Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **498,** 4518–4532 (2020) doi:10.1093/mnras/staa2598
Advance Access publication 2020 August 29

Analysing the Epoch of Reionization with three-point correlation functions and machine learning techniques

W. D. Jennings,[1]★ C. A. Watkinson [2,3] and F. B. Abdalla[1]

[1]*Department of Physics & Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
[2]*Blackett Laboratory, Imperial College, London SW7 2AZ, UK*
[3]*School of Physics and Astronomy, Queen Mary University of London, Mile End Road, London E1 4NS, UK*

Accepted 2020 August 24. Received 2020 August 13; in original form 2019 July 22

ABSTRACT

Three-point and high-order clustering statistics of the high-redshift 21 cm signal contain valuable information about the Epoch of Reionization (EoR). We present 3PCF-FAST, an optimized code for estimating the three-point correlation function (3PCF) of 3D pixelized data such as the outputs from numerical and seminumerical simulations. After testing 3PCF-FAST on data with known analytical 3PCF, we use machine learning techniques to recover the mean bubble size and global ionization fraction from correlations in the outputs of the publicly available 21CMFAST code. We assume that foregrounds have been perfectly removed and negligible instrumental noise. Using ionization fraction data, our best multilayer perceptron (MLP) model recovers the mean bubble size with a median prediction error of around 10 per cent, or from the 21 cm differential brightness temperature with median prediction error of around 14 per cent. A further two MLP models recover the global ionization fraction with median prediction errors of around 4 per cent (using ionization fraction data) or around 16 per cent (using brightness temperature). Our results indicate that clustering in both the ionization fraction field and the brightness temperature field encode useful information about the progress of the EoR in a complementary way to other summary statistics. Using clustering would be particularly useful in regimes where high signal-to-noise ratio prevents direct measurement of bubble size statistics. We compare the quality of MLP models using the power spectrum, and find that using the 3PCF outperforms the power spectrum at predicting both global ionization fraction and mean bubble size.

Key words: methods: statistical – dark ages, reionization, first stars.

1 INTRODUCTION

A few hundred million years after the big bang, the first stars and galaxies began to form (Bromm et al. 2009). The radiation emitted from these luminous structures interacted with the surrounding neutral hydrogen and caused it to become ionized. These initially isolated ionized bubbles grew over time. Around 1 billion years after the big bang the Universe became fully ionized, see for instance Becker, Rauch & Sargent (2007) and Gunn & Peterson (1965). The phase shift from a fully neutral to a fully ionized Universe occurred during the so-called Epoch of Reionization (EoR). Many particulars about this process remain unconstrained by current data, predominantly because there are precious few sources of observable radiation during this time. Another way to observe the process of reionization would be to distinguish regions of ionized hydrogen in the neutral background. The most promising probe for this is the 21 cm hyperfine transition of hydrogen, which is only observed in neutral hydrogen. Measurements of the 21 cm signal on the sky thus provide a map of which parts of the Universe were neutral. By observing this signal at different redshifts, these maps can be extended into three-dimensional maps of the neutral hydrogen. The size and clustering properties of the ionized hydrogen bubbles change throughout the EoR.

The 21 cm signal is much weaker than other foreground sources at the same frequencies. These strong foregrounds make it difficult to extract the actual 21 cm signal. Past and ongoing purpose-built experiments such as the Murchison Widefield Array (MWA; Tingay et al. 2013),[1] the Low Frequency Array (LOFAR; Patil et al. 2017),[2] and the Precision Array for Probing the Epoch of Reionization (PAPER; Ali et al. 2015)[3] have begun to place upper limits on the overall intensity of the signal. The Experiment to Detect the Global EoR Signature (EDGES)[4] last year claimed a first detection of the 21 cm signal. This 21 cm absorption profile was observed at redshifts between $15 < z < 20$ with an amplitude of 500 mK, published in Bowman et al. (2018). This exciting result has generated much attention in the past year, as the amplitude is significantly more negative than that anticipated by standard reionization models. The strongly negative amplitude is difficult to explain without considering additional cooling mechanisms or a higher background radiation than that of the cosmic microwave background (CMB). Several recent publications have considered possible modifications that could explain the discrepancy, for instance: considering dark matter interactions (Barkana 2018; Fialkov, Barkana & Cohen 2018;

