and $E_0 = 200$ eV. Giri et al. (2018b) note that the peak radius R_{max} of this distribution occurs at a scale of $R = R_{\mathrm{bubble}}/3$, and so I use $R_{\mathrm{bubble}} = 3R_{\mathrm{max}}$ as my measurement of the typical bubble size in this chapter.

4.2 Model choices

4.2.1 Hyperparameter choices

Different search strategies for comparing hyperparameters are described in Section 2.4.6. In this chapter I use a random search method with five-fold cross-validation to find the best hyperparmeters. Two of the most important hyperparameters are the number of hidden layers and the sizes of these layers, collectively known as the network architecture. The architecture affects the model's ability to represent complex functions: a network with fewer and smaller layers is only be able to model simple relationships, whereas a larger network with more layers (or larger layers) will be able to represent more complex relationships. Using a model that is too small will result in poor prediction accuracy. Using a model that is too large will result in overfitting. There are no prescribed rules for deciding what range

of architectures to consider, but a common technique is to use one's knowledge both about the complexity and the dimensionality of the function that is being modelled.

When using the correlation function measurements as the inputs, there are 30 input dimensions to the model. I choose to use networks with between one and three hidden layers, with layer sizes randomly chosen uniformly in the range $[0, 500]$. This range of layer sizes was chosen as being one order of magnitude larger or smaller than the input dimensionality while also remaining computationally feasible. The full set of varying parameters in this chapter are:

1. Number of hidden layers uniformly in the linear range $[1, 3]$
2. Size of each layer uniformly in the linear range $[0, 500]$
3. Training batch size uniformly in the linear range $[30, 500]$
4. Number of training epochs uniformly in the range $[50, 500]$
5. Initial learning rate uniformly in the log range $[10^{-4}, 10^{-2}]$
6. Learning rate either constant and adaptive with equal chance
7. Activation from RELU, TANH, or LOGISTIC with equal chance
8. Regularization parameter α from equation Equation (2.36) uniformly in the log range $[10^{-4}, 10^{-2}]$

These ranges match those suggested by the SCIKIT-LEARN website Pedregosa et al. (2011). I use fixed default values for the 'adam' parameters BETA_1 $= 0.9$, BETA_2 $= 0.999$, EPSILON $= 1e-08$ and TOL $= 0.0001$.

4.2.2 Input and output scaling

The input features to my MLP models are the correlation function measurements $\xi^{(3)}(r)$ for a range of different triangle sizes r. These correlation function values span a wide range of magnitudes. I use the MinMaxScaler method from SCIKIT-LEARN to normalise separately each correlation function bin. I also compare the effect of scaling the correlation function values by four different powers of the

binned radius values: the raw correlation functions $\xi^{(3)}$; the dimensionless correlation function $r^3\xi^{(3)}(r)$ used for more natural visualisations (see for instance Hoffmann et al. 2018); and two other powers of the radius for completeness: $r\xi^{(3)}(r)$ and $r^2\xi^{(3)}(r)$. The output features to my MLP models are either the bubble sizes $R_{\rm bubble}$ or the global ionisation fraction $\langle x_{\rm HII}\rangle$. I scale the $R_{\rm bubble}$ function using the $\sinh^{-1}$ function as described by Lupton et al. (1999).

4.3 Learning typical bubble sizes from the 3PCF

The progress of the Epoch of Reionization can be tracked by measuring the typical size of ionised regions. Ionised regions are initially small and isolated around the earliest ionising sources. The regions continually grow throughout the EoR, and the precise details of this continued growth depends on the physical interactions between ionising sources and the surrounding neutral regions. The sources themselves are seeded from the clustered non-linear density field and so show significant clustering (Hultman Kramer et al., 2006), but the details of reionization also affect the clustering of the resulting ionisation fraction field $x_{\rm HII}(\mathbf{r})$ and 21cm brightness temperature field $\delta T_{\rm b}(\mathbf{r})$. Throughout the EoR, the typical bubble size will likely boost the 3PCF at characteristic triangle sizes. Thus, the 3PCF contains information about the physics of reionization (Mcquinn et al., 2006). Similarly, higher-order clustering statistics contain information about the physical reionization parameters (see for instance Shimabukuro et al. 2017b) which affect the morphology of the $x_{\rm HII}(\mathbf{r})$ and $\delta T_{\rm b}(\mathbf{r})$ fields.

In this section, I train MLP models to predict the typical bubble size using the 3PCF from simulated data. First, I use correlation function measurements of the ionisation fraction field $x_{\rm HII}(\mathbf{r})$ to train my MLP models. The resulting model is a useful means of determining whether $\xi^{(3)}$ does indeed contain information about the typical bubble size. In practice, however, the ionisation fraction field $x_{\rm HII}(\mathbf{r})$ is difficult to disentangle from the actual results of interferometer experiments. In the second half of this section I train MLP models to predict the typical bubble size using simulated $\delta T_{\rm b}(\mathbf{r})$ data, which would be directly available from interferometer

observations.

4.3.1 Data cleaning

Data processing is a vital step in fitting machine learning models to real-world data. Raw data are often noisy and subject to systematic biases that can interfere with the model's ability to understand relationships in the data. In Section 4.1.3 I discuss grouping the three-point correlation triangle configurations into bins of similarly-sized triangles. In Section 4.1.4 I discuss collecting the distribution of mean-free paths travelled by ionising photons and binning these distances to match the pixel size. Throughout this chapter, I use one further data cleaning step. Near the end of the EoR, ionised bubbles grow to become extremely large and, due to widespread overlap, the typical bubble size becomes less clearly identifiable from the ionisation fraction field $x_{\rm HII}(\mathbf{r})$ (see Elbers and Van De Weygaert 2018 for classifications on EoR overlap regimes). I exclude data with global ionisation fraction outside the range $0.01 \leq x_{\rm HII} \leq 0.95$ to mitigate this effect.

4.3.2 Results training on $x_{\rm HII}(\mathbf{r})$ data

This subsection presents the results of training a model to learn how the typical bubble size is related to the 3PCF of ionisation fraction data $x_{\rm HII}(\mathbf{r})$. My training and testing data are from the range of simulated reionization scenarios described in Section 4.1.2, and I use the multilayer perceptron model described in Section 2.4.3. Figure 4.2 shows an example of the measured $x_{\rm HII}(\mathbf{r})$ 3PCF for a range of redshifts, showing the true typical bubble size as vertical lines. This figure is for a scenario with canonical parameter values $T_{\rm vir} = 10^4$K, $\zeta_{\rm ion} = 30$ and $E_0 = 200$ eV. The small-scale amplitude of the 3PCF decreases continually, and the amplitude on larger scale increases continually. The turnover radius at intermediate scales also increases throughout the EoR.

Input scaling types

Before running a full hyperparameter search, I first compare the different input-scaling types. The MLP models for this test all have the same architecture, namely two hidden layers both containing 100 nodes. The following values are used for the

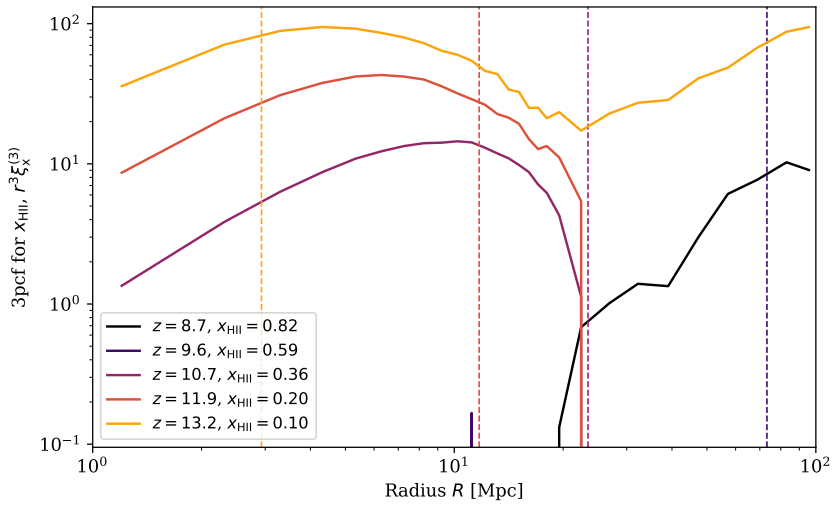

Figure 4.2: Example measurements of $r^3\xi^{(3)}$ for ionisation fraction field data $x_{\rm HII}(\mathbf{r})$. Each line shows the measured statistic for a single redshift, all taken from a simulation with $\zeta_{\rm ion} = 30.0$, $T_{\rm vir} = 10^4$K and $E_0 = 200$ eV. The redshifts and corresponding global ionisation fraction are shown for each line in the legend. Vertical dashed lines indicate the typical bubble size from mean free path measurements. The typical bubble size for $z = 8.7$ is too large to be shown on this figure.

other hyperparameters: a training batch size of 200; 200 maximum epochs; a constant learning rate of 10^{-3}; the RELU activation function; and fixed regularization parameter $\alpha = 10^{-3}$. These hyperparameters were chosen as the midpoints of the allowed random-search ranges or, for categorical choices, as the default parameters suggested by the code authors (Pedregosa et al., 2011). I train one model for each of the four possible input scaling types, namely $\xi^{(3)}$, $r\xi^{(3)}$, $r^2\xi^{(3)}$ and $r^3\xi^{(3)}$. Table 4.1 shows the resulting overall RMSE values for MLP models using each of the four different scaling types. My results indicate that scaling the 3PCF by r^2 or r^3 generates more accurate predictions than scaling by r or not scaling at all. Using $\xi^{(3)}$ or $r\xi^{(3)}$ as inputs makes it harder for my MLP models to uncover a relationship between the correlation function and the typical bubble size. Hereafter I use $r^2\xi^{(3)}(r)$ as inputs to my MLP models, since this choice gave the minimum RMSE value.

Figure 4.3 shows a more detailed description of these MLP models' prediction accuracies. This figure shows histograms of the prediction errors from each model,

Input scaling	RMSE
$\xi^{(3)}$	1.170
$r\xi^{(3)}$	0.973
$r^2\xi^{(3)}$	0.791
$r^3\xi^{(3)}$	0.806

Table 4.1: RMSE performance of four different input scaling types on unseen test data. The model using $r^2\xi^{(3)}$ inputs has the best performance, with the two lowest powers of r having the worst performance. These RMSE values are only for a single cross-validated model with the fixed hyperparameters given in Section 4.3.2, but this indicates that the relationship between $r^2\xi^{(3)}$ and the typical bubble size is easier to learn than the other inputs.

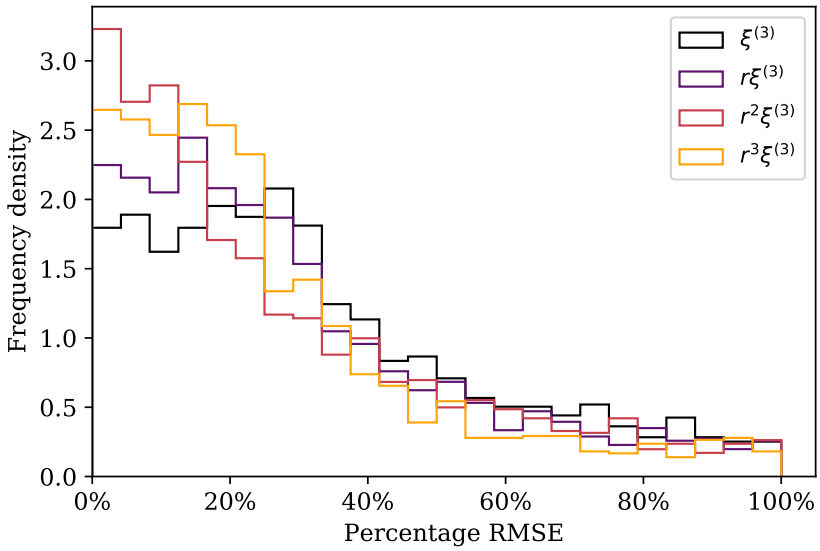

Figure 4.3: Histogram of prediction errors for different radius-value scaling types: $\xi^{(3)}$, $r\xi^{(3)}$, $r^2\xi^{(3)}$ and $r^3\xi^{(3)}$. Better model curves have errors weighted towards the left side of the histograms. In particular, there is a clear order of model performance accuracy seen in the relative frequencies of low-error predictions, indicated by the y-intercept of each line. This order of performance accuracies agrees with the overall RMSE values given in Table 4.1: the best $r^2\xi^{(3)}$ model has the highest frequency of low-error predictions; the r^3 model has the next highest frequency; and so on.

and is described in Section 2.4.8. In particular, the frequencies of low-error predictions (i.e. the y-intercept of each line) agrees with the order of model qualities in Table 4.1.

Best final model

I now find the best MLP model for predicting typical bubble sizes from the 3PCF of ioinsation fraction field data $x_{\rm HII}(\mathbf{r})$. I use the full hyperparameter search method described in Section 4.2.1, comparing 1000 randomly chosen MLP models and selecting the one with best cross-validated performance. The resulting best MLP model uses three hidden layers with sizes [198, 194, 125]; training batch size of 230; a maximum of 990 epochs (of which the model used all 990 epochs before terminating); constant learning rate starting at 1.11×10^{-3}; the RELU activation function; and L2 regularization parameter 9.24×10^{-4}.

Figure 4.4 shows the accuracy of the best MLP model's predictions for unseen testing data. I plot all predicted $R_{\rm bubble}$ values as a function of the true values. Marker colours are used to indicate the value of $\langle x_{\rm HII} \rangle (z)$ for each measurement. A model with perfect predictions would lie exactly on the dotted black diagonal line. Deviations from this diagonal represents less accurate predictions. The accuracy of the model depends strongly on the magnitude of the true bubble size. Two interesting features stand out in this figure which I discuss here.

First, the model struggles to make accurate predictions for typical bubble sizes that are larger than 100 Mpc: predictions for $R_{\rm bubble} < 100$ Mpc lie close to the diagonal, but predictions for $R_{\rm bubble} > 100$ Mpc show much larger scatter. This can be understood in terms of the relationship between the 3PCF and the typical bubble size. Near the end of the EoR, the widespread overlap of ionised bubbles gives rise to a larger average mean free path of ionising photons, but also blurs the definition of a typical bubble size. Many bubbles have merged, and thus the 'typical' bubble size is a less clear feature. The model's lessened ability to learn the typical bubble size from 3PCF measurements reflects this.

The second feature is the short vertical line of markers in the bottom-left of Figure 4.4, at true bubble sizes of $R_{\rm bubble} = 0$. My model predicts a range of typical bubble sizes from 0 Mpc up to around 40 Mpc in these scenarios, despite the true typical bubble sizes being consistently zero. These scenarios all have extremely low $\langle x_{\rm HII} \rangle$ values, as indicated by the marker colours. This feature is likely due to the

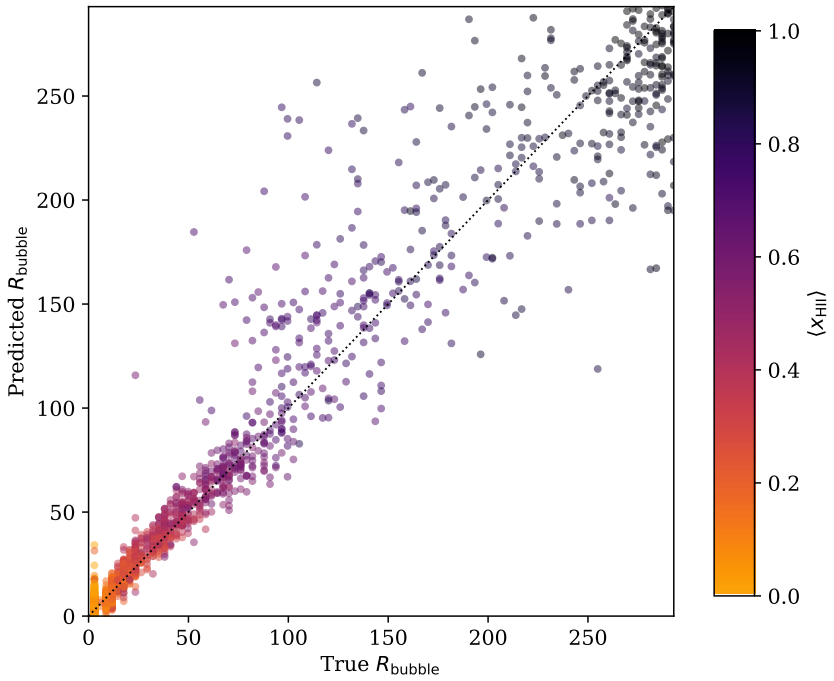

Figure 4.4: Predicted bubble size vs true bubble size for the best MLP model in Section 4.3.2. These predictions are made on unseen testing data, using only the 3PCF of ionisation fraction field data as inputs to the model. The predicted values and true values generally lie along the diagonal for values of $R_{\text{bubble}} < 100$ Mpc. Larger typical bubble sizes are harder to model and show much lager scatter away from the diagonal, as discussed in the text.

different measurement precisions of the mean-free path peak radius R_{max} and of the 3PCF from my code. My mean free path code is only able to measure bubble sizes in multiples of the simulations pixel size, specifically in multiples of 2.93 Mpc. Measurements of the typical bubble size are much noisier in these scenarios which contaminates the training data: a wide spread of inputs (the 3PCF curves) apparently lead to the same output value (the typical bubble size).

Figure 4.7 later shows the distribution of errors predicted by this model. The median prediction error from these distributions is a good measure of model performance. The model for predicting typical bubble sizes from ionisation fraction 3PCF here has a median prediction error of 19.9%.

4.3.3 Results training on $\delta T_{\text{b}}(\mathbf{r})$ data

The situation is considerably more complicated when using measurements of the 21cm differential brightness temperature field $\delta T_{\text{b}}(\mathbf{r})$ instead of the ionisation frac-

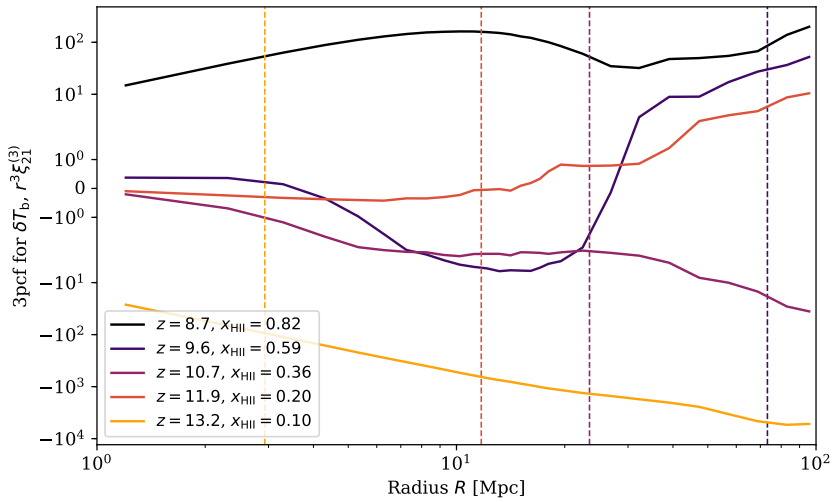

Figure 4.5: Example measurements of $r^3\xi^{(3)}$ for 21cm differential brightness temperature field data $\delta T_b(\mathbf{r})$, using the same simulation as Figure 4.2. Vertical dashed lines again indicate the typical bubble size from mean free path measurements, and the typical bubble size for $z = 8.7$ is too large to be visible.

tion field $x_{HII}(\mathbf{r})$. The 21CMFAST relationship between δT_b and the ionisation fraction x_{HII} given in Equation (1.21) is assumed to be linear, but the other terms in this equation also impact the morphology of the 21cm brightness temperature field. Most notably, local spin temperature fluctuations $T_S(\mathbf{r})$ and local density fluctuations $\delta(\mathbf{r})$ can both change the local values of $\delta T_b(\mathbf{r})$. Fluctuations in these values confuse the otherwise simple relationship between the 3PCF and the typical bubble size. Figure 4.5 shows the measured $\delta T_b(\mathbf{r})$ correlation function from a simulation with parameters $\zeta_{ion} = 30.0$, $T_{vir} = 10^4$K and $E_0 = 200$ eV. The true typical bubble sizes are shown as vertical lines. The 3PCF of the brightness temperature data has a more complex evolution over the EoR. The complex evolution of features in this figure are far less obvious than the equivalent figure for the ionisation fraction 3PCF, justifying the need for machine learning models.

Using the same method as for the ionised fraction field model, I train a model to predict the typical bubble sizes using the 3PCF of simulated $\delta T_b(\mathbf{r})$ data. The resulting best MLP model uses three hidden layers with sizes [129, 85, 141]; training batch size of 92; a maximum of 485 epochs; adaptive learning rate starting at 2.54×10^{-3}; the RELU activation function; and L2 regularlization parameter

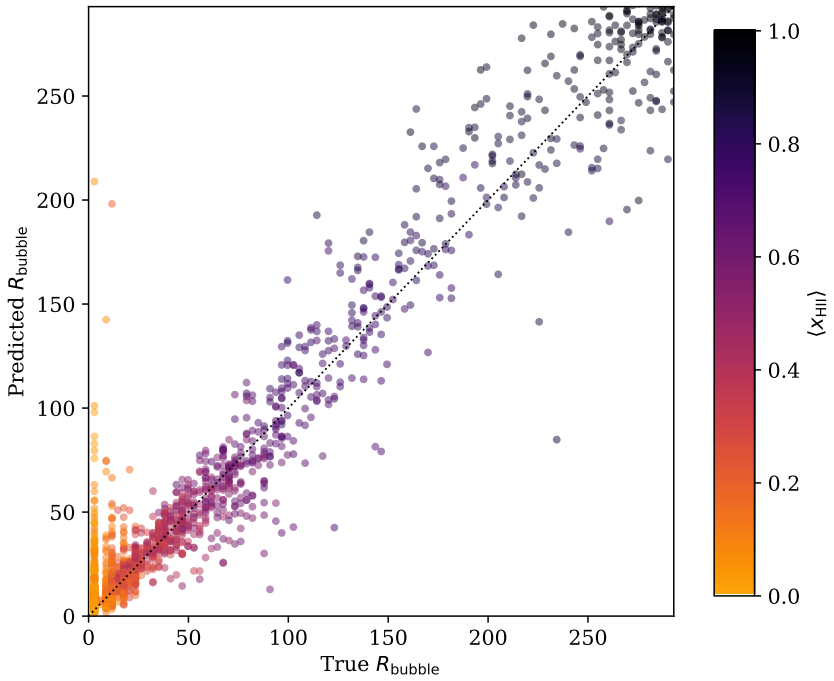

Figure 4.6: Predicted bubble size vs true bubble size for unseen testing data, using the best MLP model in Section 4.3.3. This model uses the 3PCF of $\delta T_{\text{b}}(\mathbf{r})$ data to predict the typical bubble size. The predicted values and true values generally lie along the main diagonal for middling values of R_{bubble} between 25 and 110 Mpc. The model can accurately predict the typical bubble size in these scenarios. Deviations from the diagonal line at larger and smaller bubble sizes are worse for the reasons discussed in the text.

1.68×10^{-4}.

The $\delta T_{\text{b}}(\mathbf{r})$ model is generally slightly worse than the $x_{\text{HII}}(\mathbf{r})$ model in the previous section. The median prediction error of this $\delta T_{\text{b}}(\mathbf{r})$ model is 29.3%, less accurate than the $x_{\text{HII}}(\mathbf{r})$ model's value of 19.9%. I plot the $\delta T_{\text{b}}(\mathbf{r})$ model's predicted typical bubble sizes for unseen testing data in Figure 4.6, as a function of the true typical bubble sizes. The general trend of predictions appears similar to the model using $x_{\text{HII}}(\mathbf{r})$ data: the predictions generally lie along the perfect-model diagonal, with poorer performance for both very large bubbles ($R_{\text{bubble}} > 100$ Mpc) and zero-sized bubbles ($R_{\text{bubble}} = 0$). My model is still able to recover enough information from $\delta T_{\text{b}}(\mathbf{r})$ to make fairly accurate predictions of the typical bubble size, despite the added complexities of the density and spin temperature fields.

Two features of Figure 4.6 show different behaviour from the $x_{\text{HII}}(\mathbf{r})$ model.

First, the larger scatter of points around the diagonal confirms the poorer predictions indicated by the worse RMSE value. This shows that the extra complexities of including local spin temperature fluctuations and local density field fluctuations do indeed contaminate the relationship between the typical bubble size and the data-field correlations. The model cannot distinguish between correlations of ionised regions and correlations of low density contrast regions ('under-dense' regions), because both of these scenarios give rise to lower values for $\delta T_{\rm b}$. Similarly, regions with low local values for the spin temperature $T_{\rm S}$ can mimic ionised regions.

Secondly, the performance of the $\delta T_{\rm b}(\mathbf{r})$ model appears to be more consistent across intermediate bubble sizes than the previous $x_{\rm HII}(\mathbf{r})$ model. In particular, using brightness temperature data appears to give better predictions for large typical bubble sizes ($R_{\rm bubble} > 100$ Mpc). It is not immediately obvious why this is the case. It is possible that including the effect of fluctuations in the spin temperature and density fields enhances 3PCF features in the ionisation fraction data. Such local fluctuations could increase the amplitude of the $\delta T_{\rm b}$ field and make it easier to pick out ionised bubbles.

Finally, Figure 4.7 shows the histograms of prediction errors for both final best MLP models: one using $x_{\rm HII}(\mathbf{r})$ data, and one using $\delta T_{\rm b}(\mathbf{r})$ data. Ideally, all predictions would be near zero. The distribution of errors for these two MLP models does not depend strongly on which data are used ($x_{\rm HII}(\mathbf{r})$ or $\delta T_{\rm b}(\mathbf{r})$) although, as mentioned above, each model does has different prediction accuracies for different typical bubble size regimes and the $\delta T_{\rm b}(\mathbf{r})$ model has a slightly higher median RMSE value (29.3%) than the $x_{\rm HII}(\mathbf{r})$ model's RMSE value (19.9%).

4.4 Learning global ionisation fractions from 3PCF

In the previous section I trained models to predict the typical bubble size from 3PCF measurements. The typical bubble size is a useful metric for tracking the growth of ionised regions, but the global ionisation fraction $\langle x_{\rm HII}\rangle(z)$ is a more direct measurement for the overall progress of the Epoch of Reionization. The historical evolution of $\langle x_{\rm HII}\rangle$ can be strongly affected by the reionization parameters. Different

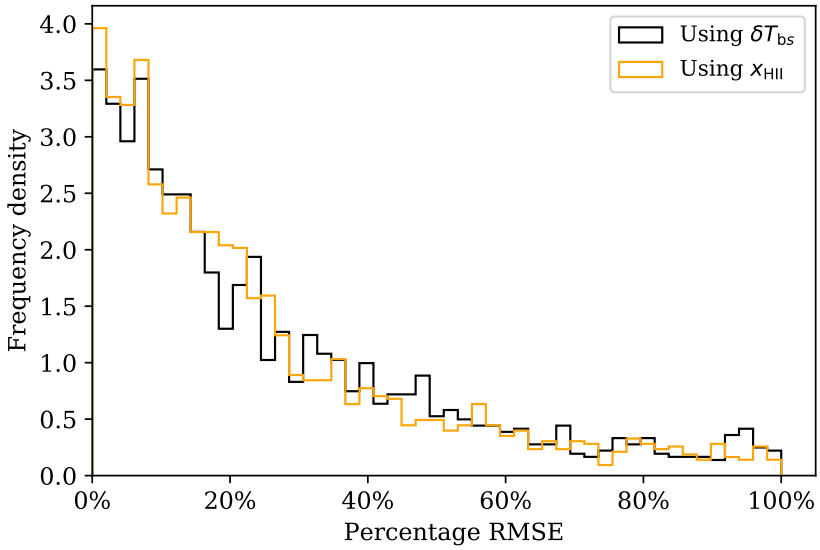

Figure 4.7: Histogram of prediction errors for typical bubble size models. Visibly, the overall distribution of errors does not depend strongly on which data are used, although the model using $x_{\rm HII}(\mathbf{r})$ data has a better median prediction error (19.9%) than the model using $\delta T_{\rm b}(\mathbf{r})$ (29.3%). The reasons for this poorer performance are discussed in the text.

ionising efficiency $\zeta_{\rm ion}$ scenarios have a different abundance of ionising photons, which affects the EoR durations. Different $T_{\rm vir}$ scenarios have different quantities of ionising sources which can also affect the EoR duration. In this section, I train models to predict the value of $\langle x_{\rm HII}\rangle(z)$ from 3PCF measurements. My models learn the relationship between the 3PCF and the global ionisation fraction by using the same simulated data in Section 4.3. Measurements of the 3PCF and mean free path use the methods described in Sections 4.1.3 and 4.1.4. The data are cleaned using the same ionisation fraction filters, namely $0.01 \leq \langle x_{\rm HII}\rangle \leq 0.95$.

4.4.1 Results training on $x_{\rm HII}(\mathbf{r})$ data

I first train a model to predict the global ionisation fraction $\langle x_{\rm HII}\rangle$ from the 3PCF of $x_{\rm HII}(\mathbf{r})$ data, using the same search strategy as in the previous section. The best MLP model uses three hidden layers with sizes [192, 150, 50]; training batch size of 261; a maximum of 365 epochs (of which the model used all epochs before terminating); adaptive learning rate starting at 2.00×10^{-3}; the RELU activation function; and L2 regularization parameter 3.72×10^{-4}.

The model has an extremely good median prediction error of 3.6%. Figure 4.8 indicates the performance from this model, showing the predicted values of $\langle x_{\rm HII} \rangle$ as a function of the true $\langle x_{\rm HII} \rangle$ values in the testing data. Marker colours show the typical bubble size. All markers lie close to the perfect-model diagonal in Figure 4.8, confirming that this model makes extremely accurate predictions. As in the previous section, the model accuracy is higher for $\langle x_{\rm HII} \rangle < 0.6$ than for $\langle x_{\rm HII} \rangle > 0.6$, with separate test RMSE for both regimes.

Ionisation fraction 3PCF measurements have a very strong relationship with the global ionisation fraction. Ionisation fraction field data contain a range of bubble sizes, and the total volume of these bubbles is an indicator of the global ionisation fraction. The 3PCF measures clustering on a range of scales and this information is apparently strong enough to provide immediate and accurate predictions for the mean ionisation fraction. The predictions begin to worsen near the end of the EoR for $\langle x_{\rm HII} \rangle > 0.6$, when overlap means that the total volume of bubbles cannot be used as an immediate estimate of the global ionisation fraction.

4.4.2 Results training on $\delta T_{\rm b}$ data

I now train a model to predict the global ionisation fraction $\langle x_{\rm HII} \rangle$ from $\delta T_{\rm b}(\mathbf{r})$ 3PCF data. I use the same search strategy as the previous subsections. The best MLP model uses three hidden layers with sizes [168, 174, 70]; training batch size of 361; a maximum of 506 epochs (of which the model used all epochs before terminating); adaptive learning rate starting at 4.44×10^{-3}; the RELU activation function; and L2 regularization parameter 3.65×10^{-3}.

The results show that it is more difficult to predict the global ionisation fraction using $\delta T_{\rm b}(\mathbf{r})$ 3PCF data than using $x_{\rm HII}(\mathbf{r})$ data. The $\delta T_{\rm b}(\mathbf{r})$ model's median prediction error is 16.0%, much worse than the error of 3.6% for the $x_{\rm HII}(\mathbf{r})$ model. Figure 4.9 gives the final prediction histograms for the two global ionisation fraction models, using either ionisation fraction data or brightness temperature field data. Predictions of the global ionisation fraction depend strongly on which data are used: the prediction errors for the model using $x_{\rm HII}(\mathbf{r})$ data are much lower than those for the $\delta T_{\rm b}(\mathbf{r})$ model.