★ E-mail: wdj240@gmail.com

[1]http://www.mwatelescope.org/telescope
[2]http://www.lofar.org/
[3]http://eor.berkeley.edu/
[4]https://www.haystack.mit.edu/ast/arrays/Edges/

Muñoz & Loeb 2018); the properties of dark matter (Fraser et al. 2018; Yang 2018; Yoshiura, Takahashi & Takahashi 2018; Lawson & Zhitnitsky 2019); axionic dark matter (Moroi, Nakayama & Tang 2018; Sikivie 2018; Lambiase & Mohanty 2020); the effects of radio-wave background (Ewall-Wice et al. 2018); and considerations of mirror neutrinos (Aristizabal Sierra & Fong 2018). Other attempts have been made to explain the amplitude in terms of the foreground analysis method (e.g. Sims & Pober 2019). However, until other 21 cm observations confirm this detection it is still sensible to continue work with the standard fiducial models that exclude such exotic physics.

Upcoming experiments such as the Hydrogen Epoch of Reionization Array (HERA; DeBoer et al. 2017)[5] and the Square Kilometre Array (SKA; Mellema et al. 2013)[6] will be able to provide more detailed measurements and should allow us to understand the processes of reionization in detail and confirm the scenarios proposed to explain the EDGES detection. The most detailed theoretical modelling of the 21 cm signal currently makes use of simulations. Numerical and seminumerical simulations encapsulate many aspects of the complex non-linear reioinization processes. Common numerical simulations include C^2-RAY (Mellema et al. 2006), which models the ionizing photons emission processes and traces these rays from source to absorption; GRIZZLY (Ghara, Choudhury & Datta 2015), which uses one-dimensional radiative transfer simulations to model the radiation profiles around different source types, and then stamps these profiles on to source locations; and many codes that use adaptive refinement (Kravtsov, Klypin & Khokhlov 1997) to model both large scales and small scales in a single simulation (see e.g. the Cosmic Reionization on Computers program, Gnedin 2014, and LICORICE, Semelin, Combes & Baek 2007). Such simulations can provide theoretical predictions for a range of possible reionization scenarios, by specifying different values for simulation input parameters.

Comparisons between 21 cm data and theory often make use of fast approximate seminumerical simulations such as 21CMFAST (Mesinger, Furlanetto & Cen 2011) and SIMFAST21 (Santos et al. 2010). By running a large number of simulations for a range of different reionization scenarios, we can determine which scenarios give rise to the best match between simulated and observed data. Two techniques can make this process more efficient. First, sampling methods such as Markov chain Monte Carlo (MCMC; Greig & Mesinger 2015, 2017a, b; Pober, Greig & Mesinger 2016; Hassan et al. 2017) reduce the total number of simulations that are needed in order to hone-in on the best regions of parameter space. Second, the simulated and observed data can be compressed before comparing them by using summary statistics. These summary statistics reduce the total size of the data while retaining much of the useful information, and are more robust to modelling and sample variance errors. Common summary statistics are the power spectrum and its higher order equivalent the bispectrum (Shimabukuro et al. 2017; Watkinson et al. 2017; Majumdar et al. 2018; Giri et al. 2019; Hutter et al. 2019; Watkinson et al. 2019). Both statistics contain information about the clustering properties of ionized hydrogen bubbles.

In this paper, we use machine learning techniques to investigate using the three-point correlation function (3PCF) as another summary statistic for 21 cm data. In particular, we determine whether the 3PCF can inform us about the mean size of ionized bubbles ($R_{\rm bubble}$) and about the global ionization fraction $\langle x_{\rm H\,II}\rangle$. These statistics provide information about the progress of the EoR and encode useful information about different physical scenarios. They also provide a means to reduce the effect of thermal noise, since they are statistical quantities that are averaged over the entire map. Although the 3PCF should be less affected by noise than full 21 cm maps, the power spectrum should be even less affected. We compare the relative performances of using either the power spectrum or the 3PCF with our methodology. This indicates whether the 3PCF likely encodes any extra information about bubble size statistics than does the power spectrum.

As well as recent work using the bispectrum, some research has focused on using the 3PCF as a tool for investigating the EoR. Gorce & Pritchard (2019) use a derived statistic from the 3PCF to concentrate on phase information. Hoffmann et al. (2019) investigate whether the 3PCF of 21 cm data can be modelled using a local bias model. Their resulting model makes predictions with around 20 per cent accuracy for large ionized regions at early times, but breaks down for other scenarios.

Machine learning has already been suggested and used for a number of different applications with 21 cm data: to emulate power spectrum outputs quickly from 21CMFAST (Kern et al. 2017; Schmit & Pritchard 2018; Jennings et al. 2019), to derive reionization parameters directly from the 21 cm power spectrum (Shimabukuro & Semelin 2017), and to derive reionization parameters from 21 cm images (Gillet et al. 2018). Jennings et al. (2019) also present a mapping between 21CMFAST and SIMFAST21 providing a proof of concept for mapping between simulations that predict different EoR histories.

In this paper, we run a large representative sample of seminumerical simulations using 21CMFAST. For each simulation, we calculate the 3PCF of the resulting 21 cm maps. We also measure the characteristic reionization features: the global ionization fraction and the size distribution of the ionized bubbles. We then use machine learning techniques to determine the relationships between the 3PCF measurements and the characteristic reionization features.

The rest of the paper is split in to the following sections. Section 2 describes the mathematical concept behind the 3PCF, and a description of the code implementation. We also test the code on data with known analytical 3PCF. In Section 3, we describe current physical models of the reionization process. We include a description of the 21CMFAST code and a summary of the range of reionization scenarios considered in this paper. Section 4 gives an overview of the machine learning techniques we use, including the search strategy that we use to find the best possible model. We also summarize the methods used to analyse the performance of the resulting models. In the remaining sections, we use our data to learn about the characteristic reionization features: the mean bubble size in Section 5 and the global ionization fraction in Section 6. We end the paper in Section 7 with our conclusions. For cosmological parameters, we use $\Omega_{\rm m} = 0.3153$, $\Omega_{\rm b} = 0.0493$, $\Omega_\Lambda = 0.6847$, $H_0 = 67.36\ {\rm km\,s^{-1}\,Mpc^{-1}}$, $n_{\rm s} = 0.9649$, and $\sigma_8 = 0.8111$, the latest results using the default *Plik* likelihood from Planck Collaboration (2018).

2 THREE-POINT CORRELATION CALCULATION

The 3PCF $\xi^{(3)}(\boldsymbol{r_1}, \boldsymbol{r_2}, \boldsymbol{r_3})$ is defined as the ensemble average over triplets of points in real space

$$\xi^{(3)}(\boldsymbol{r_1}, \boldsymbol{r_2}, \boldsymbol{r_3}) = \langle \delta(\boldsymbol{r_1})\delta(\boldsymbol{r_2})\delta(\boldsymbol{r_3})\rangle\,. \quad (1)$$

[5] http://reionization.org/
[6] https://www.skatelescope.org/

The angular brackets denote an ensemble average over a large region of space (or over a large number of universe realizations) to mitigate the effect of statistical fluctuations. If the signal is translationally invariant then the ensemble average can be replaced by a spatial average. If the signal is also rotationally invariant then the 3PCF depends only on the lengths ($r_1 = |\boldsymbol{r_1}|$) of the real-space vectors and not on their directions. These invariance assumptions are broken in real observations, in part due to the light-cone effect (for example Datta et al. 2014) and redshift-space distortions (for example Majumdar et al. 2015). Note that the three vectors $\boldsymbol{r_1}$, $\boldsymbol{r_2}$, and $\boldsymbol{r_3}$ connect three points in real space and so have a vector sum of $\boldsymbol{0}$, i.e. they form a closed triangle. In practice, $\xi^{(3)}$ measurements are actually made over *configurations* of triangles, by which we mean over sets of unique triangle side lengths in order to beat down statistical noise. The 3PCF for a single triangle configuration is an average over all triangles with those side lengths r_1, r_2, and r_3, namely