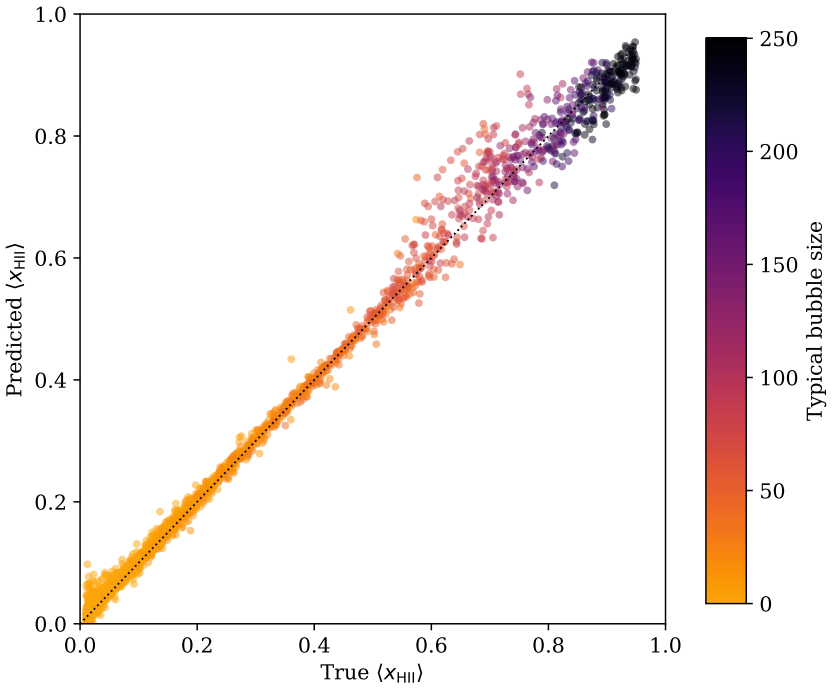

Figure 4.8: Predicted global ionisation fraction vs true global ionisation fraction for unseen testing data, using the ionisation fraction $x_{\rm HII}(\mathbf{r})$ as inputs. The predicted and true values lie very closely along the diagonal, particularly for values $\langle x_{\rm HII} \rangle < 0.6$. Predictions for $\langle x_{\rm HII} \rangle > 0.6$ are slightly worse as discussed in the text.

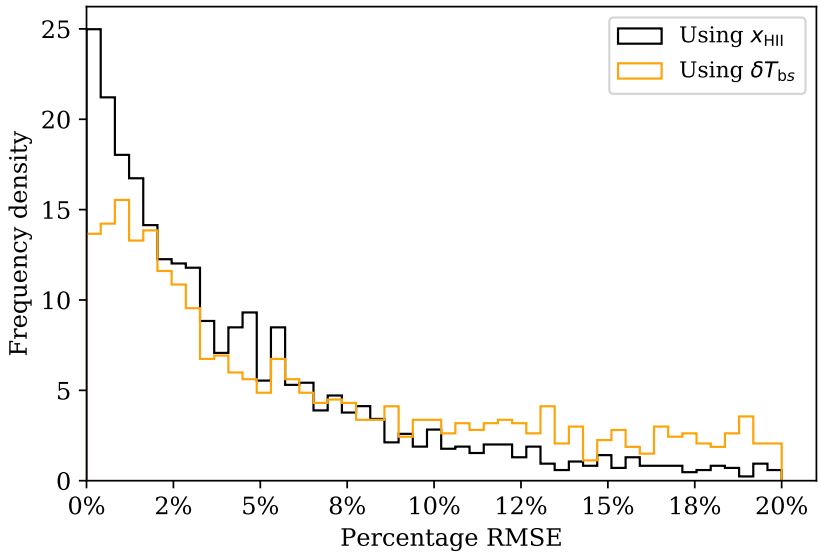

Figure 4.9: Histogram of prediction errors for predicting the global ionisation fraction. Each line shows the histogram of errors for a single model. The model using $x_{\rm HII}(\mathbf{r})$ 3PCF data has a much more accurate median prediction error (3.6%) than the model using $\delta T_{\rm b}(\mathbf{r})$ data (16.0%).

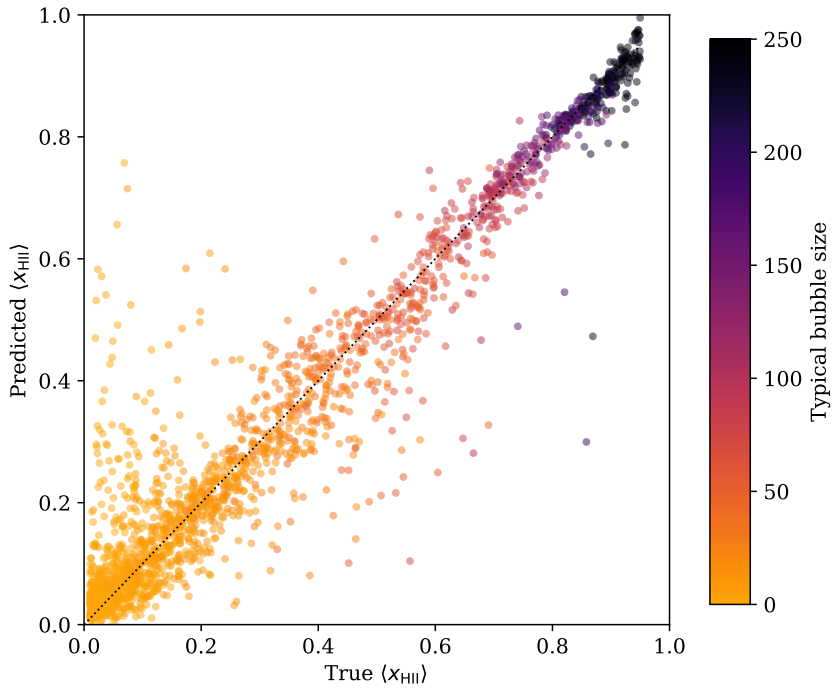

Figure 4.10: Predicted global ionisation fraction vs true global ionisation fraction for unseen testing data, using $\delta T_{\rm b}(\mathbf{r})$ as inputs. The predictions generally lie along the diagonal, but with larger scatter than using $x_{\rm HII}(\mathbf{r})$ as model inputs.

Additionally, the model predictions shown in Figure 4.10 deviate more widely from the perfect diagonal than do the predictions in Figure 4.8. For the later stages of the EoR with $\langle x_{\rm HII}\rangle > 0.6$, the $\delta T_{\rm b}(\mathbf{r})$ model's accuracy *increases* as opposed to decreasing as did the accuracy of the model using $x_{\rm HII}(\mathbf{r})$ 3PCF data. This can be understood by considering the impact of density and spin temperature fluctuations. Local fluctuations have a more significant impact on the $\delta T_{\rm b}(\mathbf{r})$ field at early times than at later times. Thus, the morphology of the $\delta T_{\rm b}(\mathbf{r})$ field is more closely linked to that of the $x_{\rm HII}(\mathbf{r})$ field at later times.

4.5 Conclusions

The three-point correlation function (3PCF) of the 21cm signal encodes valuable information about the morphology and history of the Epoch of Reionization. I use machine learning techniques and train models to recover the typical bubble size and global ionisation fraction from the measured 3PCF outputs from semi-numerical simulations. I first train models to recover the typical bubble size, from the 3PCF

of either ionisation fraction data or 21cm differential brightness temperature data. The two models are both able to determine the general trend of increasing typical bubble size and have similar overall accuracy, with median RMSE values of 19.9% and 29.3% respectively. The model using $x_{\rm HII}(\mathbf{r})$ 3PCF data has better performance at small bubble sizes (1 Mpc $< R_{\rm bubble} <$ 100 Mpc, whereas the model using $\delta T_{\rm b}(\mathbf{r})$ has better performance for larger bubble sizes ($R_{\rm bubble} >$ 100 Mpc). Both features can be understood in terms of how the data field morphologies evolve over the EoR. In particular, the morphology at early times is more strongly affected by fluctuations in the density field and the spin temperature field.

I also train a model to recover the global ionisation fraction from ionisation fraction 3PCF data. The resulting model has extremely accurate predictions and shows the three-point clustering of $x_{\rm HII}(\mathbf{r})$ data is strongly related to the evolution of the global ionisation fraction. My model is able to uncover this relationship with a median RMSE value of 3.6%, although the predictions are slightly less accurate for the later stages of the EoR with $\langle x_{\rm HII} \rangle > 0.6$. Unfortunately this model would practically not be useful in EoR analysis because the ionisation fraction field is difficult to probe directly. Instead, observations are made in terms of the differential brightness temperature. I train a fourth and final model to predict the global ionisation fraction from the 3PCF of the differential brightness temperature field. The resulting model has a median RMSE value of 16.0%, with more accurate predictions for the late stages of the EoR ($\langle x_{\rm HII} \rangle > 0.6$) than for the early stages.

As with all machine learning projects, my MLP models to predict the typical bubble size and global ionisation fraction could likely be improved by gathering more data from a wider range of reionizaion scenarios. This would allow the MLP models to learn more general connections between the 3PCF measurements and characteristic reionization features. Providing other brightness temperature field summary statistics could also improve the ability to uncover the EoR features, for instance the distribution of pixel brightnesses (Ichikawa et al., 2009) or the size distribution of bright regions (Kakiichi et al., 2017). My MLP models assume a constant value for the X-ray efficiency. Ideally this constraint should be lifted and

the X-ray efficiency allowed to vary as with the other simulation parameters. Before using any models on real data I would need to add instrumental noise and retrain them. Additionally, the mean free path method used to measure typical bubble sizes in this chapter is known to give less consistent results than other methods described in Section 2.5.1. Using another method such as granulometry (Kakiichi et al., 2017) could provide better estimates for R_{bubble} and make the relationship with the 3PCF more consistent.

There are several possible avenues of future work to build on my results. First, using similar machine learning techniques to predict the full bubble size distribution dP/dR from 3PCF data. The full bubble size distribution provides a more detailed description of the morphology than the typical bubble size alone. Secondly, training models to map from 3PCF measurements directly to parameters in a similar way to Shimabukuro and Semelin 2017. Such inference models can only make estimates of the 'best' parameters and do not provide uncertainty regions in the same way as MCMC analysis. Instead, training emulators to forward-model the 3PCF outputs directly from the simulation input parameters would effectively remove the need for further simulations.

This work presents the first attempt to predict fundamental properties of the Epoch of Reionization using the three-point correlation function and machine learning techniques. I have made my code publicly available to help the community perform similar analyses in the future.

Chapter 5

Analytic clustering model

The halo model described in Section 2.3 has been successfully used to make accurate predictions for the non-linear clustering of matter in the Universe (see for example Smith et al. 2002 and Cooray and Sheth 2002). A halo model can be made for any observable data field if the following ingredients are available: a profile function around sources; a number density distribution for the abundances of differently-sized sources; the clustering properties of the sources themselves. This chapter gives such a clustering model for the signal from the ionisation fraction field of hydrogen. The ingredients of the model are found by fitting results from semi-numerical simulations.

The rest of this chapter is structured as follows. In Section 5.1, I describe how to determine a model for the ionisation fraction profile by stacking and fitting the simulated data. I give a simple toy model for the clustering properties of randomly-placed spherical bubbles, described in Section 5.2. I then add more realistic properties to this model towards a full model for the ionisation fraction field given in Section 5.3, fitting intermediate results from SIMFAST21 simulations. The motivation for this project was as a middle ground between the simple FZH model (described in Section 2.1.1) and semi-numerical simulations, capturing the extra complexity of simulations without the higher computational cost. Section 5.4 takes the final step towards this goal by comparing the final model to actual outputs from SIMFAST21. I conclude this chapter in Section 5.5, describing what further adaptations to the model would be needed before being useful for predicting real-space correlation

functions such as the one in the previous chapter. The power spectrum calculations throughout this chapter make use of code provided by Watkinson (2017).

5.1 Stack-Fit-Predict method

The profile function $\rho(r,M)$ is a key ingredient of any halo model. The density field profiles represent the matter density centred around dark matter halos, usually the NFW profile as fitted from the results from N-body simulations (Navarro et al., 1996). Similar profiles must be found for the ionisation fraction field before making predictions for the ionisation fraction power spectrum. The profiles in this chapter are fitted to measured profiles from simulations. Fitting the profiles involves two stages: measuring the data field around halos of different masses; and fitting these measured profiles to an analytic formula so they can be used in the halo model equations.

5.1.1 Stacking

SIMFAST21 stores the locations and masses of every resolved halo in a catalog. Figure 5.1 shows a slice through a realisation of the ionisation fraction field $x_{\rm HII}(\mathbf{r})$ from SIMFAST21, with the corresponding halo locations indicated by red cross markers. The dark-regions show ionised bubbles are generally located around halos or clusters of halos.

The ionisation fraction profiles around halo centres can be measured from SIMFAST21 by stacking cubic regions around halo centres. The data and catalog are loaded for a particular redshift. Cubic regions of the ionisation fraction around all halos are extracted and sorted into bins of the underlying halo masses. Figure 5.2 shows the result of such a stacking procedure for the data in Figure 5.1. Each panel shows a slice through the mean ionisation fraction cube around halos of a unique mass. Spherically averaging these cubic regions gives measured samples of the radially-symmetric profile function $\rho(r;M,z)$ shown in Figure 5.3.

5.1.2 Fitting

The result of the stacking procedure is a set of sampled real-space profile values $\rho(r;z,M)$. The halo model however requires an analytic form $\widetilde{\rho}(k;z,M)$ for the

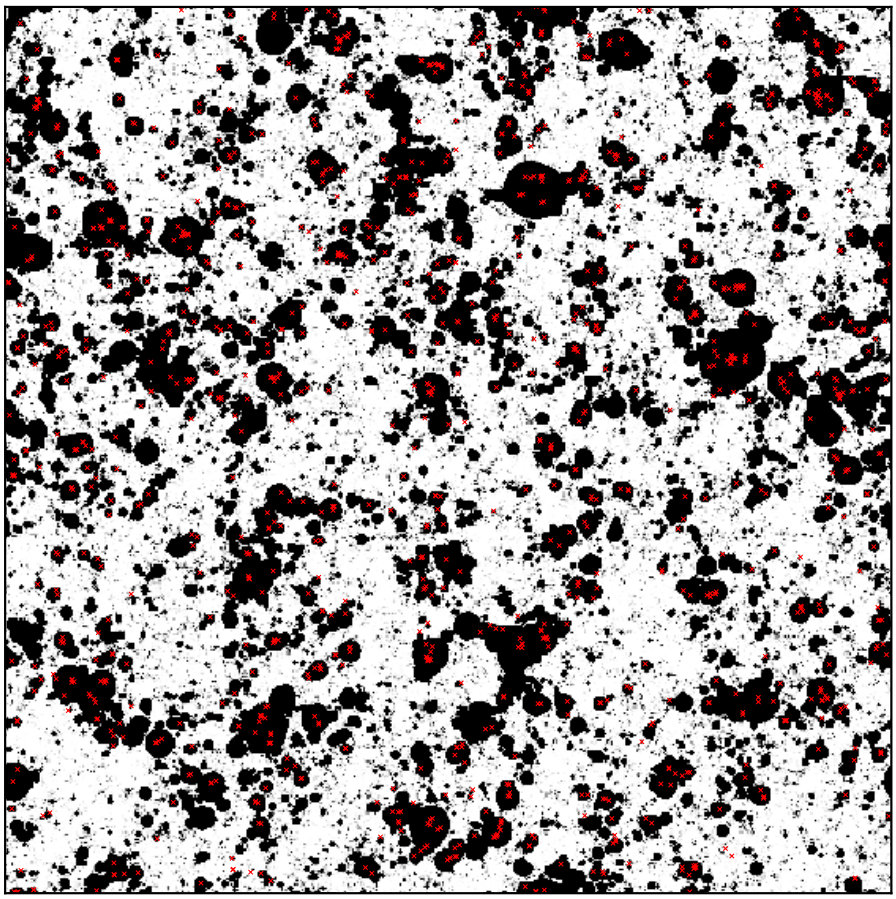

Figure 5.1: Slice through a SIMFAST21 realisation of the ionisation fraction field $x_{\rm HII}$ at $z = 11$, with halo centres shown as red crosses. Ionised bubbles are nucleated around halos and, in particular, around clusters of halos. The total length of this simulation is 500 Mpc/h.

Fourier-transformed profiles at any redshift and mass. The easiest way to find an analytic form for $\widetilde{\rho}(k;z,M)$ is to fit the sampled real-space profiles and then take the analytic Fourier transform. If the profiles are fitted to a specific function form $f(r)$ that is chosen to have an analytic Fourier-transform $\widetilde{f}(k)$, then fitting the profiles in real-space immediately allows for predicted $\widetilde{f}(k)$ profiles. The fitting procedure involves two main parts: choosing the generic form $f(r)$, and fitting this generic form to the measured profiles $\rho(r;z,M)$.

First, choosing the generic form for the profile. This is most easily done by observing the stacks and trying several possible generic forms. If the profiles appear to fall exponentially from a peak value, then $\rho(r) \propto \exp(-r)$ would be a good choice. I

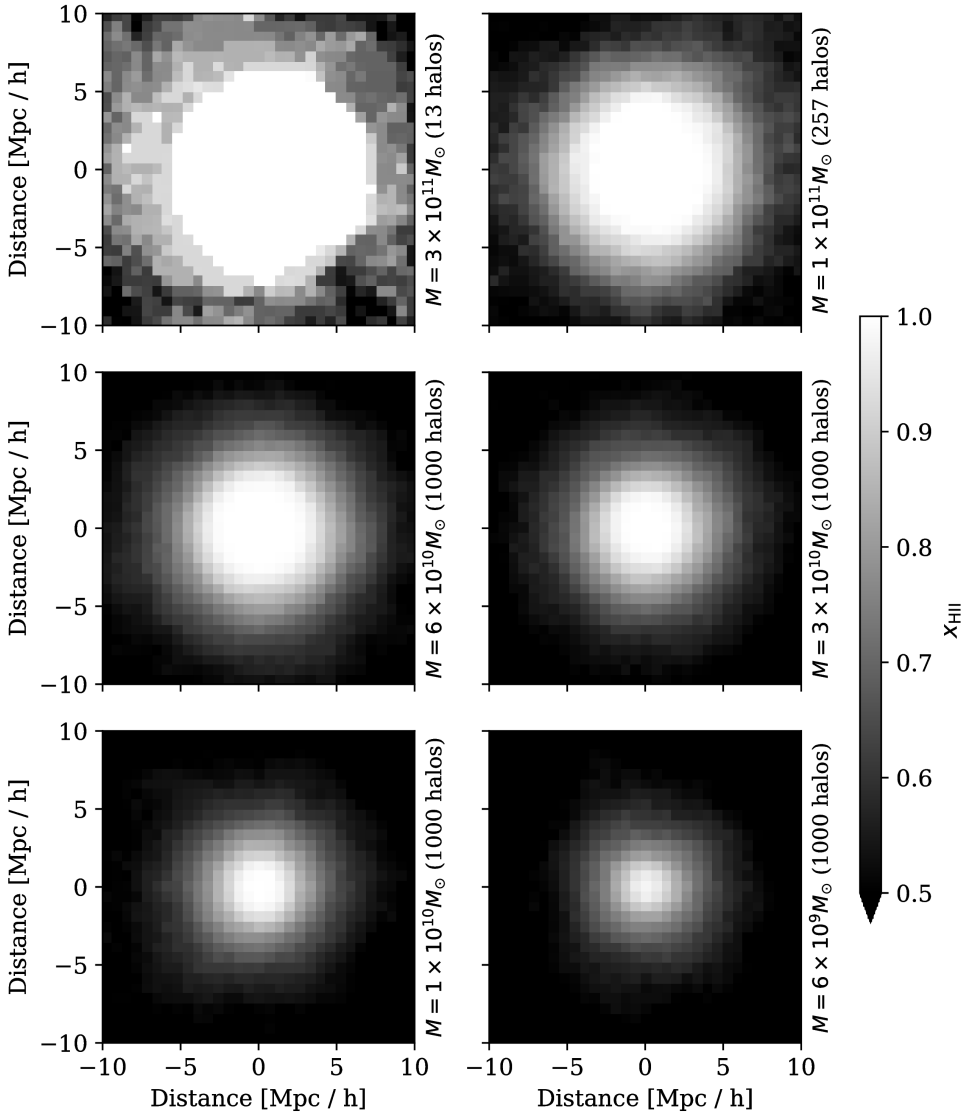

Figure 5.2: Slices through stacks of the ionisation fraction around halos from the data in Figure 5.1. Each stack shows the ionisation fraction around halos with a single unique mass. The label to the right of each panel indicates the the mass of that stack, as well as the number of matching halos over which the stacks was averaged. The ionised bubbles are larger for more massive halos.

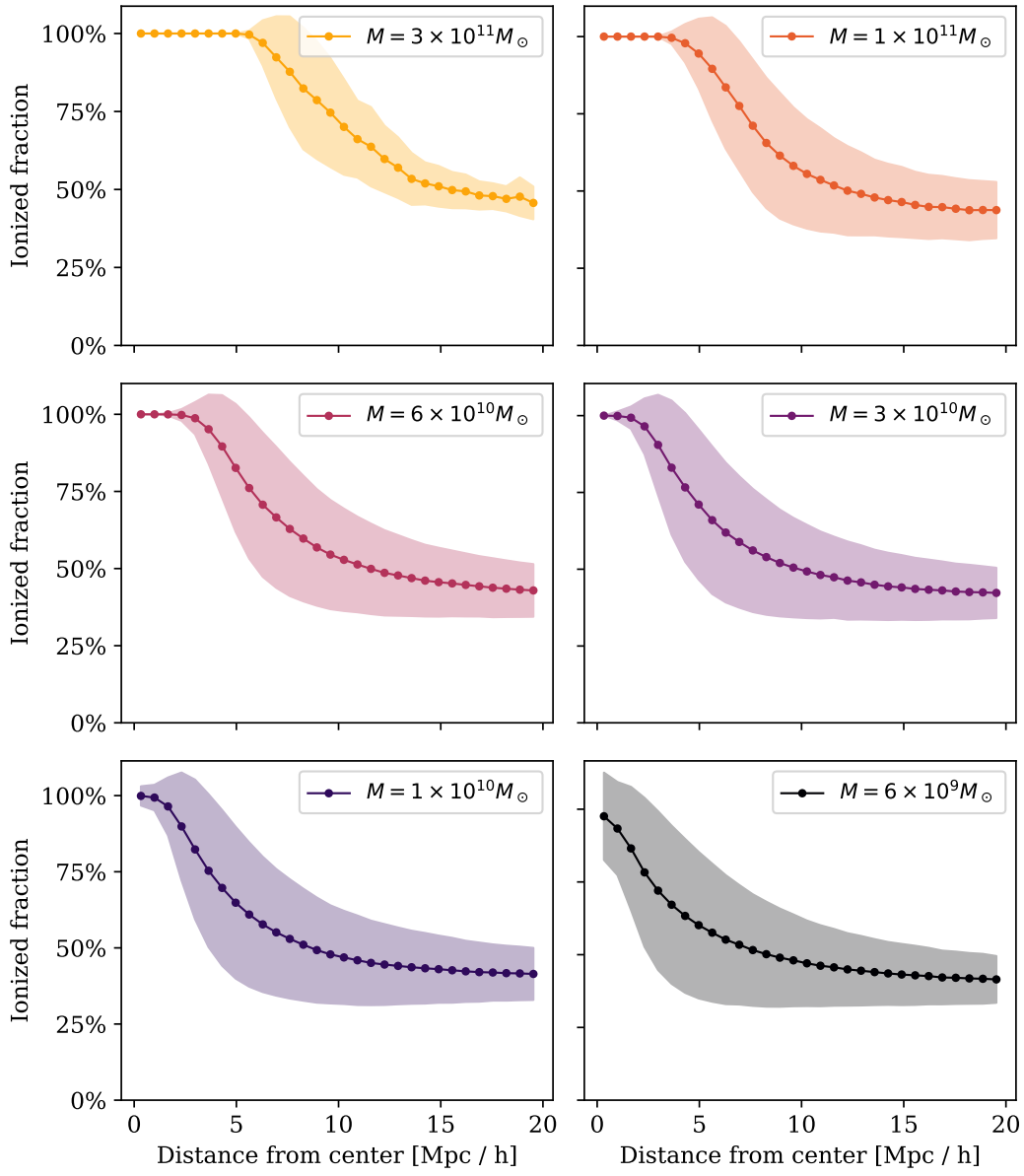

Figure 5.3: Measured ionisation fraction profiles around halo centres for the ionisation fraction data field in Figure 5.1. These are generated by spherically averaging the separate stacks in Figure 5.2, with the panels arranged in the same order. The shaded regions show the two-sigma spread in the measured profiles for each halo. The heaviest halo has a fully-ionised region for around 5 Mpc, seen as a plateau in the top-left panel. The ionised fraction then slowly falls to the global $\langle x_{\rm HII} \rangle$ at around 20 Mpc. The lightest halo (bottom-right panel) has no plateau, but immediately falls to the global $\langle x_{\rm HII} \rangle$ at a much smaller radius of around 10 Mpc. Intermediate masses show a consistent pattern of decreasing fully-ionised region size and earlier global-mean radius. This matches with the qualitative descriptions of size from the images in Figure 5.2.

add *features* to this generic form so that it becomes flexible enough to fit the profiles for several different masses and redshifts at once. The ionisation fraction profile of a halo might have a different maximum height or falloff distance depending on its mass or redshift. These two properties can be added as a peak height feature ρ_0 and a scale radius feature R. The generic form is then flexible enough to fit several profiles with different heights and falloff distances $\rho(r;R,\rho_0) = \rho_0 \exp(-r/R)$.

The sampled profiles are measured in terms of the halo mass M and redshift z. In order to fit these measured profiles $\rho(r;M,z)$ to the generic form $\rho(r;R,\rho_0)$, the physical features (R,ρ_0) must be related to the halo properties (M,z). For instance, the peak height $\rho_0(M,z)$ might be a function of the halo properties using a linear relationship $\rho_0(M,z) = \Theta_0 + \Theta_1 z + \Theta_2 \log(M)$. Similarly the scale radius might follow another linear relationship $R(M,z) = \Theta_3 + \Theta_4 z + \Theta_5 \log(M)$ with different coefficients Θ_i (also known as hyperparameters). I describe these relationships between the features and the halo properties as *feature relations*.

After the generic form and the feature relations have been chosen, fitting the measured profiles is done by finding the best values for all hyperparameters Θ_i. I use the NUMPY.OPTIMIZE.MINIMIZE function with the BFGS method to vary the values of all hyperparameters Θ_i and minimise the mean square error,

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\rho^{sampled} - \rho^{fitted}\right)^2, \tag{5.1}$$

between the sampled and fitted profiles. At each step in the minimisation process, the Θ_i values are used to convert the halo properties (M,z) into the required feature values (R,ρ_0). These feature values are then used in the generic profile form. The resulting fitted profiles are compared to the measured counterparts using Equation (5.1), and the hyperparameters updated to try and minimise any discrepancies.

This procedure is similar to the way in which NFW profiles were fitted to N-body simulations. A generic form

$$\rho_{\text{NFW}}(r) = \rho_s \left[\frac{r}{r_s}\right]^{-1} \left[1 + \frac{r}{r_s}\right]^{-2} \tag{5.2}$$

was chosen with two physical features: a scale radius $r_{\rm s}$ and a characteristic height $\rho_{\rm s}$. The feature $\rho_{\rm s}$ was in fact already constrained by mass conservation in the total integral of the profile $\int_0^{R_v} \rho(|\mathbf{r}|, M) d^3\mathbf{r} = M$. The scale radius feature $r_{\rm s}(M,z)$ was then fitted to power-law relations such as the one given in Equation (2.20).

5.1.3 Fitting & predicting: illustrative example

This subsection contains an illustrative example of the fitting and predicting procedure, using noisy mock profiles with known feature relations that I myself have chosen described in this subsection. I test my fitting procedure on these profiles, checking how well the fitted hyperparameters match the true hyperparameters that I chose at the start. Note that the profiles in this subsection are not from actual SIMFAST21 data, but were generated purely for demonstrative and testing purposes. I choose a mock profile function

$$\rho_x^{\rm sampled}(r;z,M) = \rho_0(M,z)\exp\left[-r/R(M,z)\right] + \text{Noise} \tag{5.3}$$

and mock feature relations

$$\begin{aligned} \rho_0(M,z) &= -0.40 + 0.10z \\ R(M,z) &= 10.00 + 3.00\log(M) \end{aligned} \tag{5.4}$$

so that the redshift z controls only the peak height of these mock profiles, and the mass M controls only the radial extent of the profiles. Realistic noise is added to these noiseless mock profiles, in order to test the robustness of the fitting procedure on profiles measured from data. For each curve, Gaussian noise $\mathcal{N}(0, 0.01)$ is added to the profile value for each radius. The resulting curve is then smoothed using a Savitzky-Golay window filter (Savitzky and Golay, 1964). Figure 5.4 shows these mock profiles for a grid of halo properties (z,M) with axes $z = \{8, 10, 12, 14\}$ and $M = \{10^8, 10^{10}, 10^{12}\} M_\odot$.

A generalised exponential-type profile function is used for fitting these profiles and to test the fitting procedure,

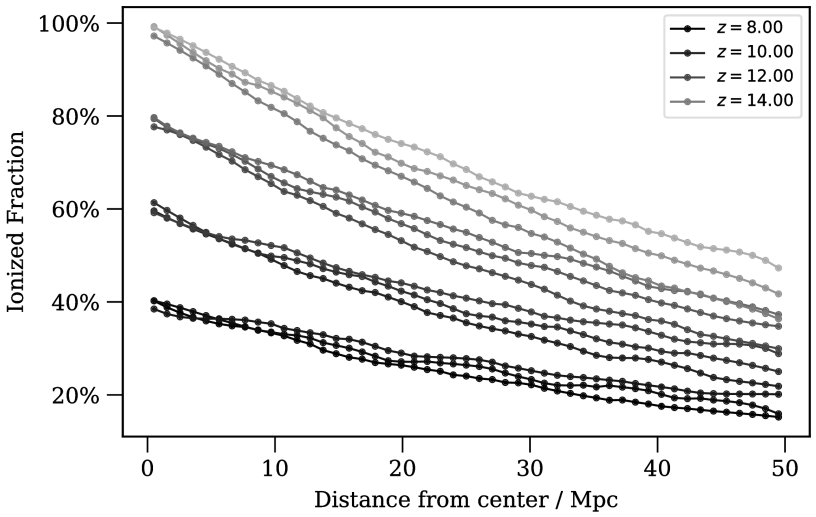

Figure 5.4: Mock profiles to demonstrate the fitting procedure, drawn from Equations 5.3 and 5.4. Random Gaussian noise was added to these mock profiles. These profiles are for four different redshifts (shown in the legend). Each redshift has profiles with three different masses $M = \{10^8, 10^{10}, 10^{12}\} M_\odot$. No error bars are shown because these mock profiles were generated (and not measured) using Equation (5.3).

$$\rho_x^{fitted}(r) = \rho_0 \, \exp(-r/R). \tag{5.5}$$

The physical features for this profile function are the peak-above-infinity height ρ_0 and the scale radius R. For fitting, each feature is allowed to depend on M and z by using the order-1 polynomial function,

$$\rho_0(M,z) = \Theta_0 + \Theta_1 z + \Theta_2 \log(M) \tag{5.6}$$

$$R(M,z) = \Theta_3 + \Theta_4 z + \Theta_5 \log(M). \tag{5.7}$$

Fitting involves finding the best values for the hyperparameters Θ_i, such that resulting fitted profile function $\rho(r)$ best matches the true profile values in Figure 5.4. Starting with random hyperparameters, Figure 5.5 shows the result of the fitting procedure for this example. Table 5.1 shows the best hyperparameter values as found by the fitting procedure, along with their correct values that I chose at the start. The

Hyperparameter	True value	Fitted value
Θ_0	-0.4	-0.402
Θ_1	0.1	0.100
Θ_2	0	0.000*
Θ_3	10	10.231
Θ_4	0	0.012
Θ_5	3	2.999

Table 5.1: Fitted and true hyperparameters for the example mock profiles. The fitted values are shown with three decimal places of precision. The values closely match the true sampled values. The asterisked (*) fitted value of Θ_2 was 6.1×10^{-6}.