$$\xi^{(3)}(r_1, r_2, r_3) = \langle \delta(\boldsymbol{r_1})\delta(\boldsymbol{r_2})\delta(\boldsymbol{r_3})\rangle_{(|\boldsymbol{r_1}|,|\boldsymbol{r_2}|,|\boldsymbol{r_3}|) = (r_1,r_2,r_3)}. \quad (2)$$

2.1 Code implementation

Calculating the 3PCF 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 3PCF is built up. The final output of the algorithm is the 3PCF estimates $\xi^{(3)}(\boldsymbol{r_i})$ at a discrete set of radius values $\boldsymbol{r_i} = (r_1, r_2, r_3)$, corresponding to a discrete set of radius bins. The 3PCF estimate for each radius bin is calculated by using a set of many triangles with similar (but not identical) side lengths. In this section, we first describe how to find these sets of triangles, by matching triangles whose side lengths lie within a given binned range of radii $R_{\rm min} \leq r < R_{\rm max}$. We also provide pseudo-code for our C++ 3PCF-FAST algorithm (publicly available on GitHub).[7] Finally, we discuss how we use the output statistics from the 3PCF-FAST code to estimate 3PCF values. Our algorithm is similar in nature to other high-order codes (see for instance Gaztañaga & Scoccimarro 2004), although we subsample both the triangle configurations and the number of lattice points and measure the level of approximation needed for robust estimates of the 3PCF.

2.1.1 3PCF-FAST for 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 pixelized set of scalar values in three dimensions. For each radius bin, we 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, we place the first vertex at the origin. We then find all possible second vertices ($\boldsymbol{r_2}$) which lie within the spherical shell $R_{\rm min} \leq |\boldsymbol{r_2}| < R_{\rm max}$ of the origin. From each of the matching second vertex points, we find the third vertex points ($\boldsymbol{r_3}$) which are a valid distance both from $\boldsymbol{r_2}$ and from the origin. This last step is effectively finding pixels which lie in the overlap of two spherical shells. Fig. 1 shows an example in two dimensions: with the first triangle vertex at the origin, the dark purple annulus indicates the allowed region for the second vertex between $R_{\rm min}$ and $R_{\rm max}$. The orange region then

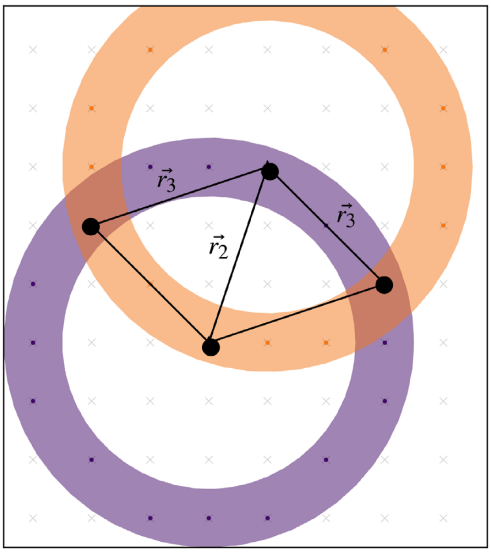

Figure 1. Triangles matching the radius bin condition 2.5 pixels $\leq r <$ 3.5 pixels. The two regions shown are the radius conditions around the first and second points. The allowed third point(s) then lie in the overlap of these annuli.