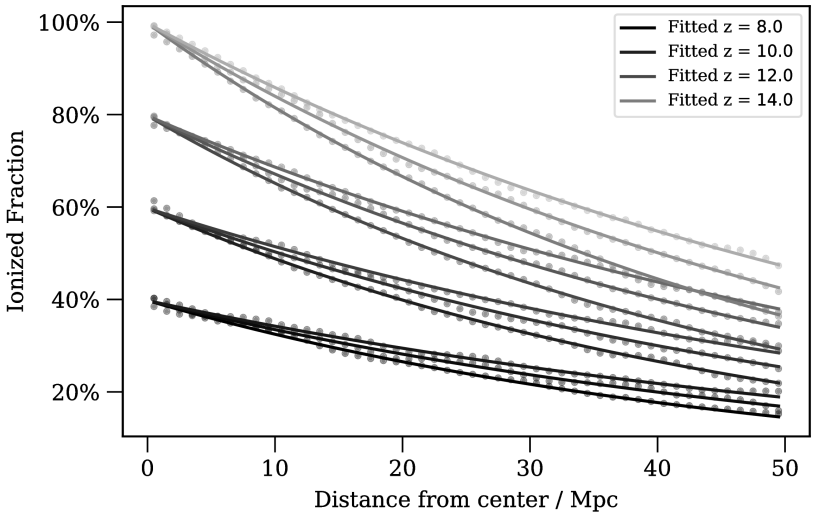

Figure 5.5: Mock profiles fitted using simple *exp*-type profile and linear feature relations. The fitted profiles lie closely on top of the stacked profiles showing that the fitting procedure has successfully fitted these data. As in Figure 5.4, no error bars are shown because the mock profiles (markers) were generated using Equation (5.3) and the predicted profiles (solid lines) were generated using Equation (5.8).

fitted values match the true values closely, showing that the fitting procedure works well for these profiles.

The final fitted profile function for these best hyperparameters is then given by

$$\rho_x^{\text{fitted}}(r,M) = [-0.402 + 0.100z]\ \exp\left(-r/\left[10.231 + 2.999\log(M) + 0.012z\right]\right) \tag{5.8}$$

which matches closely the true functional form used to generate these mock profiles

in Equations 5.3 and 5.4. Using this fitted profile in a halo model is now simple. Values for $\widetilde{\rho}_x(k;M,z)$ are found using the analytic Fourier-transform of Equation (5.5),

$$\widetilde{\rho}_x(k) = \frac{\rho_0(M,z)}{8\pi\left\{1+[kR(M,z)]^2\right\}^2}, \tag{5.9}$$

and the fitted physical feature relations,

$$\rho_0(M,z) = -0.402 + 0.100z + 0.00\log(M) \tag{5.10}$$

$$R(M,z) = 10.231 + 0.012z + 2.999\log(M). \tag{5.11}$$

Figure 5.6 shows the predicted Fourier-transformed profiles from Equation (5.9). Fourier-transformed profiles can also be predicted for (z,M) values outside the range of the originally fitted profiles. These extrapolated profiles are shown as dotted lines in Figure 5.6 and clearly extrapolate the correct patterns in both peak height and falloff radius. The final generic Fourier-transformed profile function can thus be used in the 1-halo and 2-halo terms and give predictions of the overall power spectrum.

Note that these mock profiles fall to zero as $r \to \infty$. In general I remove the global data mean value $\rho_x(r \to \infty)$ before fitting the profiles to a functional form, in order to match the normalisation method of the power spectrum calculation. The peak feature ρ_0 is then actually a 'peak-above-infinity' feature,

$$\rho_0 \equiv \rho_x(r=0) - \rho_x(r \to \infty) \tag{5.12}$$

5.1.4 Separate and joint fitting

This simple fitting procedure in the previous subsection turns out to be extremely difficult with anything other than very simple feature relations. A preparation step can be used to give good initial estimates for the hyperparameters Θ_i before running the joint fitting routine above. Each individual measured profile is first fitted to the generic form entirely independently from the other profiles. This can be done

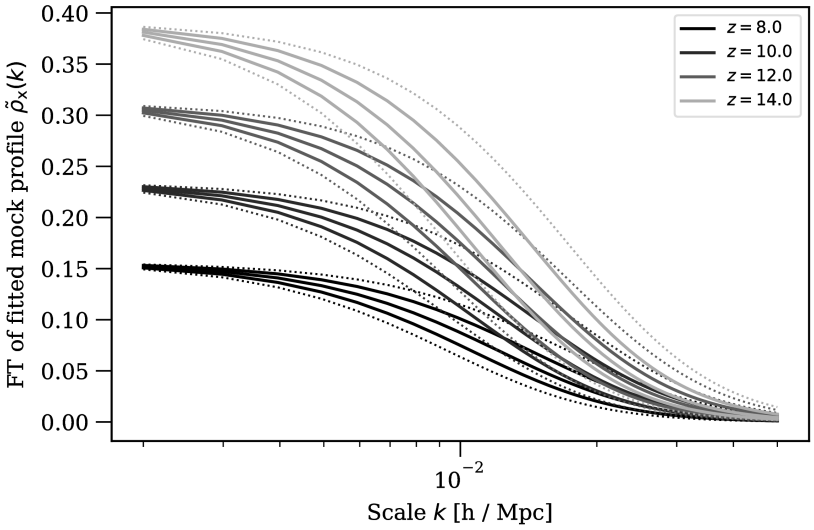

Figure 5.6: Fourier transform of the fitted mock profiles. These Fourier transformed profiles can be generated for any (M,z) pair. Solid lines show the Fourier transforms of the sampled profiles used to fit. Dotted lines show extrapolated profiles for (M,z) values that were not used for fitting. These extrapolated profiles follow the same general peak- and radius-patterns of the sampled profiles used to fit. As in Figure 5.4, no error bars are shown because these predicted profiles were generated using Equation (5.9).

either using the NUMPY.OPTIMIZE.MINIMIZE method of the main fitting procedure, or it can be simpler to find the feature values by checking a regularly spaced grid of likely values. For instance, the profiles in Figure 5.4 have peak-above-infinity feature values roughly in the range $[0.4, 1.0]$ and radius feature values (i.e. the distance at which the profile falls to $1/e \approx 0.37$ of peak value) roughly in the range $[0, 100]$. Using an exhaustive grid of such possible pairs in the given ranges, the best feature values (ρ_0, R) are found for each (M,z) pair. Each profile has been then fitted separately to Equation (5.5). To estimate the best hyperparameters, these separately fitted feature *values* (ρ_0, R) are fitted to the physical feature *relations* $(\rho_0(M,z), R(M,z))$. After determining the separate best peak-above-infinity feature values ρ_0, the feature relation

$$\rho_0(M,z) = \Theta_0 + \Theta_1 z + \Theta_2 \log(M) \tag{5.13}$$

can be fitted by finding the best values of $\{\Theta_0, \Theta_1, \Theta_2\}$. These best values can

Name	Real-space, $\rho_x(r)$	Fourier-space, $\widetilde{\rho}_x(k)$
Step	$\rho_0 * [r \leq R]$	$\rho_0\ 4\pi \left(\frac{\sin(kR) - kR\cos(kR)}{(kR)^3} \right)$
Exponential-r	$\rho_0\ \exp(-r/R)$	$\rho_0 (8\pi)^{-1} \left(1 + [kR]^2\right)^{-2}$
Exponential-r^2	$\rho_0\ \exp\left(-\left[\frac{r}{2R}\right]^{\ 2}\right)$	$\rho_0 (2\pi)^{-3/2} \exp\left(-[kR/2]^2\right)$

Table 5.2: All profile functions used for fitting in this chapter, with their corresponding three-dimensional Fourier-transforms (Hankel transform) for $k = 2\pi/r$. The features for each profile are the peak-above-infinity height ρ_0 and the scale radius R. By far the most-used profile in this chapter is the step function.

then be used as initial guesses for the higher-dimensional space of jointly-fitted hyperparameters.

5.1.5 Profile functions and feature relations

Throughout this chapter several different profile functions and feature relations are used. Table 5.2 shows the main profile functions for which fitting was attempted in this chapter, although many other real-space functions with analytic Hankel transform exist (Weisstein, 2019). These fitting functions were originally chosen as a simple first attempt to model the ionisation fraction profiles around dark matter halos. As discussed in Section 5.4, the final model in this chapter appears to be more strongly affected by the overlap of ionising bubbles than by the specific fitting function used. As such, I leave considerations of more complex fitting functions as an avenue of potential future work, if the effects of overlap were to be correctly modelled. Table 5.3 shows the feature relations used to connect the features (ρ_0, R) to the halo properties (M, z).

5.2 Ionisation fraction: toy model

This section describes a simple toy model for the clustering of ionisation fraction data. Starting with the simplest possible model, the measured power spectra of mock data cubes are compared to the predicted power spectra from a halo model. The first mock data cubes contain randomly-placed spherical bubbles of a single fixed radius. For these data the measured and halo model power spectra match

Name	Feature relation, $\phi(M,z)$
Fixed Value	1.0
Global Mean Value	$1-\langle x_{\rm HII}\rangle(z)$
Power Law	$\Theta_0\ z^{\Theta_1}\ [log(M)]^{\Theta_2}$
Polynomial	$\sum_{\alpha\beta}\Theta_{\alpha\beta}\ z^{\alpha}\ [log(M)]^{\beta}$
Radius-style	$\Theta_0[M]^{\Theta_1}$

Table 5.3: All feature relations used for fitting in this chapter. The values Θ_i are hyperparameters that can be fitted to profiles that are measured as a function of the halo properties (M,z).

closely. Multiple distinct sphere radii are added to these mock data, again giving a close match between measured and halo model spectra. This section ends with mock data cubes where many distinct sphere radii are used, with number densities drawn from a power-law relationship. Again the halo model power spectra for these data match the measured power spectra within the scatter of sample variance.

5.2.1 Single-radius model

Mock ionisation fraction data cubes are generated by using spherical bubbles of radius R_0 and fixed number density n_0. These ionised bubbles show no clustering and no redshift dependence. In these data the theoretical power spectrum $P^{\rm bb}(k;R_1,R_2)$ between the centres of two bubbles with radii R_1 and R_2 is zero for all $k\neq 0$. Thus the two-halo term $P^{\rm 2h}(k)=0$ throughout this section and only the 1-halo term contributes to the total power spectrum. The number density function for these data is given by

$$\frac{dn}{dR}(R)=n_0\ \delta_D(R-R_0). \tag{5.14}$$

The mock data are generated as follows. First a catalog of bubble centre locations is formed by randomly sampling locations uniformly within the box. The number of bubbles is chosen to match a fixed chosen value for n_0. The mock data

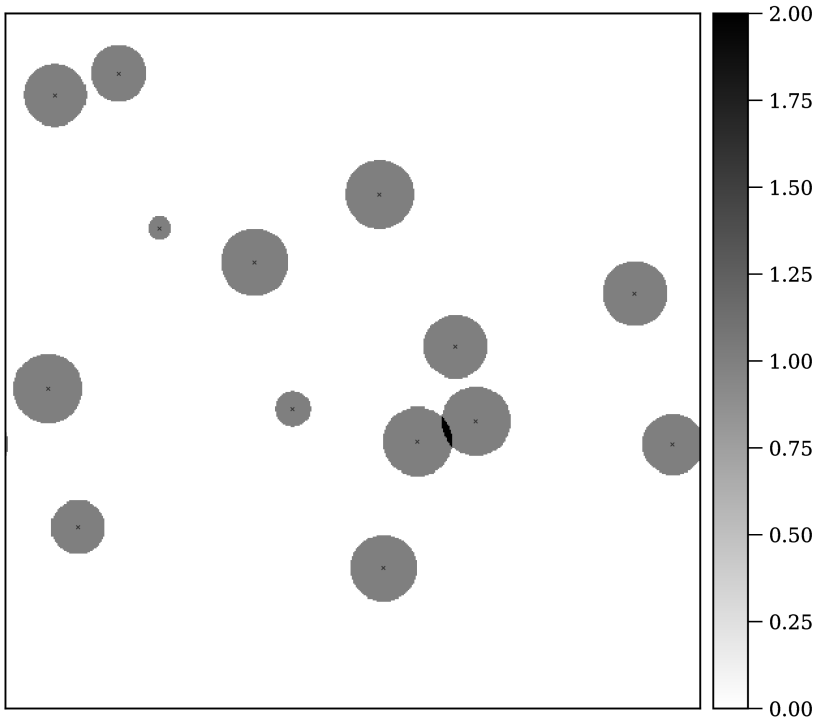

Figure 5.7: Slice though realisation for single-radius model with fixed sphere radii $R = 5$ and total box size $L = 100.0$. The number density of spheres is fixed at $n_0 = 1.9 \times 10^{-4}$, chosen so that the total volume of spheres is 10% of the box volume. Note that spheres may overlap in this model.

cube is created as a three-dimensional set of pixels, initialised to zeros. All pixels within the fixed radius R_0 of any bubble are then set to 100% ionised. In this subsection, pixels which lie within R_0 of multiple bubble centres are allowed to overlap. The issue of overlapping bubbles is a recurrent problem with ionisation models (see for instance Section 3.2 of Furlanetto et al. 2004). Allowing overlapping bubbles can lead to non-physical results where the data field exceeds 100% ionised. Provided the total volume of spheres is much less than the total box volume, however, the effect of this relaxed condition on the final correlation statistics is minimal. I first allow the overlap of bubbles in my models, and consider methods of handling the overlap are considered later in section Section 5.3.2. Figure 5.7 shows an example resulting mock data cube for this model. The data cube is stored, along with the bubble centres catalog that can then be used in the stacking procedure.

The stack-fit-predict procedure is used to fit the measured profiles from these

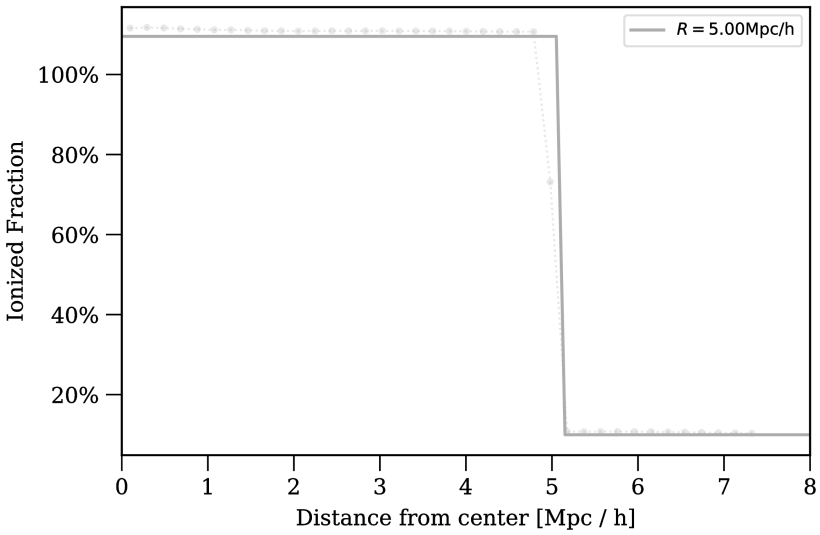

Figure 5.8: Fitted profile for data cube with spheres of a single radius $R_0 = 5$ and number density $n_0 = 1.9 \times 10^{-4}$, chosen so that the total volume of spheres is 10% of the box volume. The fitting procedure has correctly identified the radius of the spheres in this realisation.

mock data cubes. Spherically-averaged profiles are measured from the mock data cubes by using the stored catalog of bubbles centres. The stacked profiles are fitted using profile function

$$\rho_x(r;M) = \begin{cases} \rho_0(M,z), & \text{if } r \leq R(M,z) \\ 0, & \text{otherwise,} \end{cases} \tag{5.15}$$

and feature relations

$$\rho_0(M,z) = \Theta_0 \tag{5.16}$$

$$R(M,z) = \Theta_1. \tag{5.17}$$

The true values for these hyperparameters are $\Theta_0 = 1.0$ and $\Theta_1 = R_0$. Figure 5.8 shows the resulting fitted and measured profile. The dotted line with circular markers shows the measured profile and the solid line shows the fitted profile. The fitted profile has correctly identified the radius and peak of the spheres.

I use Table 5.2 to determine the analytic Fourier transform of the step-function

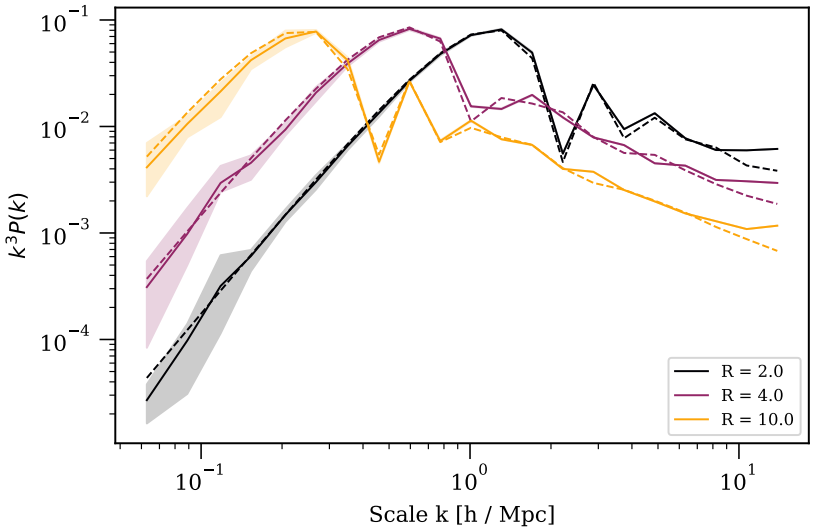

Figure 5.9: Measured and predicted power spectra for three single-radius models. The models have $R_0 = 2$ (black), $R_0 = 4$ (purple), and $R_0 = 10$ (orange). In each case the solid line shows the mean measured power spectrum of five realisations. The shaded regions show the spread of measured power spectra. The dashed lines show the halo model predictions. The realisations for these models are similar to those shown in Figure 5.7. The predictions match the measured spread except at very small scales (high k), as discussed in the text.

profile in Equation (5.15). The predicted one-halo term is then

$$P^{1h}(k) = n_0\,\rho_0 \left(\frac{4\pi\left[\sin(kR(M,z)) - kR(M,z)\cos(kR(M,z))\right]}{k^3} \right)^2 . \qquad (5.18)$$

Figure 5.9 shows the halo model power spectra using the stack-fit-predict method as dashed lines for three different scenarios. I compare these predicted spectra to the actual power spectra of the mock data cubes, as calculated using five mock data realisations. For each scenario in Figure 5.9, the shaded regions show the standard deviation of the measured spectra of the five realisations, and the solid line shows the spectrum averaged across the five realisations. For all three scenarios the dashed line from my model follows the measured spectrum closely, except at very small scales (high k). The halo model can be used to model the clustering of single-radius models.

Figure 5.10 shows the effect of different binning widths on the power spectra.

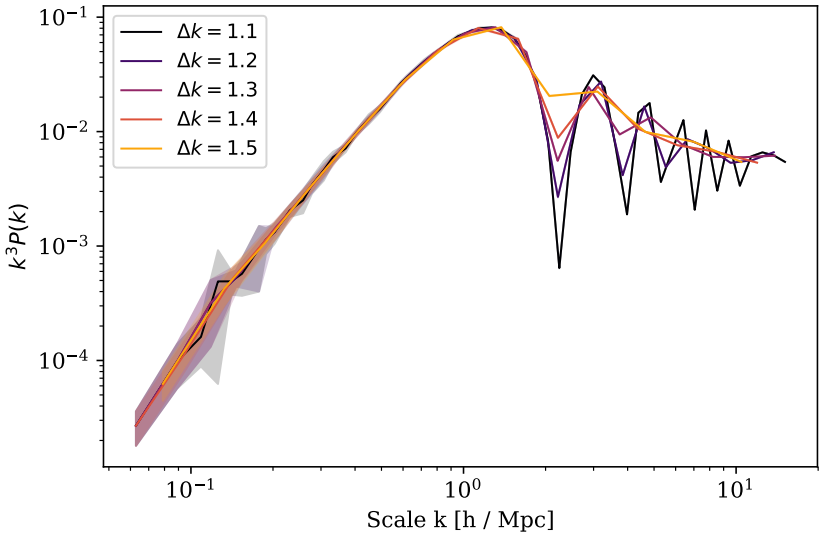

Figure 5.10: Power spectra for different binning widths in one-radius models with $R_0 = 2$. Higher values of Δk lead to smoother curves with less prominent features, in particular the fast-varying sinusoidal regions past the first power spectrum peak.

At high and low k-scales, the fast-changing power spectrum is highly sensitive to binning: using a wider binning range (higher Δk) results in a smoother curve with less-prominent sinusoidal features.

5.2.2 Two-radius model

The simple model of all ionised bubbles having the same radius is not a realistic one. In reality different bubbles have grown at different speeds and for different lengths of time, giving a distribution of spheres with different radii. I test my halo model by generating mock data cube realisations for spheres with two distinct radii R_0 and R_1, and fixed number densities n_0 and n_1 respectively. I use my stack-fit-predict method and compare the resulting spectra to the actual spectra calculated on the mock data cubes. The overall number density function is given by

$$\frac{dn}{dR}(R) = n_0 \delta_D(R - R_0) + n_1 \delta_D(R - R_1), \tag{5.19}$$

Figure 5.11 shows a slice through an example realisation for $R_0 = 5$ and $R_1 = 10$. Both small and large spheres can be seen in this figure. The number densities are chosen so that the total volume of the smaller spheres is equal to the total volume

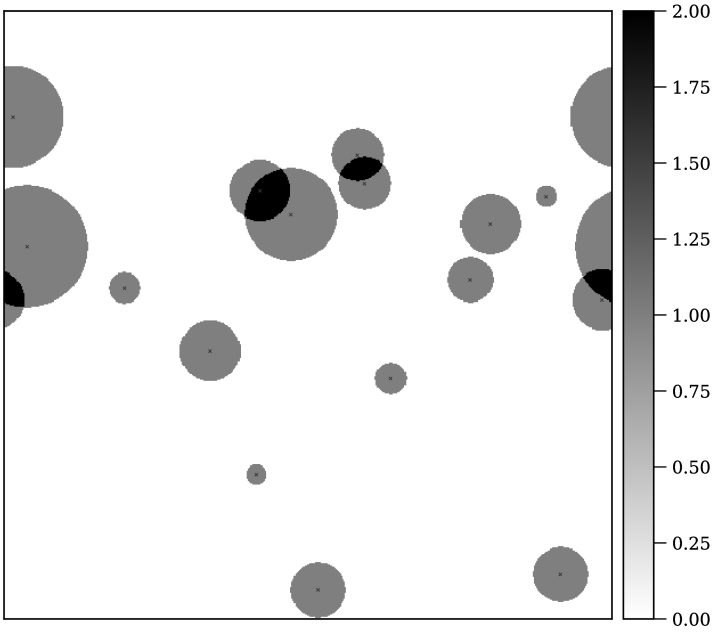

Figure 5.11: Slice through realisation for two-radius model. The two distinct sphere radii are $R_0 = 5$ and $R_1 = 10$, with matched number densities $n_0 = 1.7 \times 10^{-4}$ and $n_1 = 2.1 \times 10^{-5}$ chosen so that the total volume of spheres is around 10% of the box volume.

of larger spheres: there are more smaller spheres and fewer larger spheres. I also require that the total summed volume of all spheres should be roughly 10% of the box volume.

Feature relations

$$\rho_0(M,z) = \Theta_0 \tag{5.20}$$

$$R(M,z) = \Theta_1 [M]^{\Theta_2}, \tag{5.21}$$

are used to capture the fact that the two different sphere radii are related to the mass with a power law. The correct values for these hyperparameters are $\Theta_0 = 1$, $\Theta_1 = \frac{3}{4\pi}$, and $\Theta_2 = \frac{1}{3}$. Figure 5.12 shows the predicted and measured power spectra from scenarios with various R_0 and R_1 values. The measured spectra are again from five realisations, and the halo model predictions lie close to the measured spectra. The halo model can accurately predict the power spectra of two-radius models.

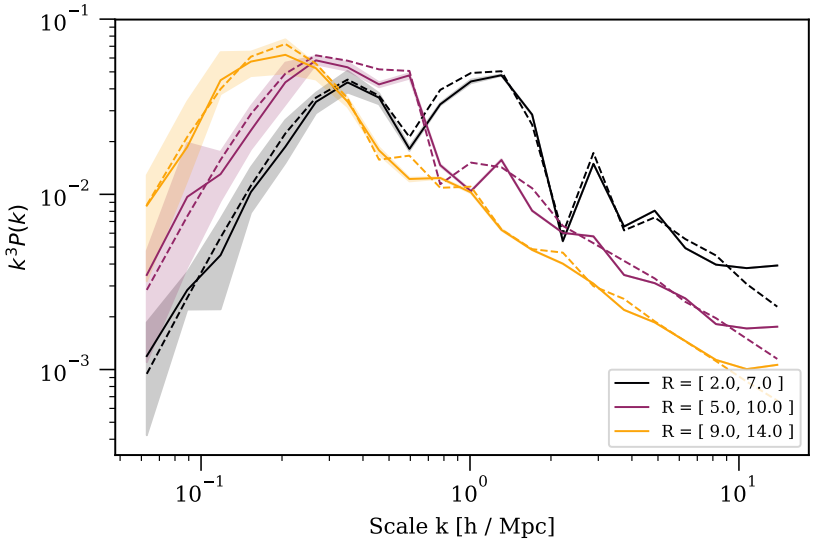

Figure 5.12: Measured and predicted power spectra for two-radius models. The two distinct sphere radii are given in the legend for each scenario. The number densities in each case were chosen so that the total volume of the smaller spheres is equal to the total volume of the larger spheres, with the summed volumes around 10% of the box volume. The predicted power spectra (dashed lines) lie close to the measured spread of spectra (shaded regions) for each scenario, except at very small scales (high k).

Slight discrepancies can be seen at middling scales around $k = 1.0\text{h/Mpc}$. These discrepancies are almost certainly caused by the binning of the power spectrum measurements. The fast-changing power spectrum at these scales is highly sensitive to the k-ranges over which the scales are binned, as discussed in the previous section.

5.2.3 Number density function model

Usually in cosmology the number density of dark matter halos is assumed to follow a specific distribution known as the halo mass function. This motivates a further realistic extension to the model of the previous subsections: setting the abundances of differently-sized spheres to follow a specific number density function. I choose the power law

$$\frac{dn}{dR}(R) = n_0 R^{-\beta}, \tag{5.22}$$

to test my halo model on data where the spheres are drawn from a number density

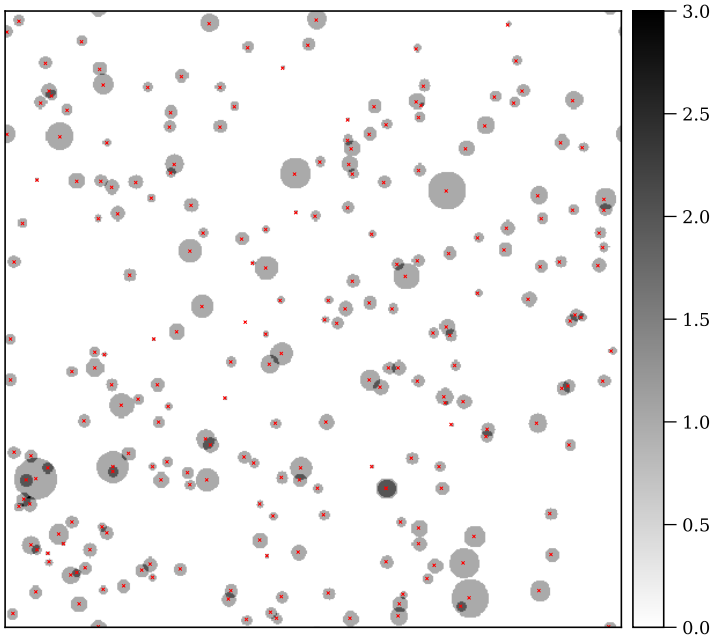

Figure 5.13: Slice through realisation for power-law number density model. The number density for each sphere radius is given by a power law $\frac{dn(R)}{dR} = n_0 R^{-6}$, with n_0 chosen so that the total sphere volumes is around 10% of the box volume.

function. Figure 5.13 shows a slice through the resulting realisation for a power-law model with $\beta = 6$.

Using the chain rule, the number density of sphere masses can be given in terms of the sphere radius,

$$\frac{dn}{dM}(M) = \frac{dn}{dR}\frac{dR}{dM} = \frac{n_0}{4\pi} R^{-\beta-2}. \tag{5.23}$$

I use the stack-fit-predict procedure for the resulting generated mock data cubes. The same feature relations are used for the fitting procedure as in Equation (5.20). Figure 5.14 shows some of the fitted and measured profiles for one realisation of this scenario using $\frac{dn(R)}{dR} = n_0 R^{-6}$ for the number densities. The measured stack values are shown as dotted lines with circular markers, and the solid lines indicate the resulting predictions from the jointly-fitted profiles. The fitted profiles have correctly identified the peak and radius features of all spheres. The stack for $R = 3.05$ Mpc/h is somewhat anomalous, exceeding 100% ionised fraction while the other stacks do not. This is because there were very few spheres with $R = 3.05$ Mpc/h and, by

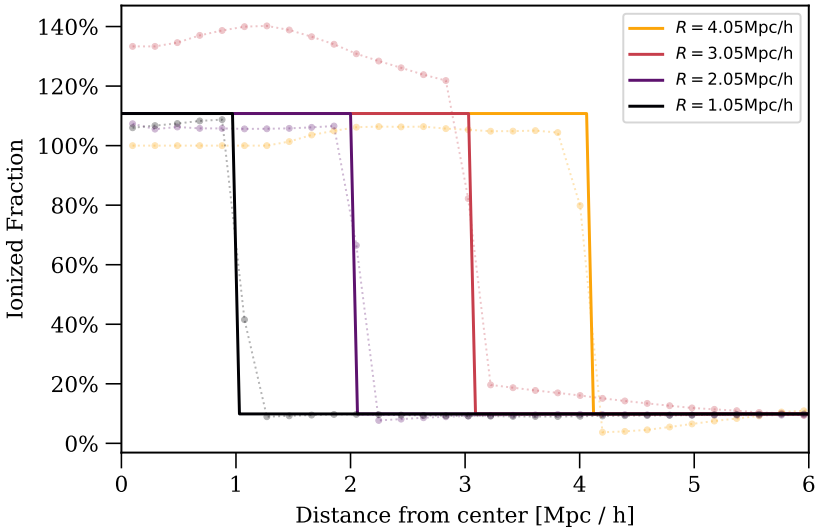

Figure 5.14: Measured and fitted profiles around bubble centres for a power-law number density model with $\frac{dn(R)}{dR} = n_0 R^{-6}$. In each case the dotted lines with circular markers show the measured profile values and the solid lines show the fitted results. The true radii for these profiles were $R \approx 1, 2, 3, 4$ which match nearly perfectly with the fitted profiles in this figure. The stack for $R = 3.05$ Mpc/h is somewhat anomalous as discussed in the text, but this does not affect the quality of the overall fitted profiles. This figure shows only a few of the many sphere radii in the realisation.

chance, these few spheres happened to overlap with several other nearby spheres. Using a single peak-above-infinity value in Equation (5.20) means that the joint fitting procedure ignores the overlapping stack and successfully gives the true underlying profile radius and peak values.

Figure 5.15 then shows the halo model and measured power spectra for models using four different power-law distributions. The halo model continues to show a good match between the measured power spectra and the halo model spectra. The halo model is able to capture the complexity of data where the spheres are drawn from a number density function. In the following section, the model is extended further to include the effect of clustering.

5.3 Ionisation fraction: including clustering

In this section I introduce clustering to the halo model. The simplest way to generate a catalog of centres with realistic clustering is to use the actual halo catalog outputs

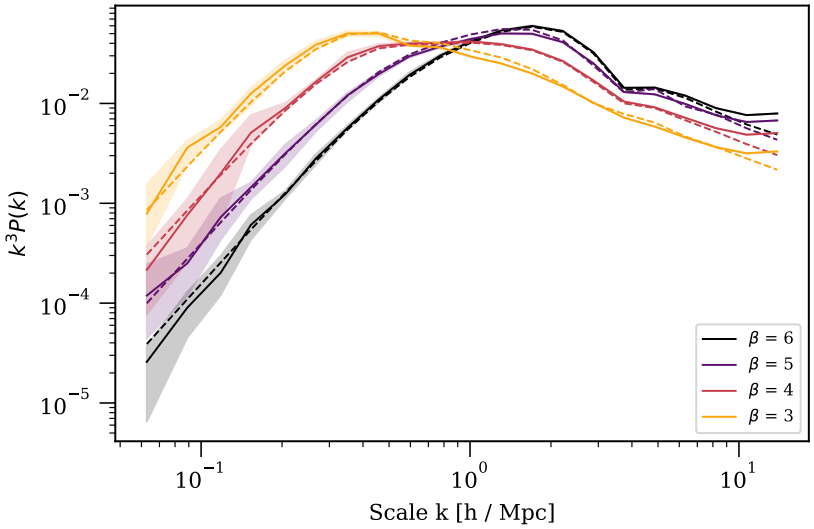

Figure 5.15: Measured and predicted power spectra for models with many distinct radii, where the number density for each sphere radius is drawn from power laws $\frac{dn(R)}{dR} = n_0 R^{-\beta}$. The predicted spectra lie within the measured spread, except at very small scales (high k).

from SIMFAST21 itself. The mock data are generated by using spheres whose radius is a fixed multiple of the underlying halo mass. The non-zero clustering between sources means that the total power spectrum $P(k) = P^{1h}(k) + P^{2h}(k)$ has terms

$$P^{1h}(k) = \int dM \, \frac{dn}{dM}(M) \, |\widetilde{\rho}_x(k;M)|^2 \tag{5.24}$$

and

$$P^{2h}(k) = P^{lin}(k) \left(\int dM \, \frac{dn}{dM}(M) b(M) \widetilde{\rho}_x(k;M) \right)^2, \tag{5.25}$$

where the bias term from Seljak (2000) has been used.

5.3.1 Allowing overlap

Including clustering in the mock data immediately causes widespread overlap of the spheres. I use the halo catalogs from a canonical SIMFAST21 simulation with parameters $M_{min} = 5 \times 10^8 M_\odot$, $\zeta_{ion} = 30.0$, and $R_{max} = 10$ Mpc. The mock data are generated by placing an ionised sphere onto every halo location. The radius of each sphere depends only on the underlying halo mass, with fixed ratio

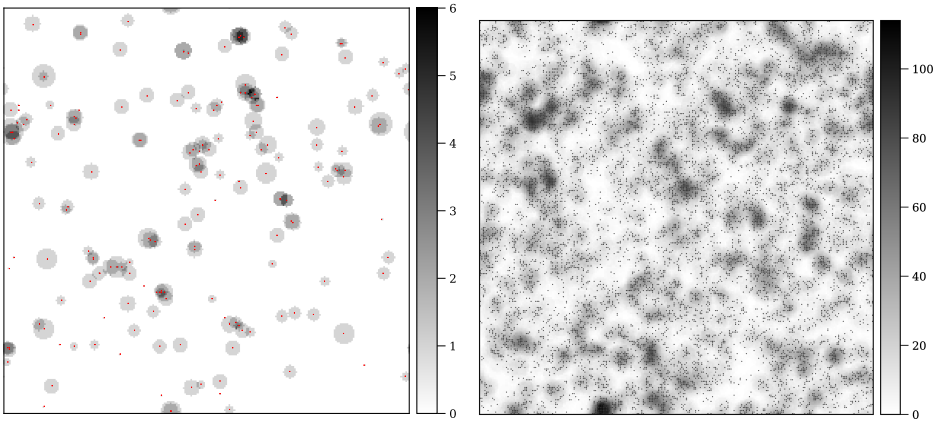

Figure 5.16: Slice through mock data realisations for the clustering ionisation fraction model, in which spheres are allowed to overlap and give values $x_{\rm HII} > 1.0$. The left panel uses a SIMFAST21 halo catalog at $z = 15$, and the right panel uses a catalog from $z = 11$. In both cases the locations of halos are indicated by red markers.

$R_{\rm ion}(M_{\rm halo}, z) = 17.5 R_{\rm halo}$ corresponding to a value of $\zeta_{\rm ion} = 30$ in the FZH model from Section 2.1.1:

$$\begin{aligned} M_{\rm ion} &= \zeta_{\rm ion} M_{\rm halo}, \\ \rho_{\rm ion} R_{\rm ion}^3 &= \zeta_{\rm ion} \rho_{\rm halo} R_{\rm halo}^3, \\ R_{\rm ion} &= \sqrt[3]{\zeta_{\rm ion} \frac{\rho_{\rm halo}}{\rho_{\rm ion}}} R_{\rm halo}, \\ R_{\rm ion} &= \sqrt[3]{\zeta_{\rm ion} \Delta_{\rm vir}} R_{\rm halo}, \end{aligned} \tag{5.26}$$

with $\Delta_{\rm vir} \approx 18\pi^2$ as described in Section 2.3.1. Figure 5.16 shows slices through two such realisations using halo catalogs at $z = 11$ and $z = 15$ from SIMFAST21. Note that these data fields are nonphysical because they are allowed to exceed 100% ionised. The number density function used in the halo model equations for this section is the true halo mass function calculated using the HMFCALC module (Murray et al., 2013) as described in Section 2.3.6. The halo bias term for each mass is also calculated using the HMFCALC module.

The fitting feature relations for this section are the same as in the previous

sections, namely

$$\rho_0(M,z) = \Theta_0 \tag{5.27}$$

$$R(M,z) = \Theta_1[M]^{\Theta_2}, \tag{5.28}$$

The resulting halo model power spectra are then shown in Figure 5.17 for halo catalogs of several different redshifts. The halo model and measured power spectra match quite closely at earlier redshifts, where the uncapped data mean $\langle \text{data} \rangle$ and capped data mean $\langle x_{\text{HII}} \rangle$ are both low. As reionization progresses and larger bubbles start to overlap $\langle \text{data} \rangle$ begins to diverge significantly from $\langle x_{\text{HII}} \rangle$. This gives rise to an increasing offset between the halo model and measured spectra, in particular on middling scales. The halo model is somewhat able to capture the clustering of these data but, in any case, the data are nonphysical because the ionisation fraction field has been allowed to exceed 100% ionised.

5.3.2 Removing overlap

Handling the overlap of bubbles in the mock data is necessary for a realistic model for the ionisation fraction. To determine the extent that overlap affects the power spectra measurements and halo model predictions, the generated data cubes in Figure 5.16 were clipped so that any pixel with $x_{\text{HII}} > 1.0$ is set to $x_{\text{HII}} = 1.0$. Figure 5.18 shows slices through the resulting clipped data.

The same stack-fit-predict procedure is used on these new clipped data cubes. Figure 5.19 shows the halo model and measured power spectra predictions for each redshift in a separate panel. The measured power spectra at earlier redshifts are similar to the measured spectra for the un-clipped data. As expected the halo model spectra for early times matches the measured ones. The effect of overlap can immediately be seen even for values of $x_{\text{HII}} \geq 0.10$: most prominently, there is a significant amplitude offset between the predicted and measured power spectra. Although I have handled the effect of overlap in the mock data, the analytic halo model in Equations 5.24 and 5.25 makes no account of the overlap between ionised bubbles

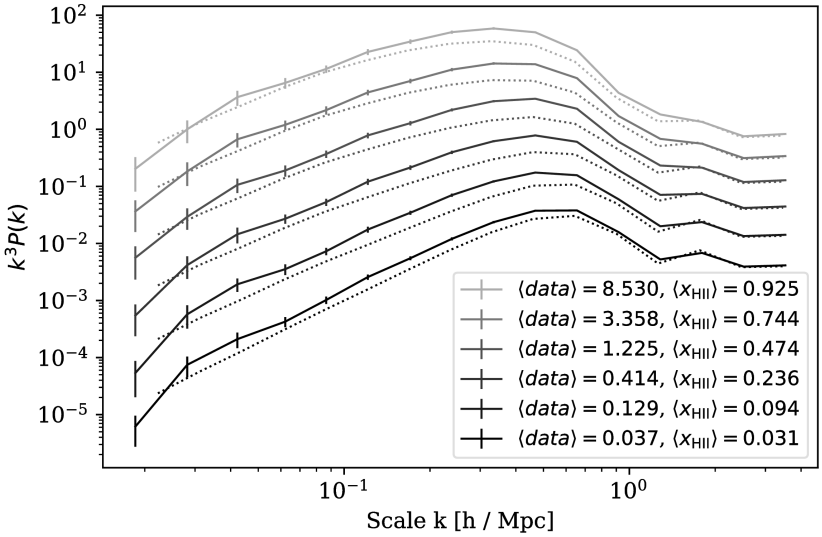

Figure 5.17: Measured and predicted power spectra for clustered ionisation fraction models allowing overlap. Dashed lines show the predicted spectra from the halo model, solid lines show the mean measured spectra from five realisations with error bars to indicate the two-sigma spread in measured spectra. The shapes of the predicted spectra qualitatively match the measured spectra, although already the predictions do not lie perfectly within the two-sigma range of the measured spectra. For later redshifts with higher global ionisation fraction, there is a greater offset in the amplitude likely caused by widespread overlap.

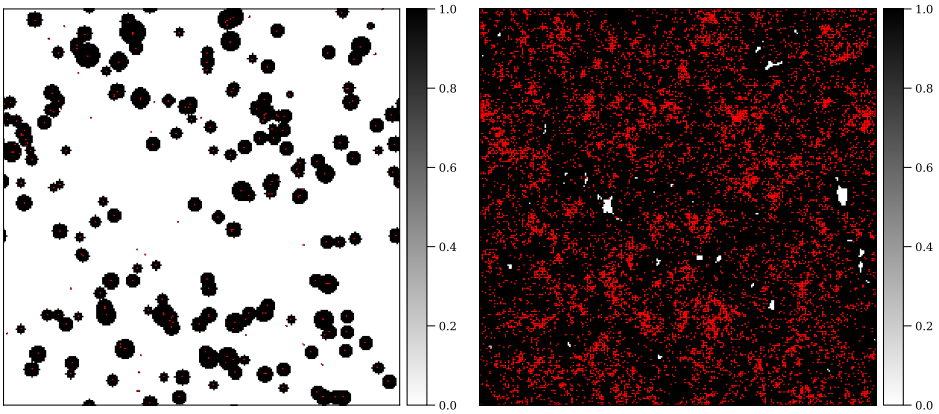

Figure 5.18: Slice through realisation for halo model with clustering, clipping the data so that each pixel contains only legal values for $0 \leq x_{\text{HII}} \leq 1.0$

around halos.

5.3.3 Handling overlap: suppression

In order to handle the effect of overlap in the analytic model, the number density of dark matter halos must be matched to the number density of actual ionising sources. In the formalism of this chapter so far, ionised bubbles have been fitted around halos locations and not ionising sources. I introduce a suppression function $f^{\mathrm{b}}(M;z)$ which links the number density function of bubbles to the number density function of halos. The function $f^{\mathrm{b}}(M;z)$ specifies the extent of bubble overlap for halos with different masses M at different redshifts z. The values $1/f^{\mathrm{b}}$ thus give the fraction of halos at any (M,z) that give rise to an ionised bubble. This function can be sampled from the un-clipped realisations in the previous subsection. I measure the extent of overlap for every halo with their respective (M,z) and fit these sampled values to a linear regression model using the SKLEARN.LINEAR_MODEL.LINEARREGRESSION class. Figure 5.20 shows the resulting suppression function, with the sampled values plotted as points and the fitted linear regression model as solid lines.

The effect of overlap is included in the halo model by suppressing the number densities of each halo mass in the catalog by the suppression function,

$$\frac{dn}{dM}(M,z) \rightarrow \frac{\frac{dn}{dM}(M,z)}{f^{\mathrm{b}}(M,z)}. \tag{5.29}$$

The resulting power spectra predictions including this suppression function are shown in Figure 5.21. The halo model using suppressed predictions is somewhat better able to account for overlap than those in Figure 5.19: the amplitude of the predicted and measured power spectra match more closely up until around 80% of the way through the reionization process ($x_{\mathrm{HII}} = 0.8$). At later times the amplitude still diverges, but to a lesser extent than without suppression. Although the predicted spectra amplitudes are much closer to the measured spectra, their shapes still differ across most redshifts. The predicted spectra are less smooth than the measured spectra, showing a more distinct peak at middling scaled around $k = 0.8\mathrm{h/Mpc}$ (corresponding to real-space scales of 8 Mpc/h).

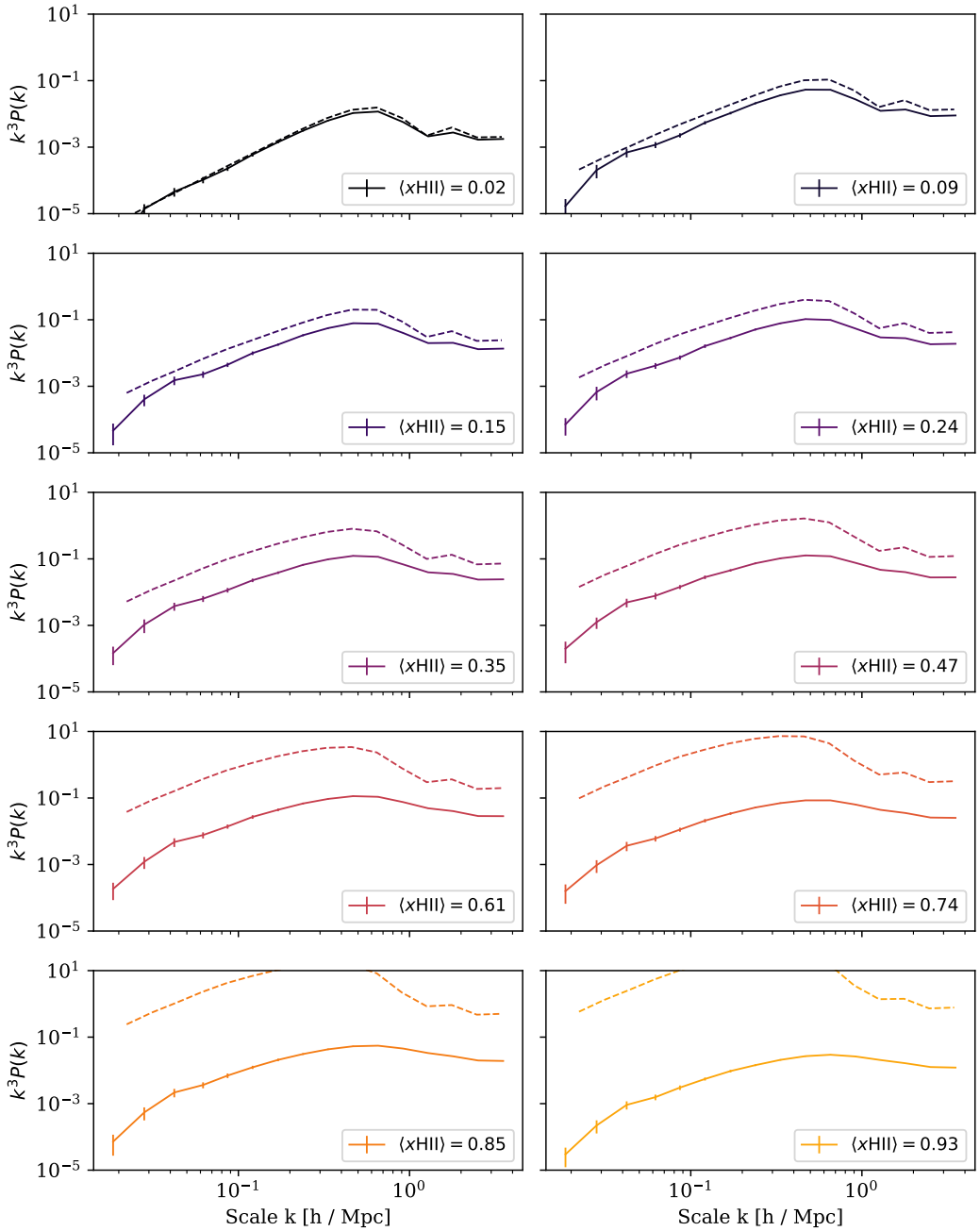

Figure 5.19: Measured and predicted power spectra of clipped data. Each redshift is shown in a separate panel. Dashed lines show the predicted spectra from the halo model, solid lines show the mean measured spectra from five realisations with error bars to indicate the two-sigma spread in measured spectra. At early redshifts, there is minimal overlap in the bubbles and the model works well. At later redshifts, the clipping has a significant effect on the model predicted power spectrum and the two diverge. Without accounting for overlap in the analytic model, the spectra amplitudes quickly diverge.

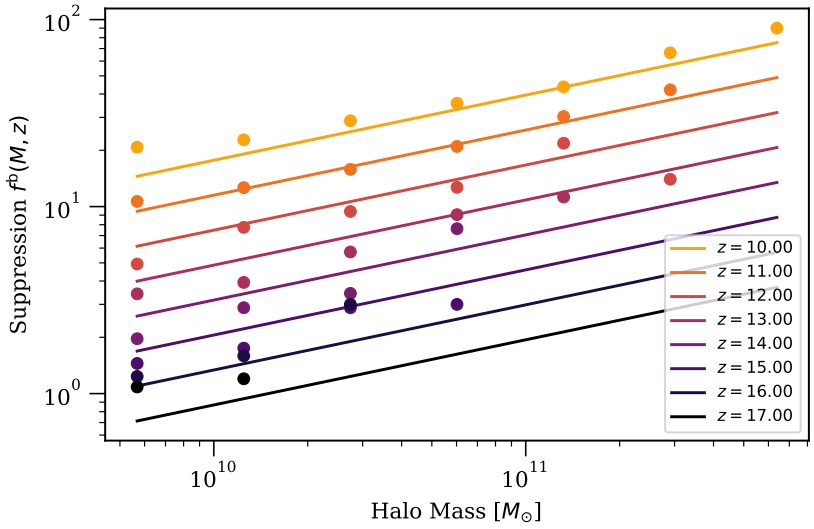

Figure 5.20: Suppression function calculated by finding the overlap in the un-clipped data at halos centres for every (M,z) pair. This function can be included in the halo model to suppress the halo mass abundances into effective ionised bubble abundances.

5.4 Ionisation fraction: towards a full model

The halo model in the previous section makes approximate predictions for the power spectrum amplitude and shape of a simplified ionisation fraction model. In the simple model every halo gives rise to a spherical fully-ionised region whose radius depends only on the underlying halo mass. The original motivation for this chapter is to replace the power spectrum predictions from SIMFAST21 by using a fitted analytic model. In this section I take the final step towards this goal: fitting a halo model to SIMFAST21 $x_{\rm HII}$ data directly. The resulting fitted model makes very approximate predictions for the amplitude of the SIMFAST21 power spectra. I discuss the possible reasons for the mismatch and consider potential avenues of future work which might resolve the differences. I use the same stack-fit-predict method with step-function profiles. For the feature relations I adjust the peak-above-infinity feature to be the correct one for ionisation fraction profiles, namely

$$\rho_0(M,z) = 1 - \langle x_{\rm HII} \rangle (z) \tag{5.30}$$

with the radius feature remaining the same as in previous sections

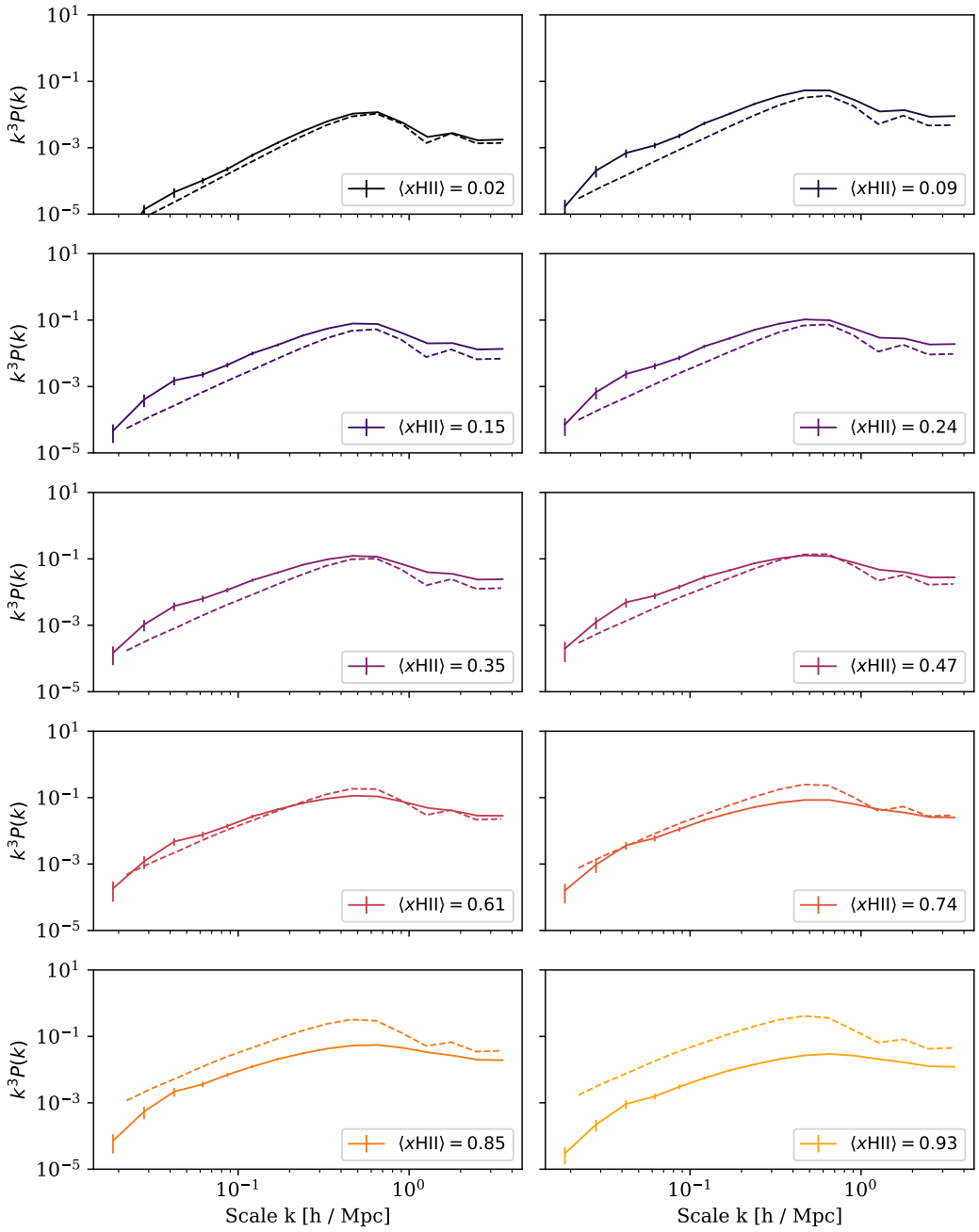

Figure 5.21: Measured and predicted power spectra using the suppression function to reduce the abundances of ionised bubbles sourced at each halo. Dashed lines show the predicted spectra from the halo model, solid lines show the mean measured spectra from five realisations with error bars to indicate the two-sigma spread in measured spectra. The predictions match the measured spectra better than the clipped model up until $x_{\rm HII} = 0.75$. The later redshifts still show significant departure between the predicted and measured spectra due to the widespread overlap of ionised bubbles. Most predictions still lie outside the two-sigma range of the measured spectra, indicating that model is still not accurately replicating the simulated spectra.

$$R(M,z) = \Theta_0[M]^{\Theta_1}. \tag{5.31}$$

Figure 5.22 shows the measured and fitted profiles for a single SIMFAST21 realisation. The measured stacks are shown as dotted lines with circular markers, and the solid lines indicate the jointly-fitted profiles. The profiles are shown at $z = 13$ for all unique halo masses that were resolved in the simulation. The jointly-fitted radii roughly match the halfway point of the measured profiles. Note that SIMFAST21 accounts for unresolved halos which are too small to be included in the halo catalogs used for fitting. One way to include the effect of these small halos would be to use higher-resolution simulations, so that all valid halo masses above the minimum mass parameter $M_{\rm min}$ are resolved. Such high resolution simulations would take a prohibitively long time to run. I instead account for these smaller halos by using $M_{\rm min}$ as the lower limit of the 1-halo and 2-halo integrals and allowing my fitted profile function to extrapolate the likely profiles of these bubbles. In practice the small mass halos made little difference to the resulting power spectrum predictions of this section.

The resulting final predicted spectra in Figure 5.23 match only loosely with the actual SIMFAST21 spectra. For early redshifts with $\langle x_{\rm HII}\rangle < 0.05$ the power spectrum shapes differ; for later redshifts with $\langle x_{\rm HII}\rangle \geq 0.05$ the shapes match more closely but the halo model quickly over-predicts the power spectrum amplitudes. Only spectra for $\langle x_{\rm HII}\rangle < 0.10$ are shown, since all power spectra for higher $\langle x_{\rm HII}\rangle$ values continue the trend of growing discrepancy between the measured and predicted power spectra. The mismatch must be caused by one or more incorrect fittings of the halo model parts: the profile function, the clustering function, the bias function, or the number density distribution. I consider each part separately to determine whether it likely contributes to the poorer predictions.

The fitted profiles in Figure 5.22 have correctly identified a reasonable profile radius for each mass in the catalog. Although the shapes of the fitted profiles differ from the measured stacks, this would be unlikely to cause such a large discrepancy in the power spectrum amplitudes. The profile function is unlikely to be the main

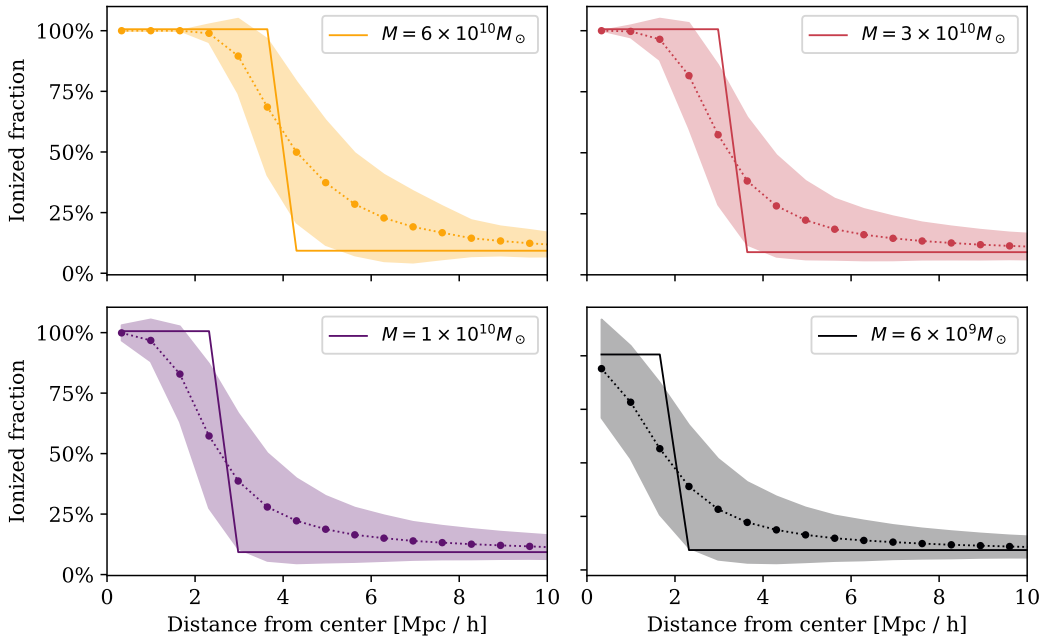

Figure 5.22: Measured and fitted profiles around halo centres for SIMFAST21 realisation using the canonical reionization parameters: $M_{\rm min} = 5 \times 10^8 M_\odot$, $\zeta_{\rm ion} = 30.0$, and $R_{\rm max} = 10$ Mpc. Each panel shows the stacked and fitted profiles for a different halo masses at a fixed redshift of $z = 13.0$. The solid line shows the fitted profile. The dotted line shows the mean of the measured stacks, with two-sigma errors indicated by the shaded region. The fitted profiles (solid lines) pick out a reasonable radius for each of the measured stacked profiles (dotted lines with circular markers) for each of the masses.

source of error. The clustering and bias functions were used in the previous section for recreated ionisation fraction data including the effects of clustering and biasing. The successful match between the measured and predicted spectra of those recreated data suggest that the clustering and bias functions are not a contributing factor to the poorer predictions. Almost certainly the cause of the mismatch is in the difference between the number density distributions of bubbles and the halo mass function. In the previous section a suppression function was successfully used to convert the halo mass function to a bubble distribution function, for the simple model of a single spherical ionised region sourced at every halo. The suppression function for SIMFAST21 clearly differs significantly from the suppression function of that simple model, causing my halo model to over-predict the abundances of ionised regions. The same logic applies for different power spectrum shapes at earlier redshifts. The

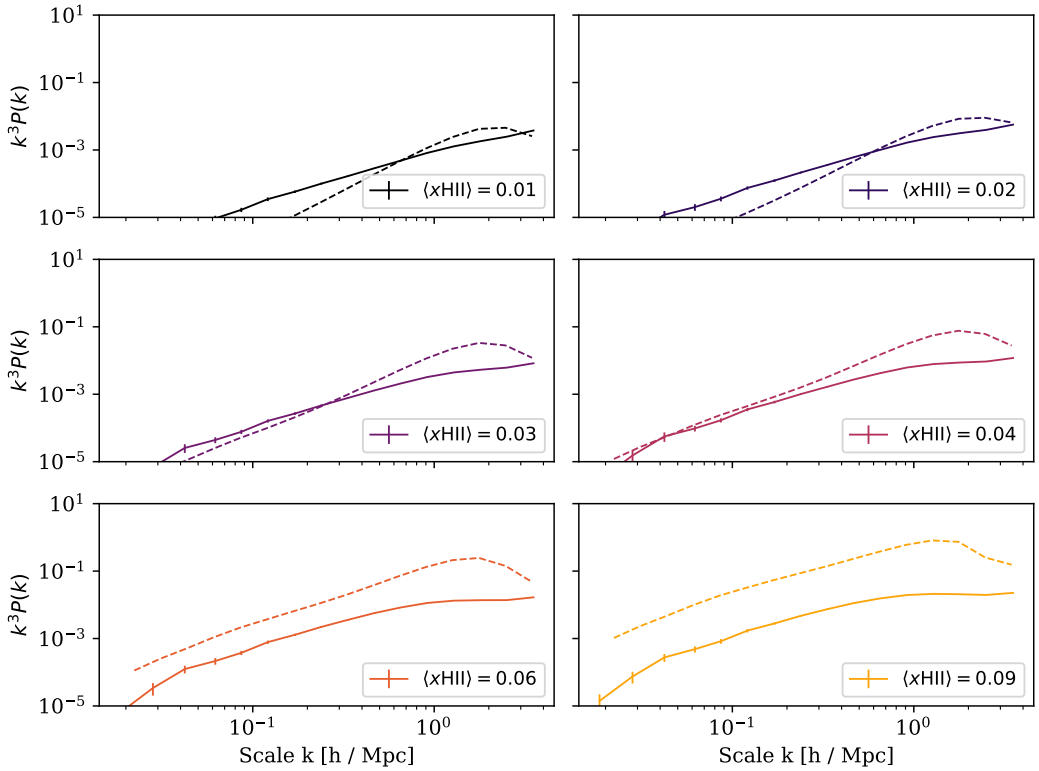

Figure 5.23: Measured and predicted power spectra predictions combining all previous properties: clustering, bias, suppression function, and fitted profiles from actual SIMFAST21 data. The measured spectra are calculated directly from the three-dimensional SIMFAST21 ionisation fraction cubes. The halo model predictions quickly diverge from the measured spectra, likely due to the suppression function as discussed in the text. Only spectra for $\langle x_{\rm HII} \rangle < 0.10$ are shown, since all power spectra for higher $\langle x_{\rm HII} \rangle$ values continue the trend of growing discrepancy between the measured and predicted power spectra.

shape of the power spectrum is dependent on the relative abundances of differently-sized bubbles: smaller bubbles have power spectrum peaks at larger k-values, and larger bubbles peak at smaller k-values. An incorrect suppression function would skew the relative abundances and give a different power spectrum shape, as is observed in the top four panels of Figure 5.23.

The mismatch would likely be resolved by using the number density distribution of bubbles directly, instead of using the halo mass function with a suppression function. This would require measuring the abundances of differently-sized ionised regions in the simulation, and then fitting the resulting number distribution to an analytic form for the halo model. Although this would certainly give a better match

between the halo model power spectra and those from SIMFAST21, this would effectively be a test of whether the simulation is correctly implementing the FZH procedure in Section 2.1.1. However, if such a bubble distribution could be fitted as a function of the reionization parameters ($\zeta_{\rm ion}$, $M_{\rm min}$) then this could add extra information from SIMFAST21 that is not available to the FZH model directly.

5.5 Conclusions

The halo model approach can be used to predict rough amplitudes and shapes for the power spectra of simplified ionisation fraction data. In Section 5.2, the procedure was seen to accurately model mock ionisation fraction data with randomly-placed ionised bubbles of various sizes. Including the effect of bubble clustering in Section 5.3 requires suppressing the number density function of halos, since not all halos give rise to a separate isolated ionised bubble. Suppressing the number density gives better predictions for all but the latest redshifts where $x_{\rm HII} \leq 0.8$. For later redshifts the model is unable to account for the effects of widespread bubble overlap. Accounting for overlap is an issue that all analytic models of ionisation struggle with (including Furlanetto et al. 2004). Finally I compared the full halo model predictions with actual spectra from SIMFAST21 outputs. The final model would need significant improvement before being useful as a full replacement for SIMFAST21. The predictions would likely be improved by measuring the bubble size distribution directly from the simulations by using methods such as granulometry (Kakiichi et al., 2017) or the mean-free path method (Mesinger and Furlanetto, 2007). Performing the fitting for many different reionization scenarios specified by a particular set of reionization parameters such as ionisation efficiency $\zeta_{\rm ion}$ and minimum halo mass $M_{\rm min}$ could provide additional information that is available in the SIMFAST21 outputs, that is not available to the analytic FZH model.

Analytic models such as the one in this chapter have a number of benefits. Most significantly, analytic models provide a direct method of understanding the underlying physics. The complex interactions within simulations can often lead to bizarre emergent properties, and it can be difficult to determine whether the unexpected re-

sults are the cause of a poor simulation or an indication of new physics. Unexpected predictions from analytic models, however, can be understood in terms of the model assumptions – highlighting what new data or assumptions could improve the model. If model interpretability is not an important requirement, then Chapter 6 describes an entirely different technique for modelling the Epoch of Reionization: using machine learning to emulate the behaviour of the simulations without any knowledge of the underlying physics. The predictions of such 'surrogate models' in Chapter 6 are much closer to the simulated power spectra than the analytic model in this chapter.

Chapter 6

Emulating Epoch of Reionization simulations

Upcoming experiments such as the SKA will provide huge quantities of data. Fast modelling of the high-redshift 21cm signal will be crucial for efficiently comparing these data sets with theory. The most detailed theoretical predictions currently come from numerical simulations and from faster but less accurate semi-numerical simulations. Semi-numerical simulations take minutes to hours to run.

In this chapter I evaluate the viability of five machine learning techniques for emulating the 21cm power spectrum from SIMFAST21. This work was published in Jennings et al. (2018). I analyse the prediction speeds of the resulting emulators and their accuracy across the standard reionization input parameter space. The best emulator is a multilayer perceptron with three hidden layers, reproducing SIMFAST21 power spectra 10^8 times faster than the simulation with 4% mean squared error averaged across all redshifts and input parameters. The other techniques (interpolation, Gaussian processes regression, and support vector machine) have slower prediction times and worse prediction accuracy than the multilayer perceptron.

See Section 2.5.4 for a review of recent machine learning techniques for the EoR. All the emulators in this chapter differ from those in Schmit and Pritchard (2018) and Kern et al. (2017), which were trained at fixed scales and fixed redshifts. Such emulators make predictions only at these fixed scales and redshifts, so that if other scales or redshifts are desired one must interpolate further. Using the scales

and redshifts directly as extra inputs to the trained models allows them to make predictions for any requested scale and redshift. This method is theoretically more flexible but results in significantly worse prediction accuracy at lower redshifts.

The rest of the chapter is split in to the following sections.Section 6.1 briefly describes the specifics of how the emulators were trained. I present the results of training the emulators in Section 6.2. Section 6.3 is a discussion of the accuracy and speed performance of the different machine learning techniques, and how their performance depends on the input parameters. The best emulator candidate is then used in Section 6.4 to present a proof-of-concept technique for determining a relationship between two different simulations. The technique is demonstrated by finding a mapping between the inputs of SIMFAST21 and those of 21CMFAST by measuring which inputs result in the most similar output power spectra. I conclude in Section 6.5 by reiterating the motivations for these emulators and describing potential use-cases for the final trained emulators.

For cosmological parameters I use $\Omega_{\rm M} = 0.270$, $\Omega_{\rm b} = 0.046$, $\Omega_{\Lambda} = 0.730$, $H_0 = 71.0 \mathrm{km\ s^{-1} Mpc^{-1}}$, $n_{\rm s} = 0.960$, $\sigma_8 = 0.810$, the default parameters in the SIMFAST21 package[1].

6.1 Emulator training

This section describes how training data were gathered and the specific choices made while training the emulators. The training results are given later in Section 6.2. All emulators are trained on the same architecture, each on a single node using 16 Xeon E5-2650 cores and 128GB RAM. The emulators are trained to reproduce correlations in fluctuations of the differential brightness temperature. See Section 2.2.2 for a review of these correlation functions.

The noise in this chapter is sample variance from randomly seeding different density fields at the start of each simulation. I do not include sources of noise from experimental factors such as instrumental noise because these emulators are intended as efficient replacements for the expensive simulations themselves. For

[1] https://github.com/mariogrs/Simfast21

comparison with observed telescope data, instrumental noise can be added in the comparison stage after running the clean emulated simulations.

6.1.1 SIMFAST21 simulations

I run 2000 SIMFAST21 simulations in total, retaining only the three input reionization parameters and the final output spherically averaged power spectra for each simulation. This number of simulations was chosen as the smallest number which gives a visible reasonable sampling of the three-dimensional input space. I test the effect of using fewer simulations on the final model accuracy in Figure 6.3. I use 1000 simulations for training, 500 for validation, and another 500 for testing the emulators which have highest prediction accuracy on the validation data. This follows the standard process of train-validate-testing described in Section 2.4. Each simulation generates three-dimensional realisations of the $\delta T_{\rm b}$ field in a cube of size 500 Mpc resolved into 512^3 pixels (smoothed from density fields resolved into 1536^3 pixels). This gives power spectra values for seven redshift values: $\{8.0, 9.5, 11.0, 12.5, 14.0, 15.5, 17.0\}$ and thirteen k-values in the range $\{0.02, 3.0\}$ hMpc^{-1}. This corresponds to 91000 overall training data points, and 45500 data points each for validation and testing. The power spectra data have size of 335MB for all 2000 simulations, compared to 7TB size of all $\delta T_{\rm b}$ boxes.

6.1.2 Training set design

The emulators map five input values to a single output target value. The target value is the $\delta T_{\rm b}$ power spectrum value for the given inputs. The first three input values are the three reionization parameters (see Section 2.1.2), which are different for each simulation. The final two inputs are the redshift z and the k-value, the values for which are constant across all simulations and are given in Section 6.1.1. The function $f(\mathbf{x})$ which the models are fitting is then the spherically averaged 21cm power spectrum $P_{\Delta T_{\rm b}}(M_{\rm min}, \zeta_{\rm ion}, R_{\rm max}, z, k)$.

I use the Latin Hypercube method (McKay et al., 1979) to choose the reionization parameter values for the simulations. The Latin Hypercube method samples the three-dimensional input space more efficiently than a naive exhaustive grid search.

I use the following ranges and scalings for the reionization parameters:

1. $M_{\rm min}$ in the logarithmic range $[10^{7.8}, 10^{9.8}]$;

2. $\zeta_{\rm ion}$ in the linear range $[5, 100]$;

3. $R_{\rm max}$ in the linear range $[5, 20]$.

These ranges match those used by the semi-numerical simulation authors, see for example Greig and Mesinger (2015). See Section 4.1.2 for a discussion of the $M_{\rm min}$ and $\zeta_{\rm ion}$ ranges. The $R_{\rm max}$ range arises from recombination models (Sobacchi and Mesinger, 2014), and only has an effect near the end of reionization when the ionised bubble sizes are comparable to $R_{\rm max}$ (Alvarez and Abel 2012, Mcquinn et al. 2006). See Figures 6.10 and 6.11 later for example power spectra across these ranges for $\zeta_{\rm ion}$ and $M_{\rm min}$ values. Another common method for evenly sampling a high-dimensional space is to choose points randomly within fixed bounds. This random sampling method is slightly less efficient than latin hypercube sampling: it can give rise to 'wasted' points which lie close to one another. Unless the data are subject to very high noise or sample variance, using two points in the same point in parameter space usually does not provide any more information than using a single point. The latin hypercube method ensures that all points are evenly spaced and not wasted.

I also test three different scaling types for the target values to determine which gives the most accurate emulation. These three are a linear function $y = P_k$, a logarithmic function $y = \log[P_k]$, and a pseudo-logarithmic function $y = \sinh^{-1}[P_k]$ sometimes called luptitude (Lupton et al., 1999). I test logarithmic scaling as an attempt to exploit the fact that power spectra appear more naturally spaced in logarithmic space $\log[P_k]$ than in linear space P_k. However a few percent of the power spectra data are zero-valued, especially at early and late redshifts where the ionisation field $x_{\rm HII}(\mathbf{r})$ becomes uniform and $\delta T_{\rm b}$ is effectively zero everywhere (Pritchard and Loeb 2012, pages 12-13). The motivation for luptitude scaling is to retain as much data as possible: a purely logarithmic scaling would require us to throw away

all zero-valued data points and reduce the size of the training data set. I comment on the effects of including or excluding these zero-valued data in Section 6.3.6.

6.1.3 k-range restriction

I exclude the largest and smallest scales from the validation and testing data, including only $0.1 \leq k \leq 2.0$ values. On large scales ($k < 0.1$ hMpc^{-1}), the power spectrum is affected by foregrounds (Datta et al., 2010). The finite resolution of the simulations means that there is little information in the power spectrum on very small scales ($k > 2.0$ hMpc^{-1}). These restrictions are common for semi-numerical simulations, see for example Greig and Mesinger (2015).

6.1.4 Goodness of fit evaluations

For validation and testing I measure the goodness of fit between predicted target values $y^*(k,z)$ and measured target values $y(k,z)$ using the mean squared error

$$\mathrm{MSE}\big[y(k,z), y^*(k,z)\big] = \frac{1}{N_\mathrm{z} N_\mathrm{k}} \sum_{z}^{N_\mathrm{z}} \sum_{k}^{N_\mathrm{k}} \left(\frac{y(k,z) - y^*(k,z)}{y(k,z)} \right)^2 \tag{6.1}$$

along with the percentage mean squared error, $100 \times$ MSE. The MSE is averaged over all N_z redshifts values and all N_k scale values in the range $0.1 \leq k \leq 2.0$, unless explicitly mentioned otherwise. For comparability I use this same error function for all different emulators during validation and testing, although the models use different error metrics for determining their training convergence (see Section 2.4 for the training objective functions for each model).

6.2 Emulator training results

After training each emulator I test its accuracy by generating predictions for a set of unseen validation data. By calculating the MSE value in Equation (6.1) between the predicted outputs and the true outputs, I determine which emulator makes the most accurate predictions. A low MSE means a high prediction accuracy.

6.2.1 Target value scaling

Here I compare the prediction accuracy for the three scaling methods of the target power spectra values: linear, logarithmic, and pseudo-logarithmic $\sinh^{-1}(x)$. As

expected, the linear function has poor prediction accuracy because the power spectra values are more naturally spaced in logarithmic space than in linear space. The logarithmic function works fairly well at intermediate redshifts for this same reason, but all of the zero-valued power spectrum values had to be discarded as $\log(0)$ is undefined. The pseudo-logarithmic function $\sinh^{-1}(x)$ has the highest prediction accuracy over all redshifts and allows us to retain all training data points (with zero-valued outputs or otherwise). I use the pseudo-logarithmic function in all emulators from here on.

6.2.2 Hyperparameter searching

Each model has a set of trainable values referred to as fitting parameters. Many models have an additional set of values which must be fixed even before starting to train, referred to as hyperparameters. See Section 2.4.6 for a review of different search strategies for deciding which hyperparameters to use. Here I describe which hyperparameters (if any) were varied for each model type, and which hyperparameters values give rise to the best prediction accuracy. For MLP models, I first choose a set of fixed hypers that remain constant through the chapter. These are: a constant learning rate of 0.001; batches of size 200; the rectified linear unit function RELU as the activation function; and an L2 regularlization value of $\alpha = 0.0001$. For each model I restrict the total training time for all hyperparameter searching to 156 CPU hours. The interpolation models involve no hyperparameters, and for the SGPR model I simply increase the number of inducing points m until the individual model's training time reaches 156 CPU hours. Increasing m should always increase the SGPR model's accuracy and so the value of m is not treated as a hyperparameter when considering the total training time. Including models with smaller m values in the total training time would give a smaller maximum value of m, making an unfair comparison with the other models.

MLP layer sizes

I use MLP models with one, two and three hidden layers. The sizes of the hidden layers were varied linearly in the range $[0, 200]$ using a simple grid-search method. This range of layer sizes was chosen as being roughly one order or magnitude larger

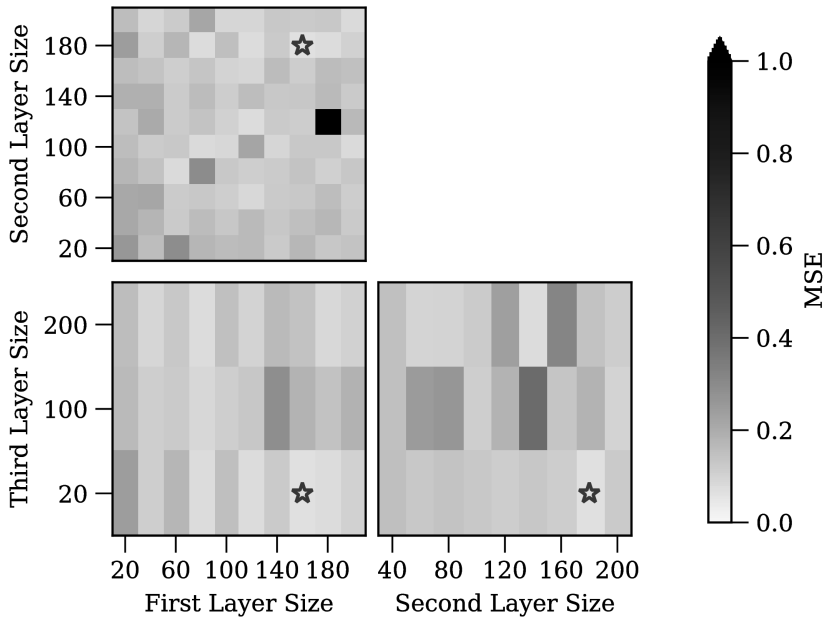

Figure 6.1: Mean squared error on the validation data for three-layer multilayer perceptron models, as a function of the sizes of each hidden layer. The star shows the layer sizes of the MLP emulator with the highest prediction accuracy: 160 neurons in the first hidden layer, 180 neurons in the second hidden layer, and 20 neurons in the final hidden layer.

than the number of inputs. This ensures that the number of trainable parameters (i.e. weights) is large enough for the model to be able to fit the complex five-dimensional function of the power spectrum. Collectively, the number of layers and the sizes of the layers are often referred to as the 'network architecture'. Generally, the emulator models with more hidden layers have higher prediction accuracy. The validation MSE values for the best one-, two- and three-layer MLP emulators are 13%, 2.3%, and 1.6% respectively. The validation MSE values for three-layer MLP emulators are shown in Figure 6.1 as a function of the sizes of each of the three hidden layers. Most three-layer MLP models have a low validation MSE near 10%. The best emulator has hidden layers sizes $160-180-20$, the architecture for which are indicated by the location of the star. The MLP models end training when the change in objective function changes more slowly than a threshold tolerance for several training epochs. Most of the MLP models achieved this in fewer than 400 training epochs, with some 1-layer models lasting up to 800 epochs.

SVM margin hyperparameters

I test a range of SVM emulators with different values for three hyperparameters controlling the margin. I vary the penalty parameter C logarithmically in the range $[10^{-3}, 10^{3}]$; the tolerance EPSILON logarithmically in the range $[10^{-3}, 10^{0}]$; and the kernel influence range GAMMA logarithmically in the range $[10^{-3}, 10^{3}]$. These hyperparameters are the suggested ranges by SKLEARN and I use a simple grid-search to find the best hyperparameters. I also test three kernel functions: RBF, sigmoid, and polynomial. Figure 6.2 shows how the validation MSE of emulators using the RBF kernel depends on the SVM hyperparameters. The different colour-map is used to emphasise that the colour range is logarithmic and has a much larger spread of MSE values between 0.2 and 2000 (or between 20% and 2×10^{5}%). The best SVM emulator has validation MSE of 20%, using hyperparameters C $= 1.0$, EPSILON $= 10^{-3}$, GAMMA $= 1.0$ and the RBF kernel. All SVM emulators with kernels other than RBF have much worse validation MSE: the best polynomial and sigmoid SVM emulators have validation MSEs of 50000% and 500% respectively.

6.2.3 Overfitting tests

For each model I determine the best hyperparameters by trying a range of values and selecting the emulator which shows the highest prediction accuracy on the validation data. By trying different hyperparameter values we can usually find a closer fit to the data. However, this process is sensitive to over-fitting: the model might fit the training data more closely, but it may not extend well to new data. I test for overfitting by training a series of emulators with increasing training dataset sizes, keeping the hyperparameters fixed at the proposed best values. Providing more training data should give rise to improved predictions for the unseen validation data. If providing more training data instead leads to a decrease in validation prediction accuracy, then overfitting has occurred: the model makes good predictions for the training data, but does not extend well to new input values. Figure 6.3 shows the results of these tests, giving the mean square error on the validation data for each model, using differently sized training datasets. Most of these mean squared errors generally decrease with increased training set size, implying that none has been overfitted.

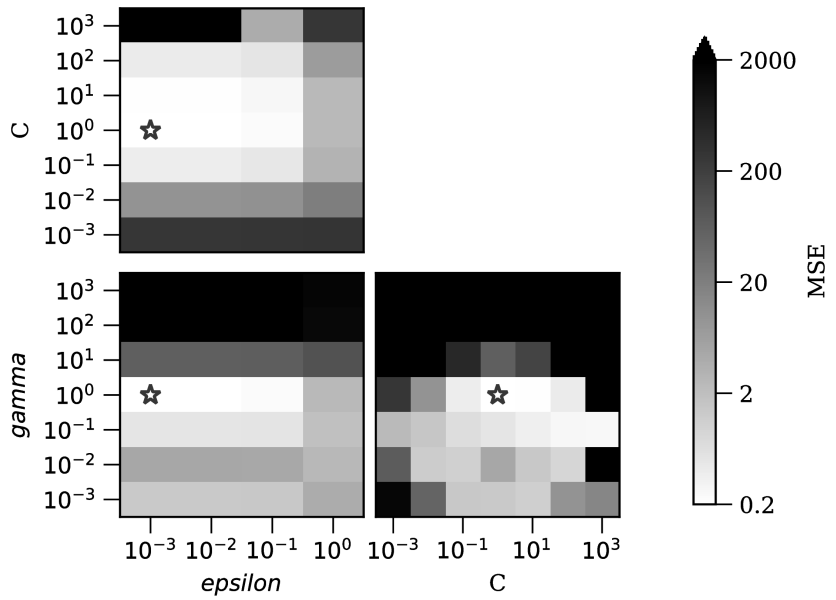

Figure 6.2: Mean squared error on the validation data as a function of the model hyperparameters, for support vector machine emulators using the RBF kernel. The hyperparameters are the penalty term C, margin tolerance EPSILON and influence range GAMMA. The spread of MSE values is much larger for SVM models, indicated by the logarithmic colour scale between MSE values of 10^{-1} and 10^{3}. The hyperparameters of the highest prediction accuracy SVM model are indicated by the star: C $= 1.0$, EPSILON $= 10^{-3}$ and GAMMA $= 1.0$.

The final MLP emulator shows a slight up-turn indicating that it may be slightly overfitted. This emulator has the best performance on the testing data, however, so the overfitting is likely minor. In general, deciding when to stop gathering training data depends on the final accuracy requirements of your model. By plotting the prediction accuracy against the training set, one can immediately see the extent that providing more training data would improve the model. Emulators are often used to replace simulations in parameter estimation methods, comparing the results of the simulations to the observed data and finding which simulated model(s) are most similar to the observed data. In this case, one way to decide the required accuracy target is to match the noise level of the observed data. With noisy observed data, the accuracy of the emulator will have less impact on the parameter constraints than the effect of the noise. On the other hand, if the data are effectively noiseless then the emulator accuracy will be extremely important. With noisy data I might only need

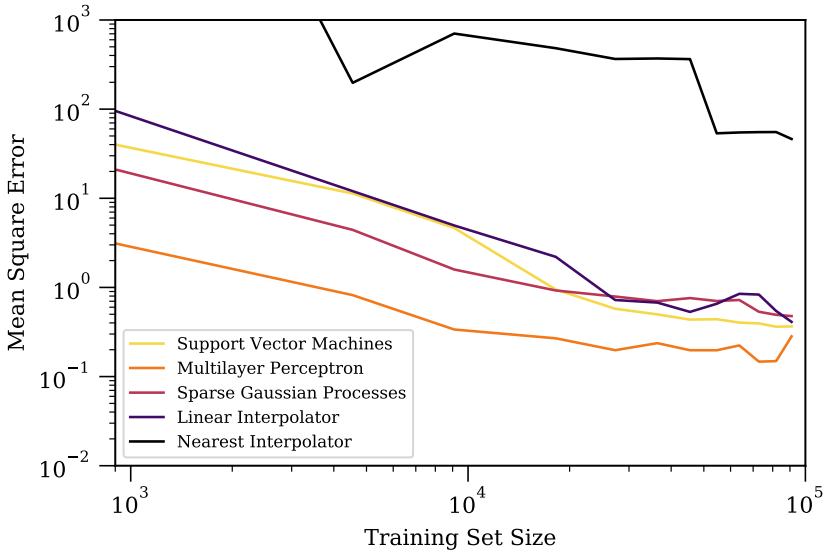

Figure 6.3: Mean squared error on the validation data as a function of training set size. The best hyperparameters were fixed for each model, and the emulator retrained with more training information. The MSE curves generally improve with more training data, implying that none has been overfitted. Note that the x-axis here can be seen as related to the number of simulations that are used: a training set size of 10^3 corresponds to roughly ten simulations, and a size of 10^5 corresponds to all thousand of the training simulations.

a mean square error accuracy of 1.0. Many of the models in Figure 6.3 achieve this accuracy and, in fact, the multilayer perceptron reaches this accuracy with fewer than half of the training data points. I could have trained my model after running only 500 simulations and saved weeks of training time. In the other extreme, if my data are less noisy I might decide that the MSE accuracy has to be better than 0.1. Figure 6.3 shows that none of the models has yet achieved this accuracy, and I would need to provide more training data (or higher quality training data) to reach that accuracy.

6.2.4 Performance on testing data

Here I test the performance of the best emulator for each model type using all 500 simulations in the testing set. Table 6.1 shows the accuracy and speed of each emulator for making predictions on the entire testing dataset. The global MSE percentage is averaged across the entire testing data set. In Section 6.3.6 later, I discuss the fact that the trained emulators have worse accuracy at lower redshifts. I

Model Type	Test MSE %	Test MSE % for $z \geq 10$	Prediction Time
Nearest Neighbour	290 % *	5.1 %	0.20s
SGPR ($m = 2730$)	36 %	0.6 %	116s
SVM	32 %	2.1 %	27s
1-layer MLP	27 %	9.2 %	0.07s
Linear Interp	17 %	1.6 %	4.1 hours *
2-layer MLP	4.5 %	2.3 %	0.14s
3-layer MLP	3.8 %	1.4 %	0.27s

Table 6.1: Speed and accuracy performance of the best emulator for each technique, using the testing data set. The percentage MSE values here are $100 \times$ MSE. The rows are sorted in order of prediction accuracy, from highest MSE (least accurate) at the top to lowest MSE (most accurate) at the bottom. I give MSE values averaged across the entire dataset and also the MSE for a subset of the testing data with $z \geq 10$ to demonstrate that most of the poor accuracy occurs at later redshifts. See Section 6.3 for a discussion on the extreme (*) values for the two naive interpolation methods. For all models except SGPR, the total time for hyperparameter searches is 156 CPU hours. For SGPR model, I run a single model with the largest possible number of inducing points m without exceeding 156 hours training time.

include a column in Table 6.1 for the percentage MSE averaged across the testing data with higher redshifts ($z \geq 10$). Figure 6.4 shows an example of the power spectra outputs from the best emulator of each model type, showing the predictions for a single test simulation near the canonical reionization parameters at $z = 9.5$.

6.3 Emulator training discussion

6.3.1 Interpolation

Figures 6.5 and 6.6 show the prediction MSE of the nearest-neighbour and linear interpolation emulators as a function of location in parameter space. These two models are the worst candidates for emulating SIMFAST21 behaviour.

The nearest-neighbour interpolation model has poor prediction accuracy both in terms of the global MSE value of 290% from Table 6.1, and the local MSE across parameter space shown in Figure 6.5. The model uses the nearest-neighbour look-up method of SCIPY.SPATIAL.KDTREE which is fast but makes no account of noise or smoothness in the simulation behaviour. The linear interpolation model emulates

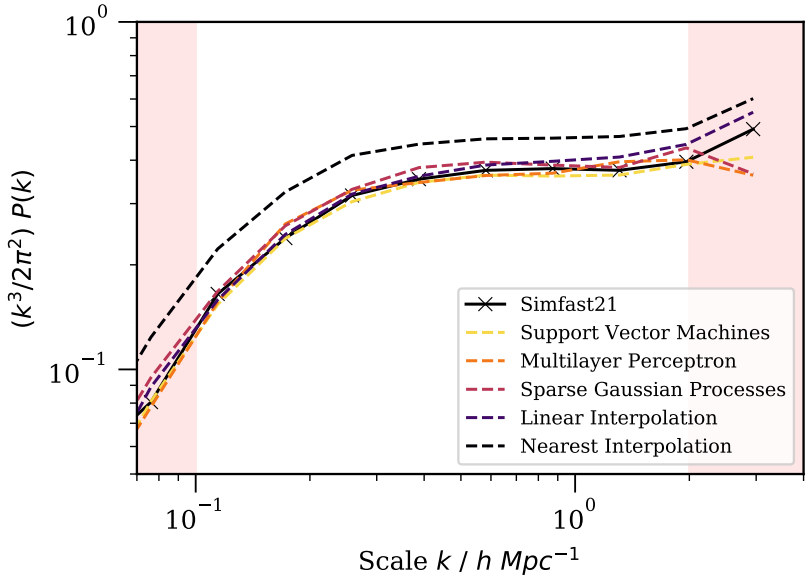

Figure 6.4: Predicted δT_b power spectra of a canonical simulation with reionization parameters $\{5 \times 10^8 M_\odot, 30.0, 10$ Mpc$\}$. Dotted lines show the predictions from the best emulator of each type. Solid line shows the power spectrum from an actual SIMFAST21 simulation. The red shaded areas indicate the k-values that were excluded from the validation and testing. This test simulation was chosen from the testing data as the nearest to the canonical reionization parameters. The model using nearest-neighbour interpolation has significantly different predictions, likely owing to the underfitting processes discussed in Section 6.3. No error bars are shown because the measured curve is from a single simulation, and the predicted curves are from deterministic ML models which do not generate uncertainty estimates.

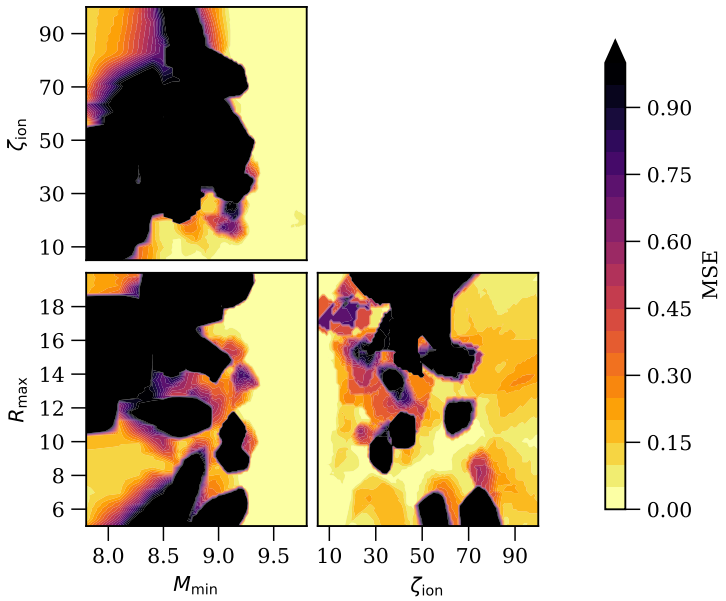

Figure 6.5: Local MSE performance for the nearest ND interpolation model

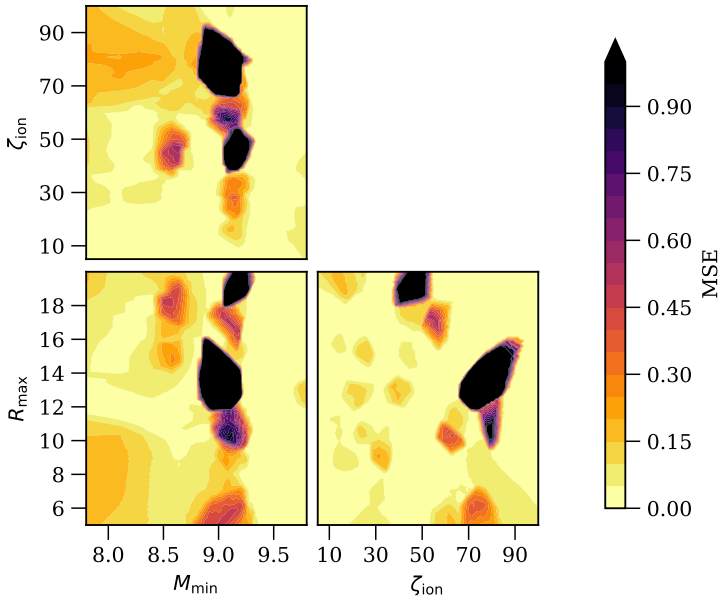

Figure 6.6: Local MSE performance for linear ND interpolation model

the SIMFAST21 behaviour more closely: the global MSE is 17% and the local MSE in Figure 6.6 shows larger regions of good accuracy. This accuracy is at the expense of much slower prediction times. The nearest neighbour model makes predictions for the whole testing dataset in less than a second, whereas the linear interpolation model takes several hours. These results indicate that interpolation methods cannot efficiently capture the complicated behaviour of SIMFAST21, justifying the need for more complex machine learning techniques.

6.3.2 Sparse Gaussian processes regression

The sparse Gaussian processes model is a poor emulator candidate. Both the local MSE in Figure 6.7 and the global MSE of 36% are poor. The accuracy of the model would almost certainly be improved by increasing the number of inducing points, which would lessen the matrix inversion approximation. However, training models with $m > 2730$ would require more than the allowed CPU time. The value of $m = 2730$ is chosen as the largest number of inducing points whose model training time does not exceed 156 hours. A hard upper limit of $m < 18000$ is found for the 128GB RAM architecture used in this chapter, since values of m larger than this cause a RESOURCEERROR in TENSORFLOW. Moreover, increasing the number of

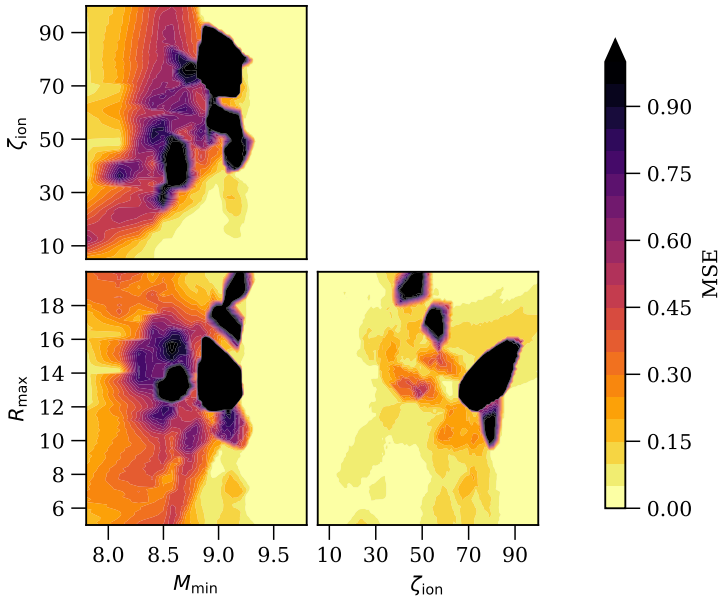

Figure 6.7: Local MSE performance for the best sparse Gaussian processes model

inducing points also increases the prediction time: using $m = 910$ takes 16 seconds to make predictions for the testing dataset; using $m = 2730$ takes 116 seconds. Increasing the number of inducing points gives better accuracy at the expense of much slower prediction times. It is likely that the model could be greatly improved by using methods such as Principle Component Analysis as in Kern et al. (2017), since Gaussian Process models scale poorly with high numbers of dimensions and large data sets.

Interestingly, if data for $z < 10$ are excluded from performance testing then the SGPR model changes from being one of the least accurate models to being the most accurate one. The reason that SGPR in particular makes poor low-redshift predictions is likely due to the effect of the MATERN32 kernel on the problematic low-redshift power spectra. The MATERN kernel tries to fit a smooth function to the sudden behaviour. Although smoothly fitting appears to work well at higher redshifts, since simulations with similar inputs should give similar output power spectra. At lower redshifts however, the emulator struggles both to enforce smoothness and to allow for the sudden jumps seen in the training data.

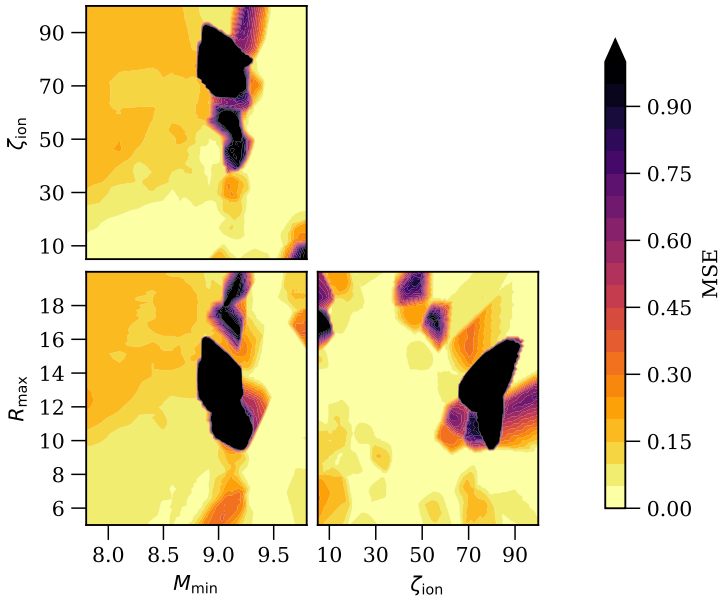

Figure 6.8: Local MSE performance for the best support vector machine model

6.3.3 Support vector machine

Figure 6.8 shows the local MSE performance of the best SVM model. This model is a poor candidate for a SIMFAST21 emulator. The global MSE of 32% from Table 6.1 is one of the worst. This model also has slow prediction speeds, taking 27 seconds to make predictions of the testing data (100 times slower than the best MLP model). It is possible that using other kernels and doing deeper hyperparameter searches would give better accuracy. Given the long prediction times for these models, it is unlikely that any SVM models would outperform the best MLP emulator, either in terms of speed or accuracy.

6.3.4 Multilayer perceptron

Figure 6.9 shows the prediction MSE of the best MLP emulator as a function of location in parameter space. The dark regions indicate the regions of parameter space which are most difficult to emulate.

The three-layer multilayer perceptron is the best candidate for emulating SIMFAST21 behaviour. Table 6.1 shows that this emulator makes fast and accurate predictions for the test dataset, taking less than a second to match the true simulation outputs within 4% mean squared error averaged across the whole input parameter

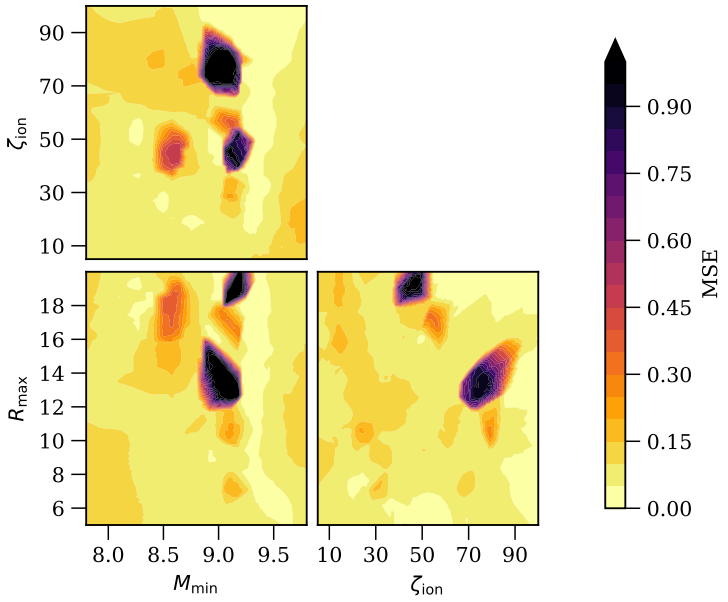

Figure 6.9: Local MSE performance for the best three-layer multilayer perceptron model

space. Figure 6.9 shows that the emulator makes accurate predictions across a wide range of input parameters. Worse performance is seen for MLP emulators using fewer hidden layers: increasing the number of layers allows MLP models to be more flexible, and the results indicate that one- and two-layer MLP models are not flexible enough to fit the simulation outputs as accurately as three-layer models.

Figures 6.10 to 6.12 show several example power spectra for a range of $\zeta_{\rm ion}$, $M_{\rm min}$ and $R_{\rm max}$ values, also showing the predicted power spectra from this best emulator. The shaded red regions in these figures indicate the ranges of excluded k-values. The simulated spectra in these figures are from the test dataset. No error bars are given for the simulated spectra because each line is for a single simulation only, so that no spread of values can be measured. Similarly each prediction is from the best emulator, which gives deterministic results such that fixed input parameters always give the same output power spectrum. It is possible to train neural networks that generate an estimate of the error bars (see for example Bayesian neural networks Hinton and Van Camp 1993, or using 'dropout' throughout the network Gal and Ghahramani 2016) but these methods require considerably more training time, and are not used in this chapter. Finally Figure 6.13 show the same predicted power

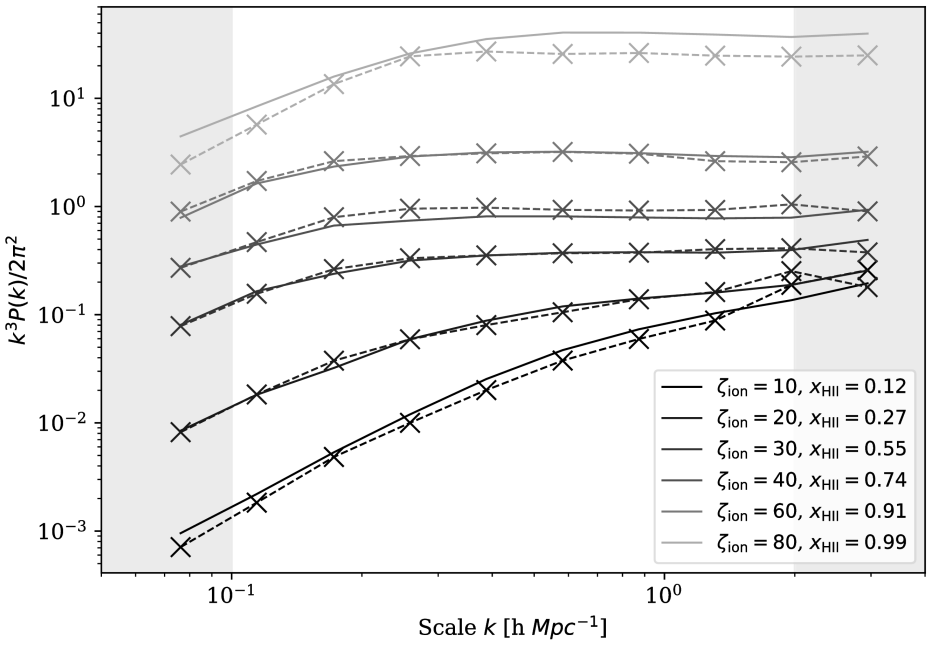

Figure 6.10: Example emulated and simulated power spectra for a range of $\zeta_{\rm ion}$ values at $z = 9.5$, for fixed $M_{\rm min} = 5 \times 10^8$ and $R_{\rm max} = 10$. Solid line shows the simulated power spectra, dotted line shows the predicted power spectra from the best emulator. The ionised fraction for each line is given in the legend.

spectra as in Figure 6.10, instead colouring by the mean-squared error of the prediction. The two predictions at the extremes of the input parameter space ($\zeta_{\rm ion} = 80$ and $\zeta_{\rm ion} = 10$) have the worst prediction accuracies. This feature of making poorer predictions on the edge of the parameter space is common for machine learning models because of the information in the training data. Predictions for parameter values near the centre of the input space generally have a large number of nearby training data points, whereas parameter values near the edge of the input space generally have far fewer nearby training data points. Predicting the power spectrum for scenarios in the centre of the input space is thus far easier than predicting the power spectrum for scenarios on the edge of the input space, as observed in Figure 6.10.

Given the benefit of most three-layer models over two-layer models, it seems likely that models using four or more layers could provide even closer fit to the training data. Such models were not investigated given the fixed upper limit on training time. Additionally, the benefit of adding more layers would likely be minimal as there is a clear case of diminishing returns for each additional layer: the

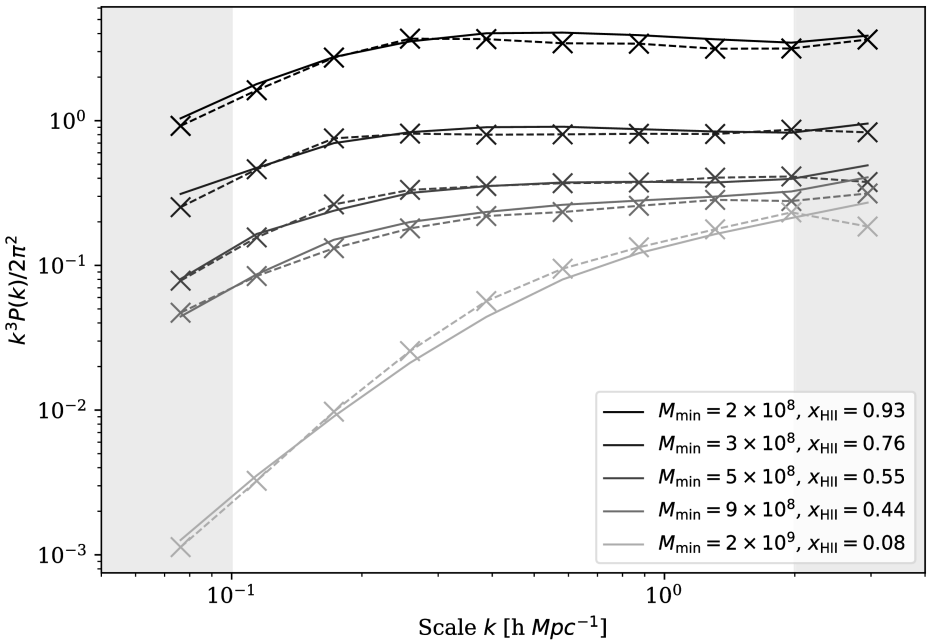

Figure 6.11: Example emulated and simulated power spectra for a range of $M_{\rm min}$ values at $z = 9.5$, for fixed $\zeta_{\rm ion} = 30$ and $R_{\rm max} = 10$. Solid line shows the simulated power spectra, dotted line shows the predicted power spectra from the best emulator. The ionised fraction for each line is given in the legend.

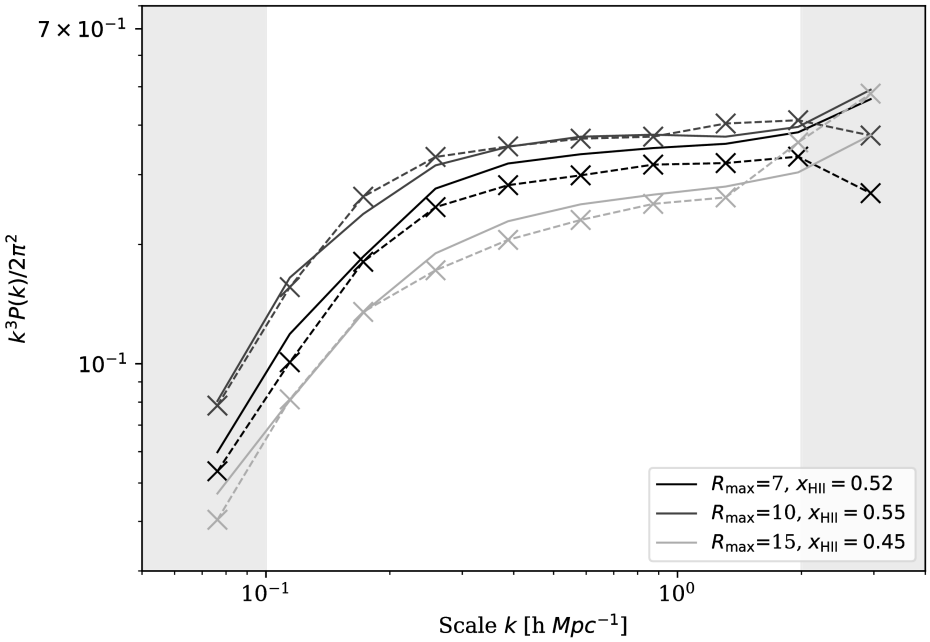

Figure 6.12: Example emulated and simulated power spectra for a range of $R_{\rm max}$ values at $z = 9.5$, for fixed $\zeta_{\rm ion} = 30$ and $M_{\rm min} = 5 \times 10$. Solid line shows the simulated power spectra, dotted line shows the predicted power spectra from the best emulator. The ionised fraction for each line is given in the legend.

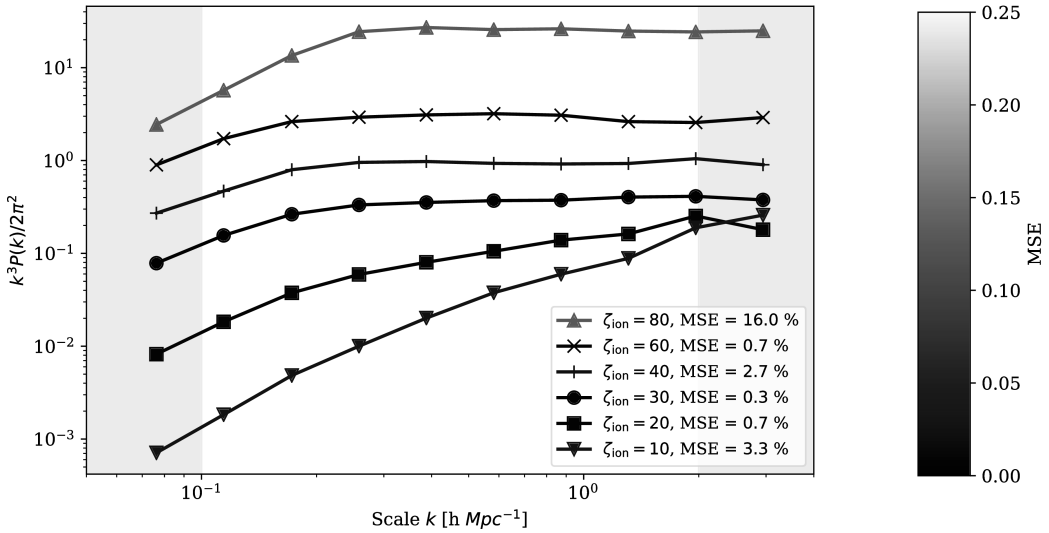

Figure 6.13: Example emulated power spectra for a range of $\zeta_{\rm ion}$ values at $z = 9.5$, for fixed $M_{\rm min} = 5 \times 10^8$ and $R_{\rm max} = 10$. Colours indicate the overall MSE of the prediction, with markers added to discriminate between the different lines. The MSE for each line is given in the legend. The two predictions at the extremes of the parameters space ($\zeta_{\rm ion} = 80$ and $\zeta_{\rm ion} = 10$) have the worst prediction accuracies, which is a common feature of machine learning models as discussed in the text.

best MSE for one layer was 27%; two layers gave 4.5% MSE; and three layers gave 3.8%.

6.3.5 Mass turnover performance

All of Figures 6.5 to 6.9 show a region of poorer prediction accuracy for inputs near $M_{\rm min} = 10^9$. This is likely due to the finite mass resolution of the SIMFAST21 simulations. For values of $M_{\rm min}$ near the mass resolution, the simulation switches between containing both resolved and unresolved halos (if $M_{\rm min} < 5 \times 10^9$), and containing only resolved halos (if $M_{\rm min} > 5 \times 10^9$). The change in behaviour appears to be difficult to emulate for all model types.

6.3.6 Low redshift performance

The prediction accuracy of the emulators is worse for lower redshifts than for higher redshifts. If data for $z < 10$ are excluded from performance testing, then all emulators improve significantly: for instance, the three-layer multilayer perceptron's percentage MSE improves from 3.8% to 1.4%. The improved values using only

high-redshift power spectra are presented in the third column of Table 6.1. There are two effects that could be causing the worse accuracy at lower redshift, which I discuss here.

First, my emulators differ from those of Kern et al. (2017) and Schmit and Pritchard (2018) in that my models are trained using the redshift and k-scales as extra input dimensions. The motivation for including redshift and k-scales was to allow for immediate predictions at any redshift or k-scale. Without including these input dimensions the trained models would only make prediction at the fixed redshifts and k-scales of the training data. Making predictions at other input values with such fixed-input emulators would require further interpolation afterwards. Although using z as an input allows for more flexible predictions, this flexibility is likely a cause of poorer emulation at lower redshifts. Without more computing power or faster training algorithms, the results in this chapter would suggest that future attempts to emulate the power spectrum should be done for fixed z inputs.

Secondly, a feature of the actual simulated power spectra could be another source of poor emulator accuracy. The amplitude of the power spectrum is highly sensitive to the global 21cm brightness temperature $\langle \delta T_{\rm b} \rangle$. At low redshifts near the end of reionization, $\langle \delta T_{\rm b} \rangle$ approaches zero (see Pritchard and Loeb 2012 for a review). This low value of $\langle \delta T_{\rm b} \rangle$ amplifies even small fluctuations in $\delta T_{\rm b}$, causing a sharp increase in the fluctuation field $\Delta T_b(\mathbf{r})$ from which the power spectrum is calculated. Soon after, reionization finishes and $\Delta T_b(\mathbf{r}) = 0$ everywhere so that the amplitude of $P_{\Delta T_{\rm b}}(k)$ jumps suddenly from high-amplitude to zero-amplitude. These two sudden features are difficult to emulate: the sharp increase in power spectrum amplitude near the end of reionization, and the sudden drop thereafter from high-amplitude to zero-amplitude. For $z \geq 10$, none of the simulations in the training data has completed reionization and thus none contain these problematic zero-valued power spectra. For $z < 10$, some models begin to near the end of reionization, especially those with high $\zeta_{\rm ion}$ and low $M_{\rm min}$. For $z = 8$, a large fraction of the parameter space shows these zero-amplitude power spectra in the training data.

It is interesting to note that all of Figures 6.5 to 6.9 show a region of poorer

prediction at around $\zeta_{\rm ion} = 70$, $M_{\rm min} = 10^9 M_\odot$, $R_{\rm max} = 14$ Mpc. By looking at Figure 6.14, it can be seen that this region of poor performance is particularly prominent in the later redshifts. It is likely caused by a combination of the mass turnover and low redshift performance issues: these regions have both a sudden change in the halo-mass calculation algorithm (including or not including any unresolved halos), and also are near the sudden jump in the power spectrum near the end of the EoR. Understanding precisely why this region has occurred is a difficult problem related to the interpretability of machine learning models as discussed in Section 2.4: the predictions are clearly less accurate, but it is extremely difficult to determine which weight-combinations of the MLP model, or which aspects of the training data, have caused this to be the case.

Figure 6.14 shows the best emulator's local MSE separately for the three lowest redshifts $z = \{8.0,\ 9.5,\ 11.0\}$. The local prediction accuracy at $z = 8$ shows large regions of poor predictions, especially for larger ionisation efficiencies and lower minimum halo masses. Training using $\delta T_{\rm b} - \langle \delta T_{\rm b} \rangle$ as the target values rather than $\Delta T_{\rm b}$ could remove the sudden changes in power spectra magnitudes and would be easier to emulate. Note for instance that Kern et al. (2017) were able to emulate down to $z = 5$ without reporting any issues for emulating these redshifts. Figure 6.15 shows normalised histograms of the global ionisation fraction values for all simulations in the data, separating into low redshift ($z < 10$) and high redshift ($z \geq 10$). For the higher redshifts, very few scenarios have finished reionization, and so their power spectra are easier to emulate. For low redshifts, many — but not all — scenarios have finished reionization, giving a sudden feature in the input parameter space and making it harder to emulate.

6.3.7 Extending to more parameters

In this chapter I have restricted my investigation to the standard three-parameter reionization model: $\zeta_{\rm ion}$, $M_{\rm min}$ and $R_{\rm max}$. Is is highly likely that future investigations will require theoretical models that include a larger number of parameters, for instance allowing different reionization heating scenarios. We might like to be able to use emulators to make predictions for these new scenarios too. This will require

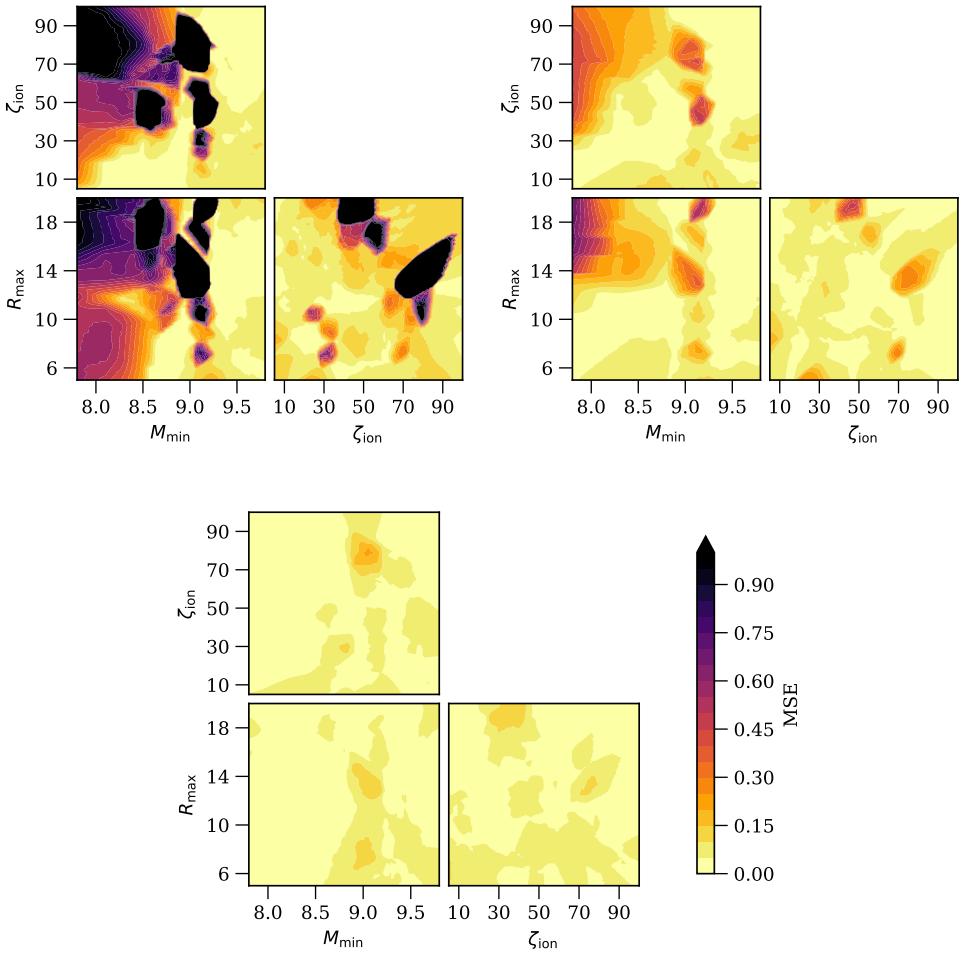

Figure 6.14: Mean squared error on testing dataset for the best MLP model as a function of prediction location, similar to Figure 6.9 but without averaging over redshift. The top left panel shows the performance for $z = 8.0$; the top right panel shows performance for $z = 9.5$; the bottom panel shows performance for $z = 11.0$. The prediction quality is much worse at low redshift than it is at high redshift, shown by the large darker regions at low redshift. The percentage MSE for $z \geq 11.0$ is better than 5% across almost all of the input parameter space. I omit plotting panels for each $z > 11.0$ here as they all look similar to that for $z = 11.0$.

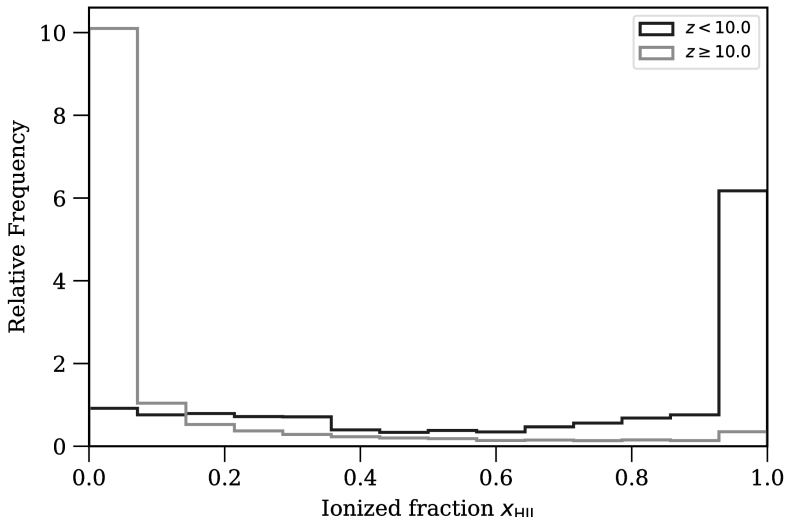

Figure 6.15: Normalised histogram of the ionised fraction for all simulations, for low redshifts ($z < 10$) and for higher redshifts ($z \geq 10$). The 21cm power spectrum is sensitive to the neutral fraction. The fact that many simulations are fully ionised for $z < 10$ could be one reason for the poor performance of the emulators at low redshifts.

generating new training data, increasing the training data set size and including a range of different heating scenarios. In order to make accurate predictions, emulators require a training data set that is well-sampled across the whole input space. With a higher-dimensional input space, the emulator will almost certainly require a much larger number of simulations to fulfil this well-sampled requirement. It will also likely be necessary to increase the size of the models, for instance increasing the number of layers in the multilayer perceptron model. Model size is related to the complexity of the function that the model is representing: small models with only a few trainable parameters can only mimic simple functions, whereas larger models can capture more complex functions. Increasing the number of parameters would give rise to a more complex function, hence larger models with more trainable parameters would be needed to capture this increase complexity. Both of these problems (training data size and model size) will require much more computational power for training, and so it is natural to search for ways to try to make use of our existing models. For this, we can use *transfer learning*. Transfer learning is the repurposing of an existing model for a new (but ideally similar) problem. I describe

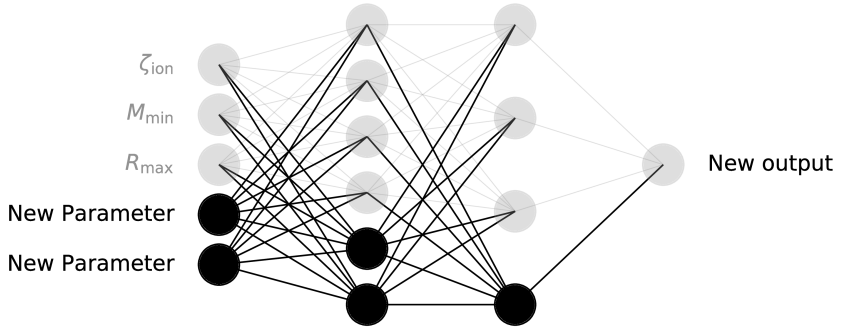

Figure 6.16: Transfer learning by adding new nodes to the existing network. This overall model is fitted to new training data. During training the existing weight-connections (shown in gray) are fixed, and only the new training weight connections (shown in black) are allowed to change. This allows the network to retain its former knowledge, while at the same time learning the effect of the new input parameters.

two similar methods for transfer learning here.

The first method is to add new layers or nodes to the existing model, and only learn the best values for the new weights. Figure 6.16 shows a schematic of this process. The weights and nodes in the existing network are fixed, so that the model retains its former knowledge. All new weights are trained using the new training data, so that the model learns how the new parameters affect the power spectrum. A second similar approach might be to add a completely new network after the existing network. Figure 6.17 shows a schematic of this process, with the old network prediction acting as an input node to the new secondary network. In this case, the secondary network learns how the new parameters can be used to *modify* the existing network's predictions. Either of these methods will likely reduce the required training time, since the model does not have to relearn its understanding of how the old parameters affect the power spectrum.

6.4 Mapping between SIMFAST21 and 21CMFAST

21CMFAST and SIMFAST21 are two common semi-numerical simulations for generating predicted 21cm maps during the Epoch of Reionization. In this section the best emulator is used to investigate the extent to which the two simulations give similar outputs for similar inputs. The motivation for this is to demonstrate

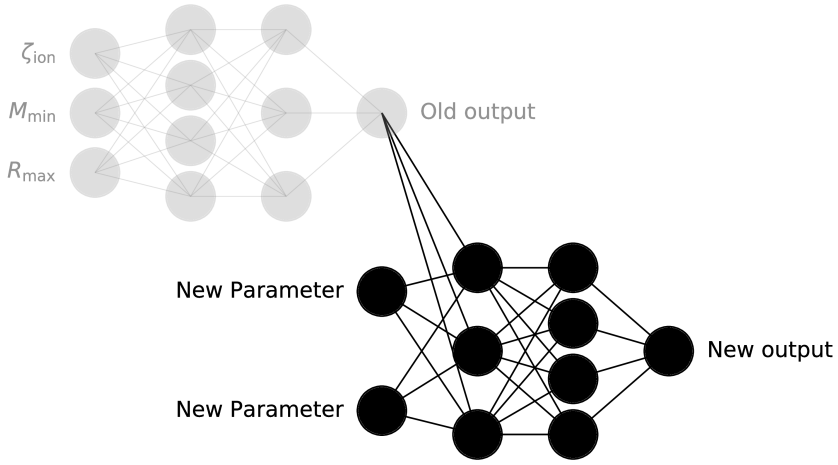

Figure 6.17: Transfer learning by adding a secondary network (shown in black) after the existing network (shown in gray). The new parameters are used as inputs to the secondary network, with the power spectrum output from the old network being used as an extra input. The overall model is fitted to new training data, with only the new weight-connections being varied. The secondary network learns only how the new parameters modify the previous power spectrum prediction.

a method for creating a mapping between any two simulations which generate the same output statistic. In particular, this method could be extended to give a mapping between SIMFAST21 power spectra and those from more accurate (but slower) three-dimensional radiative transfer simulations, such as C^2-RAY (Mellema et al., 2006a). Although numerical simulations have different input parameters, it would still be possible to map between them by finding the parameters which give the most similar power spectra.

When analysing huge datasets, SIMFAST21 could be used to give coarse constraints on reionization parameters. Using the mapping between SIMFAST21 and the more accurate numerical simulation, the coarse contours could be mapped to their equivalent regions of the numerical simulation inputs. This would allow more detailed exploration of this smaller region of parameter space with the numerical simulations.

6.4.1 Matching reionization histories

Section 2.1.3 describes the default procedure of 21CMFAST, in particular highlighting how it differs from the SIMFAST21 algorithm. In this subsection I discuss which of these differences I retain when creating the mapping. Using identical reionization and cosmological parameters, and keeping all other input parameters at their default values from the GitHub packages, 21CMFAST version 1.2[2] and SIMFAST21 version 1.0 result in different reionization histories, as expected due to the different default bubble-finding algorithms. The motivation in this section is to demonstrate a method for mapping between the input parameters of two similar (but not identical) simulations. Using the default implementations, the output power spectra of the two simulations at a single fixed redshift are not comparable because the two simulations have reached different stages of reionization. Before making the mapping I choose input parameters of 21CMFAST which more closely matched the SIMFAST21 algorithm, so that the output power spectra are similar enough that making a mapping is meaningful, but not so similar that they give identical results. The following is a list of significant input parameters in 21CMFAST that I adjust from the default values:

1. FIND_BUBBLE_ALGORITHM $= 1$
2. ION_TVIR_MIN $= -1$, instead using ION_M_MIN
3. INHOMO_RECO $= 0$

Appendix A lists all parameters used in both simulations for repeatability. The most significant change from default was in the algorithm for finding ionised bubbles, setting FIND_BUBBLE_ALGORITHM $= 1$. Without making a judgement on which method is more realistic I used the SIMFAST21 algorithm of painting the entire sphere as ionised, rather than painting only the central pixel. I fix the minimum mass $M_{\rm min}$ for collapse using ION_M_MIN, rather than using the default 21CMFAST functionality of a fixed virial temperature $T_{\rm vir}$ using ION_TVIR_MIN. I also turn off calculations involving inhomogeneous recombinations by setting

[2] https://github.com/andreimesinger/21cmFAST

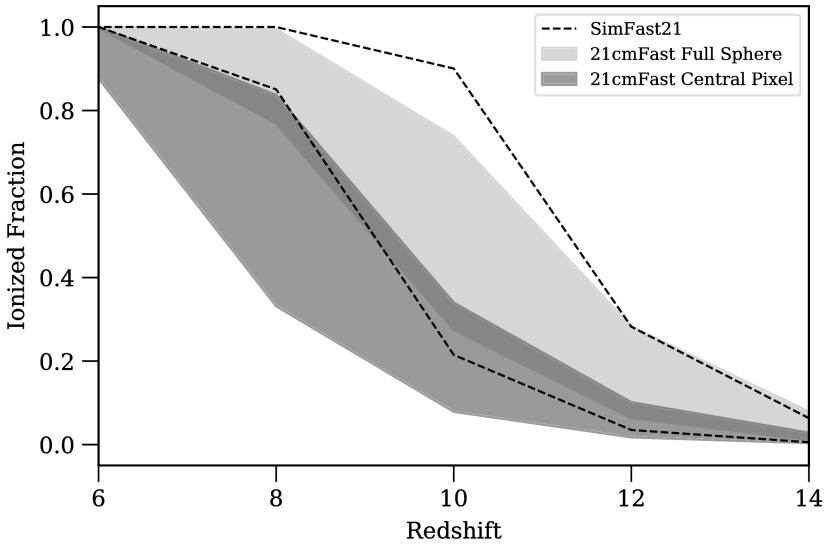

Figure 6.18: Ranges of reionization histories that result from SIMFAST21 and 21CMFAST, with $M_{\rm min}$ varying from $10^8 M_\odot$ to $10^9 M_\odot$. The region between the black dotted curves indicates the range of histories from SIMFAST21. The two coloured regions show the range of histories from 21CMFAST, both before (darker red) and after (lighter orange) matching the algorithms. The other reionization parameters are fixed at $\zeta_{\rm ion} = 30.0$ and $R_{\rm max} = 10.0$. The bubble-finding algorithm has a significant impact on the resulting reionization history, and even after matching algorithms there is a slight difference between SIMFAST21 and 21CMFAST.

INHOMO_RECO $= 0$, since the version of SIMFAST21 that I use does not have this option although later versions do (Hassan et al., 2016). Figure 6.18 shows the resulting ranges of reionization histories from a spread of minimum halo mass scenarios between $10^8 M_\odot - 10^9 M_\odot$. Each minimum mass scenario is averaged across five realisations. The histories are shown for SIMFAST21 (dotted) and for 21CMFAST with both bubble-finding algorithms: ionising the central pixel only (darker red region) and ionising the full sphere (lighter orange region) to match SIMFAST21. The only remaining major differences between the default SIMFAST21 simulation and the changed 21CMFAST simulation are in the specifics of implementation discussed above. Figure 6.18 shows that the differences in implementation still result in different reionization histories even after matching the bubble-finding algorithms, although the bubble-finding algorithm is the most dominant effect.

6.4.2 Using $x_{\rm HII}$ as input

I made the above changes to match the evolution of the global ionisation fraction in the simulations, so that I could compare power spectra at fixed redshift values. Instead of changing the default implementations, it would also be possible to account for the differing reionization histories by comparing power spectra at fixed ionisation fraction values (instead of fixed redshift values). This would require training an emulator to use the global ionisation fraction as an input, and would investigate how the shape of the power spectrum differs at a fixed point through the reionization process. I choose to match at fixed redshifts, to include the effect of differing reionization histories in the comparison.

6.4.3 Determining a mapping between simulations

Here I describe how the best emulator can be used to determine a mapping between the inputs of the modified 21CMFAST and the inputs of SIMFAST21. I use the same k-space restrictions as in Section 6.1.3, using only $0.1 \leq k \leq 2.0$ since the large scales are subject to foregrounds and the small scales are subject to shot noise from the finite simulation resolution. I also restrict the comparison to higher redshifts $z \geq 10$ for which the emulator exhibits higher prediction accuracy. I emphasise that this is a proof-of-concept method showing how to make a mapping between simulations solely using the output power spectra.

Figure 6.19 shows an example of one such mapping. I explain how to interpret the mapping here. Suppose a reference 21CMFAST simulation has already been run using the parameters specified by the white star: namely, a 21CMFAST simulation with parameters $M_{\rm min} = 3 \times 10^8 M_\odot$, $\zeta_{\rm ion} = 30.0$, $R_{\rm max} = 10$ Mpc. According to the mapping in Figure 6.19, any SIMFAST21 simulation using parameters within the orange contour will result in power spectra which are similar to the reference 21CMFAST spectra. I classify two simulations as similar if the mean squared error between their output power spectra is better than 30%. The orange contour thus shows the region of SIMFAST21 parameters which should be used, if the desired result is to exhibit similar power spectra to the reference 21CMFAST simulation. I generate the reference power spectra in Figure 6.19 by running five 21CMFAST

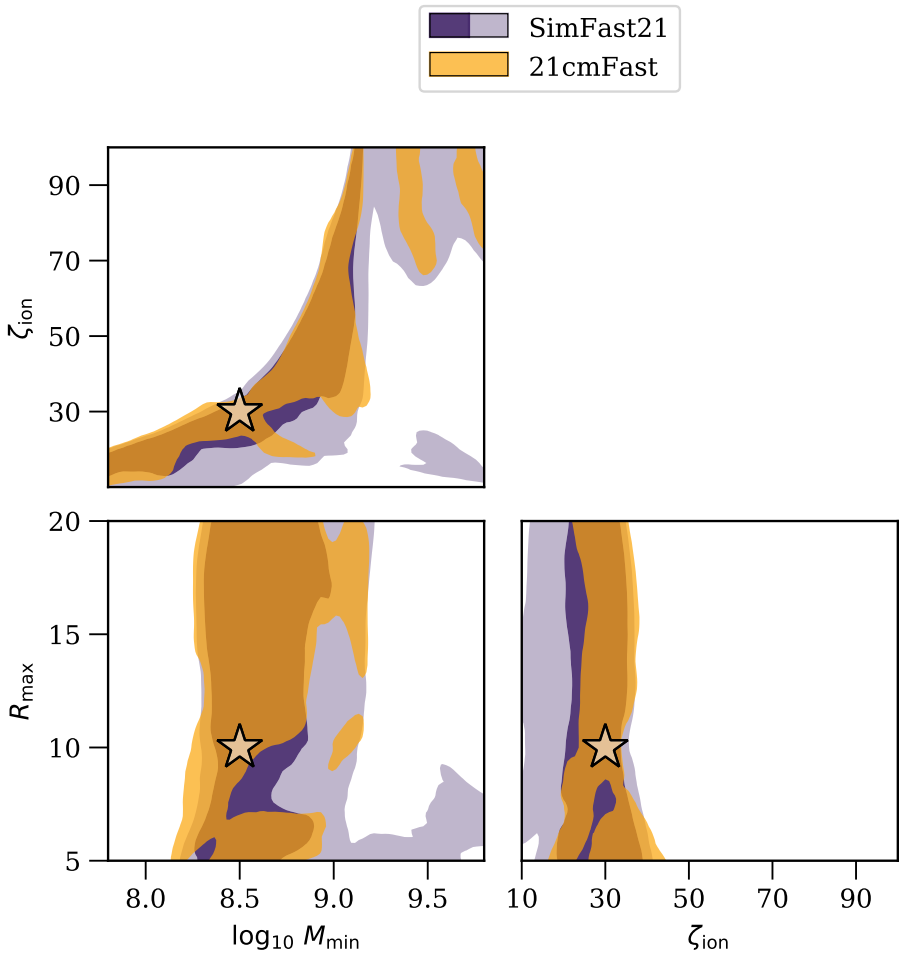

Figure 6.19: Mean squared error between emulated SIMFAST21 power spectra and measured power spectra from both simulations. The star indicates the fixed simulation parameters. The orange contour indicates the regions where emulated SIMFAST21 power spectra are within 30% MSE of the fully-simulated 21CMFAST power spectra. For comparison, the purple contours indicate the same regions for comparing emulated SIMFAST21 power spectra with fully-simulated SIMFAST21 power spectra, using 30% MSE (lighter contour) and 15% MSE (darker contour).

simulations, and taking the average to reduce the effect of sample variance. I refer to this type of figure as a similarity plot. Most importantly, if the orange contour does not overlap with the white star, then this indicates that SIMFAST21 and 21CMFAST result in different output power spectra for the same input parameters.

Two features of the orange contours are immediately apparent. First, the extended contours in the $R_{\rm max}$ direction. The $R_{\rm max}$ parameter is known to have little effect on the output power spectra for high redshifts (Mesinger et al., 2011). This is an inherent property of the power spectrum, regardless of which simulation is used. A second clear feature is the large curved contour in the $M_{\rm min}$-$\zeta_{\rm ion}$ parameter space of the top left panel in Figure 6.19. I investigated both features, to confirm whether they arise as an inherent property of the power spectrum itself, or if they arise from differences in the two simulations. To do this, I perform the same similarity analysis as above, but using SIMFAST21 itself as the reference simulation. The purple contours in Figure 6.19 then give the regions of SIMFAST21 parameters which result in similar power spectra to the reference SIMFAST21 simulation. The lighter purple contours use a MSE threshold of 30%. The darker purple contours use a stricter threshold of 15% MSE. The curved feature appears in both orange and purple contours, indicating that it is not due to a difference in the simulations. This curved degeneracy has been observed previously, see for example Greig and Mesinger (2015) and Schmit and Pritchard (2018). Note that I do not include a dark orange contour for the the stricter 15% MSE threshold because the power spectra for 21CMFAST differ from those of SIMFAST21 enough that no 21CMFAST contours are visible for an MSE threshold of 15%.

It is interesting to note that the substructure in the contours is not identical for the two simulations. The contours for SIMFAST21 are smooth and fairly contiguous, mostly forming one large region with smooth edges that surrounds the true white-star location. The contours for 21CMFAST, however, are patchy and segmented into multiple regions. To understand this, first recall that the emulator was trained on SIMFAST21 data. Any substructure in the true parameter space of the simulation's power spectrum would likely have been captured by the trained emulator.

This explains why the purple contours are smooth and contiguous: they are closer to measurements of the inherent properties of the 21cm power spectrum. In order to explain the discontiguous orange regions, recall that Figure 6.19 compares the power spectra by matching redshifts of the two simulations, and that the simulations show a difference in the reionization histories as show in Figure 6.18 earlier. Indeed, this is one of the main points in making these comparison plots – visualising the different behaviours of the two simulations. The observed substructure in the orange contours indicates that measuring the power spectrum at a different location in the parameter space can mimic the effect of measuring the power spectrum at a slightly earlier or later stage of reionization. The precise relationship that controls the appearance of these substructures is too complex to be interpreted by eye.

Figure 6.20 shows similarity plots for several other reference simulations, where the parameters for each reference simulation is again indicated by the location of the white star. I show the contours in the two-dimensional $M_{\rm min}$-$\zeta_{\rm ion}$ space, ignoring the less interesting $R_{\rm max}$ direction. I find that the orange contour does not always lie on top of the white star. This indicates that SIMFAST21 and 21CMFAST do not always result in similar output power spectra. I use the same contour levels as in Figure 6.19, namely 15% and 30% for the darker and lighter purple contours, and 30% for the 21CMFAST contours. Again, no 15% MSE contour is shown for 21CMFAST because the SIMFAST21 power spectra differ from the 21CMFAST power spectra by more than 15% everywhere.

For several of these scenarios, there is an offset between the orange contour and the white star. The offset is small near the canonical parameters in the central panels, but gets larger at lower $M_{\rm min}$ and higher $\zeta_{\rm ion}$. The most likely reason for this offset is the difference in the reionization histories. This offset would mean that the choice of using SIMFAST21 or using 21CMFAST would affect the outcome of parameter estimation methods, such as maximising χ^2 values (Shimabukuro and Semelin, 2017) or using MCMC methods (Schmit and Pritchard 2018 and Kern et al. 2017). Note that the two simulations in this comparison needn't share the same types of input parameters. For instance, it would be possible to generate

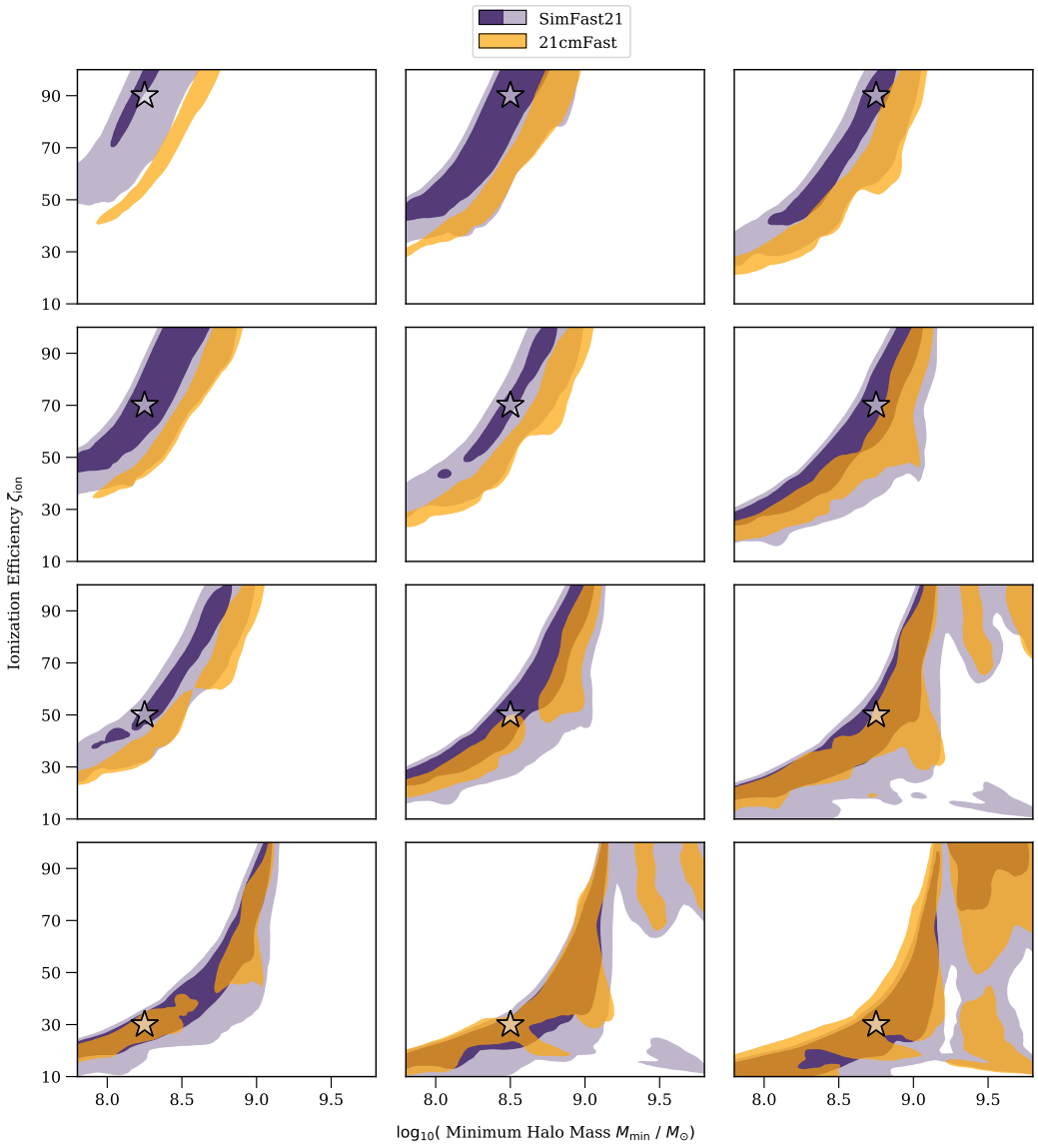

Figure 6.20: Similarity plots between emulated SIMFAST21 power spectra and fully-simulated power spectra from 21CMFAST (orange contours) and SIMFAST21 (purple contours). In each panel, the white star indicates the scenario parameters of the fully-simulated power spectra. The orange contour shows the regions in which emulated SIMFAST21 power spectra differ by less than 30% from the fully-simulated 21CMFAST power spectra. The lighter- and darker-purple contours show the equivalent regions for comparing emulated SIMFAST21 power spectra to fully-simulated SIMFAST21 power spectra, within 30% and 15% MSE respectively. An offset can be seen for several of these different scenarios.

the reference power spectra using a numerical radiative transfer simulation, and determine how its inputs map to SIMFAST21.

6.5 Conclusions

Fast modelling of the 21cm signal will become a significant problem in analysing the huge datasets from upcoming radio interferometry experiments. Ideally we would be able to compare numerical radiative transfer simulations with these data. Current numerical simulations are too slow to sample the input parameter space efficiently. Semi-numerical simulations are faster but can still only be used to constrain a small number of parameters. One potential solution to this problem is to replace current semi-numerical simulations with emulated models, reproducing the simulation outputs in a fraction of the original simulation time.

In this chapter I have trained and compared emulators using five different machine learning techniques. The two naive interpolation methods are not feasible as emulators, since they have either slow prediction times (linear interpolation model) or poor accuracy (nearest neighbour interpolation model). Of the three more sophisticated models, one model performs much better than the others: the multilayer perceptron. This trained model makes predictions of the outputs from 500 SIMFAST21 simulations to within 4% mean squared error averaged across all output points, reducing the modelling time from around 3000 hours to less than a second. The simulations include the effect of sampling noise from randomly-seeded density fields, but do not include other sources of experimental noise which would be needed if the emulators were to be used on observed data. If CPU training time is not a factor, then the accuracy of the sparse Gaussian processes regression or support vector machine models could potentially be improved with deeper hyperparameter searches. However, given their already relatively long prediction times and the accurate performance of the multilayer perceptron, these models are unlikely to give an improvement over the three-layer multilayer perceptron. There are several possible reasons why the MLP might perform better than the other methods. First, the MLP is exceptionally easy to implement with the SCIKIT-LEARN module.

Whereas other methods like SGPR required more complex approximations which required more in-depth hyper-parameter searching and led to many implementation issues. Second, the neural network is a more natural choice for the regression problem in this chapter than methods which were originally designed for classification problems. Using classification algorithms such as SVM or decision trees often leads to 'blocky' or step-like predictions. Due to the local-searching of these algorithms, such models often separate the parameter space into segments and make predictions within these segments. Finally, neural networks are extremely flexible in terms of their ability to represent complex functions: if the model appears to have poor predictions (underfitting the data), one can simply increase the size of the network by adding more layers or more neurons in each layer. For the problem in this chapter, even the default out-of-the-box model with a single hidden layer with 100 neurons had a decent prediction accuracy.

These emulators use redshift and k-scales as extra input dimensions. This makes the models more flexible but gives rise to less accurate emulation especially near the end of reionization at lower redshifts. I also use $\Delta T_{\rm b} = \frac{\delta T_{\rm b}}{\langle \delta T_{\rm b} \rangle} - 1$ as the target values of the emulators. This gives rise to sudden features in the power spectra near at the end of reionization and is harder to emulate than using $\delta T_{\rm b} - \langle \delta T_{\rm b} \rangle$. All simulated data ignores the effect of spin temperature calculations. Including the effect of spin temperature would likely have two effects. First, the simulations take significantly longer to run, meaning that the size of the training dataset would likely have to be reduced. Secondly, the relationship between parameters and power spectrum would likely become more complex. This would require comparing larger models with a wider range of hyperparameters.

I use the best emulator to determine a relationship between two different reionization algorithms, using SIMFAST21 and a version of 21CMFAST with non-default inputs. I find some noticeable offsets in which input parameters match the power spectra outputs of SIMFAST21 with those of 21CMFAST. I provide a graphical description of how this offset depends on location in parameter space, so that users could roughly determine which SIMFAST21 input parameters should be used if the

desired result is to match the 21cm power spectrum of an existing 21CMFAST simulation. Although the results in this chapter are for a version of 21CMFAST with non-default inputs, this method has potential for bridging between fast semi-numerical simulations and more accurate three-dimensional radiative transfer codes, such as C^2-RAY (Mellema et al., 2006a) and LICORICE (Semelin et al., 2017) (Kulkarni et al., 2016). However, Majumdar et al. (2014) noted that there can be a 25% difference between the power spectrum outputs of C^2-RAY and semi-numerical codes. Given this discrepancy, it is likely that mapping between numerical and semi-numerical simulations will be considerably more challenging and it may be necessary to emulate numerical codes directly using the same techniques in this chapter.

Chapter 7

Conclusions

The Epoch of Reionization (EoR) is a fascinating but relatively unexplored time in the Universe's history. We have yet to make precise measurements of when this process started, how long it lasted, and which types of sources contributed the most. The 21cm emission line from neutral hydrogen is an exciting probe for the EoR. The next generation of radio interferometers such as the Square Kilometre Array will for the first time observe three-dimensional maps of 21cm radiation during the EoR. The SKA will provide terabytes of raw data every second. Interpreting these data will require fast and accurate theoretical modelling methods. In this thesis I have presented three projects both for efficient modelling of the EoR and for analysing the results of EoR experiments.

The first step for efficient data analysis is often to compress and summarise the data before comparing with theoretical predictions. I developed an efficient code for calculating one such summary compression statistic: higher-order clustering functions. I present this publicly available code in Chapter 3. I tested the code by calculating three-point correlation function for an example distribution of points-on-spheres data, the theoretical correlation function solution for which was kindly provided by Lorne Whiteway (2018). The code reproduced the theoretical values over a wide range of different model parameters such as the size and number density of the spheres. I used the code to calculate the three-point correlation function of equilateral triangles for ionisation fraction data from semi-numerical simulations. The equilateral three-point correlation function shows a clear evolu-

tion with redshift, likely tracing the mean size of ionised bubbles over time. The effects of two reionization parameters on the three-point statistics was also investigated. The ionisation efficiency affected the growth speed of ionised bubbles, and gave rise to distinctive change in the position of the three-point correlation peak. The minimum halo mass parameter also affected the correlation amplitude, in particular changing the amplitude of the low-radius peak. The dependencies indicate that the three-point code could be a useful addition to analysis of real-space data from experiments. It could be included in the likelihoods of MCMC analysis as an attempt to break degeneracies between parameters.

In Chapter 4, I use my code for the three-point correlation function to see whether the 21cm signal encodes information about the morphology and history of the Epoch of Reionization. I use machine learning techniques to recover the typical bubble size and global ionisation fraction from the measured 3PCF outputs from semi-numerical simulations. In general, my models recover these properties with similar accuracies to other existing methods. My models recover the typical bubble size from the 3PCF of either ionisation fraction 3PCF data (with median RMSE value of 19.9%) or 21cm differential brightness temperature 3PCF data (with median RMSE 29.3%). My models for the global ionisation fraction have better performance, recovering the global ionisation fraction from ionisation fraction 3PCF data with median RMSE of 3.6%, or from the brightness temperature with median RMSE of 16.0%. This work presents the first attempt to predict these physical EoR properties using the three-point correlation function and machine learning techniques. I have made my code publicly available on GitHub to help the community perform similar analyses in the future.

Efficient modelling of the 21cm signal will be a crucial limiting factor in analysis of 21cm data. Theoretical models for the EoR are of three main types: analytical, numerical simulations, and semi-numerical simulations. Analytical models give less detailed predictions than simulations but are much faster. In Chapter 5, I developed a middle-ground between analytic and semi-numerical simulations: fitting an analytical model to the outputs from semi-numerical simulations. This model made

predictions for the second-order clustering in Fourier-space, but could also be used to make predictions for the three-point correlation function in Chapter 3. My final halo model was reached through a series of increasingly complex steps. Starting with a very simple approximation of identical ionised bubbles, I added more complexity to the model by fitting results from SIMFAST21. At each stage I checked that the predictions match the measured power spectrum of mock generated data. These stages were using differently-sized bubbles, improving the bubble size distribution to a power law, and adding clustering to the bubble centres. For the first two stages the fitted model matches the mock generated data spectra. Clustering caused a mismatch between the theory and recreated data, which was resolved for all but the latest redshifts by including a fitted suppression function measured from SIMFAST21 simulations. The amplitude of the resulting halo model spectra showed better agreement with the measured values. A final step of comparing the halo model predictions to fully simulated power spectra was attempted. The final model gives approximate predictions for the amplitude of the ionisation fraction spectra for the earliest parts of reionization.

The model in Chapter 5 would need significant improvement before being useful as a full replacement for semi-numerical simulations. In particular it would need to be extended so that theoretical predictions could be made for any reionization scenario specified by a particular set of reionization parameters such as ionisation efficiency $\zeta_{\rm ion}$ and minimum halo mass $M_{\rm min}$. In Chapter 6, I used machine learning techniques to replicate the predictions semi-numerical simulations. The best emulator reproduced the SIMFAST21 power spectra outputs 10^8 times faster than the simulation with only 4% mean squared error averaged across all redshifts and input parameters. This emulator could safely be used as a replacement for the semi-numerical simulation with only minimal loss in prediction accuracy. I analysed the prediction speeds and accuracies of emulators using different machine learning techniques. The best emulator used a multilayer perceptron with three hidden layers. The other machine learning techniques (interpolation, Gaussian processes regression, and support vector machine) all had slower prediction times or worse

prediction accuracy than the multilayer perceptron. At the end of Chapter 6 I use my best emulator to compare the outputs between SIMFAST21 and 21CMFAST. These simulations have many similarities, but a few key differences give rise to a small offset in the simulation outputs which depends on the simulation input parameters.

Developments in surrogate modelling will almost certainly have a great impact on the wider community in the coming decades. The ability to replace expensive calculations, accurately and almost instantaneously, is a dramatic change in the status quo. With tools and methods like those in this thesis, we can face the formidable datasets from future EoR experiments with a fighting chance of getting answers to our unsolved questions. With enough international collaborative effort and continued interest in Astrophysics, we will soon finally get a glimpse of reionization in action and satisfy a part of our two-thousand-year-old curiosity.

Appendix A

Simulation parameters

I list all relevant user-changeable parameters used for all 21CMFAST and SIMFAST21 simulations in this paper. For further descriptions of these parameters see Mesinger et al. (2011) and Santos et al. (2010). I exclude parameters relating to spin temperature calculations since I did not use this functionality.

A.1 Cosmology

Parameter	Value
σ_8	0.810
Hubble h	0.710
Ω_M	0.270
Ω_Λ	0.730
Ω_b	0.046
Ω_n	0.0
Ω_k	0.0
Ω_R	0.0
Ω_{tot}	1.0
Y_{He}	0.245
n_s	0.960
Sheth-Tormen b	0.34
Sheth-Tormen c	0.81
Helium II z_{reion}	3
Maximum Redshift	17.00
Minimum Redshift	8.00
Redshift Step	1.50
Simulation Length	500.00
Star Formation Rate	0.025
Velocity Component	3
Critical Overdensity	1.680

A.2 SIMFAST21

Parameter Name	Value
USE_CAMB_MATTERPOWER	False
USE_FCOLL	True
HALO_RMAX	40
HALO_MMIN	Various
ION_EFF	Various
BUBBLE_RMAX	Various
USE_LYA_XRAYS	False

A.3 21CMFAST

Parameter Name	Value
ION_M_MIN	Various
ION_TVIR_MIN	-1 (off)
HII_EFF_FACTOR	Various
EFF_FACTOR_PL_INDEX	0
R_BUBBLE_MAX	Various

A.4 Other 21CMFAST

Parameter Name	Value
P_CUTOFF	0
M_WDM	2
G_X	1.5
INHOMO_RECO	0
ALPHA_UVB	5
T_STAR	0.5
EVOLVE_DENSITY_LINEARLY	0
SMOOTH_EVOLVED_DENSITY_FIELD	1
R_SMOOTH_DENSITY	0.2
SECOND_ORDER_LPT_CORRECTIONS	0
HII_ROUND_ERR	1e-3
FIND_BUBBLE_ALGORITHM	1
R_BUBBLE_MIN	L_FACTOR*1
USE_HALO_FIELD	0
N_POISSON	-1
T_USE_VELOCITIES	1
MAX_DVDR	0.2
DIMENSIONAL_T_POWER_SPEC	0
DELTA_R_FACTOR	1.1
DELTA_R_HII_FACTOR	1.1
R_OVERLAP_FACTOR	1.0
DELTA_CRIT_MODE	1
HALO_FILTER	0
HII_FILTER	1
OPTIMIZE	0
OPTIMIZE_MIN_MASS	1e11
SIZE_RANDOM_SEED	-23456789
LOS_RANDOM_SEED	-123456789
USE_TS_IN_21CM	0
CLUMPING_FACTOR	50
POP	2
POP2_ION	4361

Bibliography

Abadi, Martín, Barham, Paul, Chen, Jianmin, et al. (2016), "TensorFlow : A System for Large-Scale Machine Learning". Proc 12th USENIX conference on Operating Systems Design and Implementation.

Abbott, T. M. C., Abdalla, F. B., Annis, J., et al. (2017), "Dark Energy Survey Year 1 Results: A Precise H0 Measurement from DES Y1, BAO, and D/H Data".

Agarwal, Shankar, Abdalla, Filipe B., Feldman, Hume A., et al. (2014), "PkANN-II. A non-linear matter power spectrum interpolator developed using artificial neural networks". Monthly Notices of the Royal Astronomical Society, 439(2):2102–2121.

Ali, Zaki S., Parsons, Aaron R., Zheng, Haoxuan, et al. (2015), "Paper-64 constraints on reionization: the 21cm power spectrum at z = 8.4". Astrophysical Journal, 809(1):61.

Alpher, R. A., Bethe, H., and Gamow, G. (1948), "The Origin of Chemical Elements". Physical Review, 73(7):803–804.

Alvarez, Marcelo A. and Abel, Tom (2012), "The effect of absorption systems on cosmic reionization". Astrophysical Journal, 747(2).

Bacon, David J, Battye, Richard A, Bull, Philip, et al. (2018), "Cosmology with Phase 1 of the Square Kilometre Array". Publications of the Astronomical Society of Australia.

Banados, Eduardo, Venemans, Bram P, Mazzucchelli, Chiara, et al. (2018), "An 800-million-solar-mass black hole in a significantly neutral Universe at redshift 7.5". Technical report.

Barber, C. Bradford, Dobkin, David P., and Huhdanpaa, Hannu (1996), "The quick-

hull algorithm for convex hulls". ACM Transactions on Mathematical Software, 22(4):469–483.

Barkana, Rennan and Loeb, Abraham (2001), "In the beginning: The first sources of light and the reionization of the universe". Physics Report, 349(2):125–238.

Beardsley, A. P., Hazelton, B. J., Sullivan, I. S., et al. (2016), "First Season MWA EoR Power Spectrum Results at Redshift 7".

Becker, George D., Rauch, Michael, and Sargent, Wallace L. W. (2007), "The Evolution of Optical Depth in the Lyα Forest: Evidence Against Reionization at z $˜$ 6". The Astrophysical Journal, 662(1):72–93.

Becker, Robert H., Fan, Xiaohui, White, Richard L., et al. (2001), "Evidence for Reionization at z ˜ 6: Detection of a Gunn-Peterson Trough in a z=6.28 Quasar". The Astronomical Journal, 122(6):2850–2857.

Bernal, Jose Luis, Verde, Licia, and Riess, Adam G. (2016), "The trouble with H_0". Technical report, ICC, University of Barcelona, IEEC-UB, Mart´ı i Franqu‘es.

Bernardi, G., Zwart, J. T. L., Price, D., et al. (2016), "Bayesian constraints on the global 21-cm signal from the Cosmic Dawn".

Bolejko, Krzysztof (2018), "Emerging spatial curvature can resolve the tension between high-redshift CMB and low-redshift distance ladder measurements of the Hubble constant". Physical Review D, 97(10):103529.

Bolton, J. S., Haehnelt, M. G., Warren, S. J., et al. (2011), "How neutral is the intergalactic medium surrounding the redshift z = 7.085 quasar ULAS J1120+0641?". Monthly Notices of the Royal Astronomical Society: Letters, 416(1):L70–L74.

Bowman, Judd D., Rogers, Alan E. E., Monsalve, Raul A., et al. (2018), "An absorption profile centred at 78 megahertz in the sky-averaged spectrum".

Bromm, Volker, Kudritzki, Rolf P, and Loeb, Abraham (2001), "GENERIC SPECTRUM AND IONIZATION EFFICIENCY OF A HEAVY INITIAL MASS FUNCTION FOR THE FIRST STARS". Technical report.

Bullock, James S., Kolatt, Tsafrir S., Sigad, Yair, et al. (2001), "Profiles of dark

haloes: Evolution, scatter and environment". Monthly Notices of the Royal Astronomical Society, 321(3):559–575.

Burles, Scott, Nollett, Kenneth M., and Turner, Michael S. (1999), "Big-Bang Nucleosynthesis: Linking Inner Space and Outer Space". arXiv eprints.

Canetti, Laurent, Drewes, Marco, and Shaposhnikov, Mikhail (2012), "Matter and Antimatter in the Universe". New Journal of Physics, 14(9).

Cohen, Aviad, Fialkov, Anastasia, Barkana, Rennan, et al. (2016), "Charting the Parameter Space of the Global 21-cm Signal".

Collister, Adrian A. and Lahav, Ofer (2003), "ANNz: estimating photometric redshifts using artificial neural networks". The Publications of the Astronomical Society of the Pacific, 116(818):345–351.

Cooray, Asantha and Sheth, Ravi (2002), "Halo models of large scale structure". Physics Report, 372(1):1–129.

Das, Arpan, Mesinger, Andrei, Pallottini, Andrea, et al. (2017), "High-mass X-ray binaries and the cosmic 21-cm signal: impact of host galaxy absorption". Monthly Notices of the Royal Astronomical Society, 469(1):1166–1174.

Datta, A., Bowman, J. D., and Carilli, C. L. (2010), "Bright source subtraction requirements for redshifted 21cm measurements". Astrophysical Journal, 724(1):526–538.

De Bernardis, P., Ade, P. A.R., Bock, J. J., et al. (2000), "A flat Universe from high-resolution maps of the cosmic microwave background radiation". Nature, 404(6781):955–959.

Dijkstra, Mark, Mesinger, Andrei, and Stuart Wyithe, J B (2011), "The detectability of Lyα emission from galaxies during the epoch of reionization". Mon. Not. R. Astron. Soc, 414:2139–2147.

Dodelson, Scott (2003), Modern Cosmology. 2003. Academic Press.

Elbers, Willem and Van De Weygaert, Rien (2018), "Persistent topology of the reionisation bubble network. I: Formalism & Phenomenology". Technical report.

et al DeBoer, David R. (2017), "HYDROGEN EPOCH OF REIONIZATION

ARRAY (HERA)". Publications of the Astronomical Society of the Pacific, 129(974).

Faber, S. M. and Gallagher, J. S. (1979), "Masses and Mass-To-Light Ratios of Galaxies". Annual Review of Astronomy and Astrophysics, 17(1):135–187.

Faessler, Amand, Hodak, Rastislav, Kovalenko, Sergey, et al. (2016), "Can one measure the Cosmic Neutrino Background?". International Journal of Modern Physics, 26(1n02).

Fialkov, Anastasia, Barkana, Rennan, and Cohen, Aviad (2018), "Constraining Baryon–Dark Matter Scattering with the Cosmic Dawn 21-cm Signal".

Fialkov, Anastasia, Barkana, Rennan, and Visbal, Eli (2014), "The observable signature of late heating of the Universe during cosmic reionization". Nature, 506(7487):197–199.

Fialkov, Anastasia, Cohen, Aviad, Barkana, Rennan, et al. (2015), "Constraining the redshifted 21-cm signal with the unresolved soft X-ray background". Technical report.

Fialkov, Anastasia and Loeb, Abraham (2013), "The 21-cm Signal from the Cosmological Epoch of Recombination".

Field, George (1958), "Excitation of the Hydrogen 21-CM Line". Proceedings of the IRE, 46(1):240–250.

Friedman, A. (1922), "Über die Krümmung des Raumes". Zeitschrift für Physik, 10(1):377–386.

Furlanetto, Steven R, Oh, S. Peng, Briggs, Frank H, et al. (2006), "Cosmology at Low Frequencies: The 21 cm Transition and the High-Redshift Universe". Physics Reports, 433(4-6):181–301.

Furlanetto, Steven R., Zaldarriaga, Matias, and Hernquist, Lars (2004), "The Growth of HII Regions During Reionization". The Astrophysical Journal, 613(1):1–15.

Gal, Yarin and Ghahramani, Zoubin (2016), "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning Zoubin Ghahramani". Technical report.

George, E. M., Reichardt, C. L., Aird, K. A., et al. (2014), "A measurement of secondary cosmic microwave background anisotropies from the 2500-square-degree SPT-SZ survey".

Ghosh, Abhik, Koopmans, Léon V E, Chapman, E, et al. (2015), "A Bayesian analysis of redshifted 21-cm HI signal and foregrounds : Simulations for LOFAR". Monthly Notices of the Royal Astronomical Society, 000(June).

Gillet, Nicolas, Mesinger, Andrei, Greig, Bradley, et al. (2018), "Deep learning from 21-cm images of the Cosmic Dawn". arXiv eprints.

Giri, Sambit K., D'Aloisio, Anson, Mellema, Garrelt, et al. (2018a), "Position-dependent power spectra of the 21-cm signal from the epoch of reionization".

Giri, Sambit K, Mellema, Garrelt, Dixon, Keri L, et al. (2018b), "Bubble size statistics during reionization from 21-cm tomography". MNRAS, 473:2949–2964.

Glorot, Xavier and Bengio, Yoshua (2010), "Understanding the difficulty of training deep feedforward neural networks". Technical report.

Gnedin, Nickolay Y (2014), "COSMIC REIONIZATION ON COMPUTERS. I. DESIGN AND CALIBRATION OF SIMULATIONS". The Astrophysical Journal, 793(12pp):29.

Gorce, Adélie and Pritchard, Jonathan R. (2019), "Studying the morphology of reionisation with the triangle correlation function of phases".

Greig, Bradley and Mesinger, Andrei (2015), "21CMMC: an MCMC analysis tool enabling astrophysical parameter studies of the cosmic 21 cm signal". Monthly Notices of the Royal Astronomical Society, 000(0000):0–0.

Greig, Bradley and Mesinger, Andrei (2017a), "Simultaneously constraining the astrophysics of reionisation and the epoch of heating with 21CMMC". Proceedings of the International Astronomical Union, 12(S333):18–21.

Greig, Bradley and Mesinger, Andrei (2017b), "The global history of reionization". Monthly Notices of the Royal Astronomical Society, 465(4):4838–4852.

Greig, Bradley, Mesinger, Andrei, and Pober, Jonathan C. (2016), "Constraints on the temperature of the intergalactic medium at z = 8.4 with 21-cm observations". Monthly Notices of the Royal Astronomical Society, 455(4):4295–4300.

Gunn, James E. and Peterson, Bruce A. (1965), "On the Density of Neutral Hydrogen in Intergalactic Space.". The Astrophysical Journal, 142:1633.

Guth, Alan H. (1981), "Inflationary universe: A possible solution to the horizon and flatness problems". Physical Review D, 23(2):347–356.

Hassan, Sultan, Davé, Romeel, Finlator, Kristian, et al. (2016), "Simulating the 21 cm signal from reionization including non-linear ionizations and inhomogeneous recombinations". Monthly Notices of the Royal Astronomical Society, 457(2):1550–1567.

Hassan, Sultan, Davé, Romeel, Finlator, Kristian, et al. (2017), "Epoch of reionization 21 cm forecasting from MCMC-constrained semi-numerical models". Monthly Notices of the Royal Astronomical Society, 468(1):122–139.

Hinton, Geoorey E and Van Camp, Drew (1993), "Keeping Neural Networks Simple by Minimizing the Description Length of the Weights". Technical report.

Hoffmann, Kai, Mao, Yi, Xu, Jiachuan, et al. (2018), "Signatures of Cosmic Reionization on the 21cm 2- and 3-point Correlation Function I: Quadratic Bias Modeling".

Hubble, E (1929), "A RELATION BETWEEN DISTANCE AND RADIAL VELOCITY AMONG EXTRA-GALACTIC NEBULAE.". Proceedings of the National Academy of Sciences of the United States of America, 15(3):168–73.

Hultman Kramer, Roban, Haiman, Zoltán, and Peng, S OH (2006), "FEEDBACK FROM CLUSTERED SOURCES DURING REIONIZATION". Technical report.

Hutter, Anne (2018), "The accuracy of seminumerical reionization models in comparison with radiative transfer simulations". Monthly Notices of the Royal Astronomical Society, 477(2):1549–1566.

Ichikawa, Kazuhide, Barkana, Rennan, Iliev, Ilian T, et al. (2009), "Measuring the History of Cosmic Reionization using the 21-cm PDF from Simulations". Technical report.

Iliev, Ilian T., Mellema, Garrelt, Pen, Ue-Li, et al. (2005), "Simulating Cosmic Reionization at Large Scales I: the Geometry of Reionization".

Jennings, William D., Watkinson, Catherine A., Abdalla, Filipe B., et al. (2018), "Evaluating machine learning techniques for predicting power spectra from reionization simulations". Monthly Notices of the Royal Astronomical Society, 483(3):2907–2922.

Jones, Eric, Oliphant, Travis, Peterson, Pearu, et al. (2007), "SciPy: Open source scientific tools for Python".

Kakiichi, Koki, Dijkstra, Mark, Ciardi, Benedetta, et al. (2015), "Lyα-Emitting Galaxies as a Probe of Reionization: Large-Scale Bubble Morphology and Small-Scale Absorbers". Technical Report 23.

Kakiichi, Koki, Majumdar, Suman, Mellema, Garrelt, et al. (2017), "Recovering the HII region size statistics from 21-cm tomography". Monthly Notices of the Royal Astronomical Society, 471(2):1936–1954.

Kaurov, Alexander A and Gnedin, Nickolay Y (2018), "RECOMBINATION CLUMPING FACTOR DURING COSMIC REIONIZATION". Technical report.

Kern, Nicholas S., Liu, Adrian, Parsons, Aaron R., et al. (2017), "Emulating Simulations of Cosmic Dawn for 21cm Power Spectrum Constraints on Cosmology, Reionization, and X-ray Heating". The Astrophysical Journal, 848(1).

Kingma, Diederik P. and Ba, Jimmy (2014), "Adam: A Method for Stochastic Optimization". In 3rd International Conference for Learning Representations.

Kohn, Saul A., Aguirre, James E., La Plante, Paul, et al. (2018), "The HERA-19 Commissioning Array: Direction Dependent Effects".

Kravtsov, Andrey V., Klypin, Anatoly A., and Khokhlov, Alexei M. (1997), "Adaptive Refinement Tree - a new high-resolution N-body code for cosmological simulations".

Kulkarni, Girish, Choudhury, Tirthankar Roy, Puchwein, Ewald, et al. (2016), "Models of the cosmological 21 cm signal from the epoch of reionization calibrated with Ly A and CMB data". Monthly Notices of the Royal Astronomical Society, 463(3):2583–2599.

Lahav, O. (1995), "Galaxy classification by human eyes and by artificial neural networks". Astrophysical Letters and Communications, 31:73.

Landy, Stephen D. and Szalay, Alexander S. (1993), "Bias and variance of angular correlation functions". The Astrophysical Journal, 412:64.

Liddle, Andrew R. and Leach, Samuel M. (2003), "How long before the end of inflation were observable perturbations produced?". Technical Report 10, Astronomy Centre, University of Sussex.

Lin, Yin, Oh, S. Peng, Furlanetto, Steven R., et al. (2015), "The Distribution of Bubble Sizes During Reionization".

Liu, Adrian, Pritchard, Jonathan R., Allison, Rupert, et al. (2016), "Eliminating the optical depth nuisance from the CMB with 21 cm cosmology". Physical Review D, 93(4).

Lochner, Michelle, McEwen, Jason D., Peiris, Hiranya V., et al. (2016), "Photometric Supernova Classification With Machine Learning". The Astrophysical Journal Supplement Series, 225(2).

Loeb, Abraham and Furlanetto, Steven R. (2013), The first galaxies in the universe.

Lorne Whiteway (2018), "Private Correspondence".

Lupton, Robert H., Gunn, James E., and Szalay, Alexander S. (1999), "A Modified Magnitude System that Produces Well-Behaved Magnitudes, Colors, and Errors Even for Low Signal-to-Noise Ratio Measurements". The Astronomical Journal, 118(3):1406–1410.

Macaulay, E., Nichol, R. C., Bacon, D., et al. (2018), "First Cosmological Results using Type Ia Supernovae from the Dark Energy Survey: Measurement of the Hubble Constant". Monthly Notices of the Royal Astronomical Society.

Majumdar, Suman, Mellema, Garrelt, Datta, Kanan K., et al. (2014), "On the use of seminumerical simulations in predicting the 21-cm signal from the epoch of reionization". Monthly Notices of the Royal Astronomical Society, 443(4):2843–2861.

Majumdar, Suman, Pritchard, Jonathan R., Mondal, Rajesh, et al. (2018), "Quantifying the non-Gaussianity in the EoR 21-cm signal through bispectrum". Monthly Notices of the Royal Astronomical Society, 476(3):4007–4024.

Mao, Yi, Tegmark, Max, McQuinn, Matthew, et al. (2008), "How accurately can 21 cm tomography constrain cosmology?".

Mather, J. C., Cheng, E. S., Cottingham, D. A., et al. (1994), "Measurement of the cosmic microwave background spectrum by the COBE FIRAS instrument". The Astrophysical Journal, 420:439.

Matheron, G. (Georges) (1974), Random sets and integral geometry. Wiley.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979), "Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". Technometrics, 21(2):239–245.

Mcquinn, Matthew, Lidz, Adam, Zahn, Oliver, et al. (2006), "The Morphology of HII Regions during Reionization". Technical Report 1.

Mellema, Garrelt, Iliev, Ilian T., Alvarez, Marcelo A., et al. (2006a), "C2-ray: A new method for photon-conserving transport of ionizing radiation". New Astronomy, 11(5):374–395.

Mellema, Garrelt, Iliev, Ilian T, Pen, Ue-Li, et al. (2006b), "Simulating Cosmic Reionization at Large Scales II: the 21-cm Emission Features and Statistical Signals". Technical Report 1.

Mellema, Garrelt, Koopmans, León, Abdalla, Filipe, et al. (2012), "Reionization and the Cosmic Dawn with the Square Kilometre Array". Experimental Astronomy, 36(1-2).

Mesinger, Andrei (2018), "Was reionization complete by z 5-6?". Technical Report 0000.

Mesinger, Andrei and Furlanetto, Steven (2007), "Efficient Simulations of Early Structure Formation and Reionization". The Astrophysical Journal, 669(2):663–675.

Mesinger, Andrei, Furlanetto, Steven, and Cen, Renyue (2011), "21cmfast: A fast, seminumerical simulation of the high-redshift 21-cm signal". Monthly Notices of the Royal Astronomical Society, 411(2):955–972.

Mitra, Sourav, Choudhury, T. Roy, and Ferrara, Andrea (2015), "Cosmic reionization after Planck".

Monsalve, Raul A., Fialkov, Anastasia, Bowman, Judd D., et al. (2019), "Results from EDGES High-Band. III. New Constraints on Parameters of the Early Universe". The Astrophysical Journal, 875(1):67.

Monsalve, Raul A., Rogers, Alan E. E., Bowman, Judd D., et al. (2017), "Results from EDGES High-band. I. Constraints on Phenomenological Models for the Global 21 cm Signal". The Astrophysical Journal, 847(1):64.

Mörtsell, Edvard and Dhawan, Suhail (2018), "Does the Hubble constant tension call for new physics?".

Muller, Mervin E. and E., Mervin (1959), "A note on a method for generating points uniformly on n-dimensional spheres". Communications of the ACM, 2(4):19–20.

Murray, Steven, Power, Chris, and Robotham, Aaron (2013), "HMFcalc: An Online Tool for Calculating Dark Matter Halo Mass Functions". Astronomy and Computing, 3:23–24.

NASA Goddard Space Flight Center (2019), "CMB Spectrum".

Navarro, Julio F., Frenk, Carlos S., and White, Simon D. M. (1996), "A Universal Density Profile from Hierarchical Clustering". The Astrophysical Journal, 490(2):493–508.

Ocvirk, Pierre, Gillet, Nicolas, Shapiro, Paul R., et al. (2015), "Cosmic Dawn (CoDa): the First Radiation-Hydrodynamics Simulation of Reionization and Galaxy Formation in the Local Universe".

Oesch, P A, Brammer, G, Van Dokkum, P G, et al. (2016), "A REMARKABLY LUMINOUS GALAXY AT Z=11.1 MEASURED WITH HUBBLE SPACE TELESCOPE GRISM SPECTROSCOPY".

O'Luanaigh, Cian (2015), "First images of collisions at 13 TeV".

Paciga, Gregory, Albert, Joshua G., Bandura, Kevin, et al. (2013), "A simulation calibrated limit on the HI power spectrum from the GMRT Epoch of Reionization experiment".

Park, Jaehong, Mesinger, Andrei, Greig, † Bradley, et al. (2018), "Inferring the astrophysics of reionization and cosmic dawn from galaxy luminosity functions and the 21-cm signal". Technical report.

Parsons, Aaron R., Backer, Donald C., Bradley, Richard F., et al. (2009), "The Precision Array for Probing the Epoch of Reionization: 8 Station Results".

Patil, A. H., Yatawatta, S., Koopmans, L. V. E., et al. (2017), "Upper limits on the 21-cm Epoch of Reionization power spectrum from one night with LOFAR". The Astrophysical Journal, 838(1).

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, et al. (2011), "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research, 12(Oct):2825–2830.

Penzias, A. A. and Wilson, R. W. (1965), "A Measurement of Excess Antenna Temperature at 4080 Mc/s.". The Astrophysical Journal, 142:419.

Perlmutter, S., Aldering, G., Goldhaber, G., et al. (1998), "Measurements of Omega and Lambda from 42 High-Redshift Supernovae". The Astrophysical Journal, 517(2):565 –586.

Planck, M. (1899), "Naturlische Masseinheiten". Der Koniglich Preussischen Akademie Der Wissenschaften, page p. 479.

Planck Collaboration (2016), "Planck intermediate results. XLVII. Planck constraints on reionization history". Astronomy & Astrophysics, 596.

Planck Collaboration (2018), "Planck 2018 results. VI. Cosmological parameters". arXiv eprints.

Pober, Jonathan C., Greig, Bradley, and Mesinger, Andrei (2016), "Upper limits on the 21 cm power spectrum at z = 5.9 from quasar absorption line spectroscopy". Monthly Notices of the Royal Astronomical Society: Letters, 463(1):L56–L60.

Press, William H. and Schechter, Paul (1974), "Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation". The Astrophysical Journal, 187:425.

Pritchard, Jonathan R. and Loeb, Abraham (2008), "Evolution of the 21 cm signal throughout cosmic history". Physical Review D - Particles, Fields, Gravitation and Cosmology, 78(10).

Pritchard, Jonathan R. and Loeb, Abraham (2012), "21 cm cosmology in the 21st century".

Pritchard, Jonathan R and Pierpaoli, Elena (2008), "Constraining massive neutrinos using cosmological 21 cm observations". Technical report.

Rasmussen, C. E. and Williams, C. K. I. (2006), Gaussian Processes for Machine Learning. The MIT Press.

Razoumov, Alexei O. and Sommer-Larsen, Jesper (2006), "Escape of Ionizing Radiation from Star-forming Regions in Young Galaxies". The Astrophysical Journal, 651(2):L89–L92.

Riess, Adam G., Casertano, Stefano, Yuan, Wenlong, et al. (2018), "Milky Way Cepheid Standards for Measuring Cosmic Distances and Application to Gaia DR2: Implications for the Hubble Constant". The Astrophysical Journal, 861(2).

Riess, Adam G., Filippenko, Alexei V., Challis, Peter, et al. (1998), "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant". Astronomical Journal, 116(3):1009–1038.

Robertson, Brant E., Ellis, Richard S., Furlanetto, Steven R., et al. (2015), "No Title". The Astrophysical Journal, 802(2):L19.

Ross, Hannah E., Dixon, Keri L., Iliev, Ilian T., et al. (2016), "Simulating the Impact of X-ray Heating during the Cosmic Dawn".

Rumelhart, David E., Hinton, Geoffrey E., and Williams, Ronald J. (1986), "Learning representations by back-propagating errors". Nature, 323:533.

Sadeh, I., Abdalla, F. B., and Lahav, O. (2016), "ANNZ2: Photometric redshift and probability distribution function estimation using machine learning". Publications of the Astronomical Society of the Pacific, 128(968).

Santos, M. G., Ferramacho, L., Silva, M. B., et al. (2010), "Fast large volume simulations of the 21-cm signal from the reionization and pre-reionization epochs". Monthly Notices of the Royal Astronomical Society, 406(4):2421–2432.

Savitzky, Abraham. and Golay, M. J. E. (1964), "Smoothing and Differentiation of Data by Simplified Least Squares Procedures.". Analytical Chemistry, 36(8):1627–1639.

Schmit, Claude J and Pritchard, Jonathan R (2018), "Emulation of reionization

simulations for Bayesian inference of astrophysics parameters using neural networks". Monthly Notices of the Royal Astronomical Society, 475(1):1213–1223.

Seljak, Uros (2000), "Analytic model for galaxy and dark matter clustering". Monthly Notices of the Royal Astronomical Society, 318(1):203–213.

Semelin, B., Combes, F., and Baek, S. (2007), "Lyman-alpha radiative transfer during the epoch of reionization: contribution to 21-cm signal fluctuations". Astronomy & Astrophysics, 474(2):365–374.

Semelin, Benoit (2015), "Detailed modelling of the 21-cm Forest".

Semelin, Benoit, Eames, Evan, Bolgar, Florian, et al. (2017), "21SSD: a public database of simulated 21-cm signals from the epoch of reionization". Monthly Notices of the Royal Astronomical Society, 472(4):4508–4520.

Sheth, Ravi K., Mo, H. J., and Tormen, Giuseppe (2001), "Ellipsoidal collapse and an improved model for the number and spatial distribution of dark matter haloes". Monthly Notices of the Royal Astronomical Society, 323(1):1–12.

Shimabukuro, Hayato and Semelin, Benoit (2017), "Analysing the 21cm signal from the Epoch of Reionization with artificial neural networks". Monthly Notices of the Royal Astronomical Society, 468(4):3869–3877.

Shimabukuro, Hayato, Yoshiura, Shintaro, Takahashi, Keitaro, et al. (2015), "Studying 21cm power spectrum with one-point statistics". Monthly Notices of the Royal Astronomical Society, 451(1):467–474.

Shimabukuro, Hayato, Yoshiura, Shintaro, Takahashi, Keitaro, et al. (2017a), "21 cm line bispectrum as a method to probe cosmic dawn and epoch of reionization". Monthly Notices of the Royal Astronomical Society, 568(2):1542–1550.

Shimabukuro, Hayato, Yoshiura, Shintaro, Takahashi, Keitaro, et al. (2017b), "Constraining the EoR model parameters with the 21cm bispectrum". Technical report.

Singh, Saurabh, Subrahmanyan, Ravi, Shankar, N Udaya, et al. (2018), "SARAS 2 Constraints on Global 21 cm Signals from the Epoch of Reionization".

Smith, R. E., Peacock, J. A., Jenkins, A., et al. (2002), "Stable clustering, the halo model and nonlinear cosmological power spectra". Monthly Notices of the Royal Astronomical Society, 341:1311–1332.

Sobacchi, Emanuele and Mesinger, Andrei (2014), "Inhomogeneous recombinations during cosmic reionization". Monthly Notices of the Royal Astronomical Society, 440:1662–1673.

Sumit Saha (2018), "A comprehensive guide to convolutional neural networks".

Thomas, Rajat M, Zaroubi, Saleem, Ciardi, Benedetta, et al. (2009), "Fast Large-Scale Reionization Simulations". Monthly Notices of the Royal Astronomical Society, 393(1):32–48.

Tingay, S. J., Goeke, R., Bowman, J. D., et al. (2013), "The murchison widefield array: The square kilometre array precursor at low radio frequencies". Publications of the Astronomical Society of Australia, 30(1).

Titsias, Michalis (2009), "Variational Learning of Inducing Variables in Sparse Gaussian Processes". In 12th International Conference on Artificial Intelligence and Statistics (AISTATS), pages 567–574.

van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. (2013), "LOFAR: The LOw-Frequency ARray".

Verde, L., Treu, T., and Riess, A. G. (2019), "Tensions between the Early and the Late Universe".

Watkinson, Catherine (2017), "Private Correspondence".

Watkinson, Catherine A., Giri, Sambit K., Ross, Hannah E., et al. (2019), "The 21cm bispectrum as a probe of non-Gaussianities due to X-ray heating". Monthly Notices of the Royal Astronomical Society, 482(2):2653–2669.

Watkinson, Catherine A, Majumdar, Suman, Pritchard, Jonathan R, et al. (2017), "A fast estimator for the bispectrum and beyond - a practical method for measuring non-Gaussianity in 21-cm maps". Monthly Notices of the Royal Astronomical Society, 472(2):2436–2446.

Watkinson, Catherine A. and Pritchard, Jonathan R. (2015), "The impact of spin-temperature fluctuations on the 21-cm moments". Monthly Notices of the Royal Astronomical Society, 454(2):1416–1431.

Weisstein, Eric W. (2019), "Hankel Transform".

Werbos, P (1974), Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Ph.D. thesis, Harvard University.

Wise, John H. and Cen, Renyue (2008), "Ionizing Photon Escape Fractions from High Redshift Dwarf Galaxies". The Astrophysical Journal, Volume 693, Issue 1, pp. 984-999 (2009)., 693:984–999.

Wouthuysen, S. A. (1952), "On the excitation mechanism of the 21 cm interstellar hydrogen emission line".

Wyithe, Stuart, Loeb, Abraham, and Geil, Paul (2007), "Baryonic Acoustic Oscillations in 21cm Emission: A Probe of Dark Energy out to High Redshifts".

Yoshiura, Shintaro, Shimabukuro, Hayato, Takahashi, Keitaro, et al. (2015), "Sensitivity for 21 cm bispectrum from Epoch of Reionization". Monthly Notices of the Royal Astronomical Society, 451(1):266–274.

Zahn, Oliver, Lidz, Adam, McQuinn, Matthew, et al. (2006), "Simulations and Analytic Calculations of Bubble Growth During Hydrogen Reionization". The Astrophysical Journal, 654(1):12–26.

Zaroubi, Saleem (2019), "THE REDSHIFTED 21 CM AS A PROBE OF THE EoR".

Zeldovich, YA. (1970), "Gravitational Instability: An Approximate Theory for Large Density Pertubations". Astronomy & Astrophysics, 5(5):84–89.

Zwicky, F. (1933), Helvetica physica acta., volume 6. E. Birkhaauser.