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. To prevent repeated calculations, we use a PYTHON script to search for these matching triangles and store the resulting pairs of vectors ($\boldsymbol{r_2}$, $\boldsymbol{r_3}$) in a binary file. This binary file can be loaded by the main C++ algorithm many times. we refer to these binary files as *vertices* files. Measurements of the 3PCF are calculated by looping over possible lattice points $\boldsymbol{r_1}$ and summing the contributions for all triangle configurations ($\boldsymbol{r_2}$, $\boldsymbol{r_3}$) at that pixel. Both the number of triangle configurations and the number of lattice points are sampled to give faster calculations, and we investigate different levels of sampling in Section 2.3.

The 3PCF of a data field is usually calculated in comparison to a random field without clustering. The correlation function then quantifies the extent to which the data field is more clustered than the random unclustered field. The purpose of the random field is to create a comparison for the data field. Using a uniform field can thus be seen as a method for counting the number of triangle configuration occurrences. The outputs from our three-point 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 autocorrelation 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 3PCF. For a scalar data field, the random field should be uniform with mean equal to the data mean. Instead, it is practically simpler and mathematically identical to normalize the *data* field to have a mean of unity, so that the random field averaged within in each pixel is also everywhere unity. This allows our code to skip the correlation calculations for the random field, since the value of RRR is equal to the known integer count of triangles. Algorithm 1 shows the pseudo-code for our algorithm, taking as inputs a data field $D[\boldsymbol{r}]$ and a binned *vertices* file, and outputting the three-point correlation statistics (DDD, DDR, DRR, and RRR) for each radius bin.

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

Algorithm 1 Three-point correlation algorithm

```
procedure 3PCF(D[r], R_min, R_max)
    DDD, DDR, DRR, RRR ← 0                    ▷ Initialise to zero
    Load (r_2, r_3)                           ▷ using (R_min, R_max) vertices file
    for all r_1 do                            ▷ over all data pixels
        for each r_2, r_3 pair do             ▷ over matching triangles
            DDD += D[r_1] × D[r_1 + r_2] × D[r_1 + r_3]
            DDR += D[r_1] × D[r_1 + r_2]
            DRR += D[r_1]
            RRR += 1
        end for
    end for
    return DDD, DDR, DRR, RRR
end procedure
```

Algorithm 1 outputs the correlation statistics (DDD, DDR, DRR, RRR). An estimate of the 3PCF is found by combining these statistics. The simplest such estimator is given by ratios of the data- and random-field autocorrelations

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

Another estimator

$$\xi^{(3)} = \frac{DDD - 3DDR + 3DRR - RRR}{RRR}, \tag{4}$$

from Landy & 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.

The number of triangles found by this matching algorithm can quickly exceed hundreds of thousands for side lengths larger than around 10 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 3PCF can be obtained efficiently by subsampling a small number of triangles from all valid matches. In this section, we sample 5000 occurrences of triangle configurations from the total available number of valid matches. In Section 2.3, we discuss the effect of sampling and why a value of 5000 was chosen.

2.2 Testing 3PCF-FAST using points on spheres

We test our code by generating three-dimensional realizations for a distribution with a known 3PCF. We compare the measured 3PCF from our code to the theoretical form, to get an indication of the regimes in which the code has good accuracy and precision. Our testing distribution consists of three-dimensional realizations 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 pixelized grid and counting the number of points in each grid: zero for no points, one for a single point, and so on. For all realizations in this section, the data are represented as a box with side length 100 arbitrary units pixelized into 512^3 pixels. The amplitude and shape of the theoretical 3PCF for these realizations depend on the sphere radius R and the number density of spheres n_s. We describe a scenario as a particular pair of these two parameters. We also use the number of spheres $N_s = n_s \times 10^6$, since all realizations in this section have a fixed box size of 100 arbitrary units in each of the three dimensions. The equilateral 3PCF of points-on-sphere realizations has a closed analytical form (Lorne Whiteway, private correspondence). For a scenario with parameters n_s and R, the 3PCF for equilateral triangles as a function of the triangle side length r is given by

$$\xi^{(3)}(r; R, n_s) = \begin{cases} \dfrac{1}{16\pi^3 R^3 n_s^2 r^2 \sqrt{3R^2 - r^2}} & \text{if } r < R\sqrt{3}, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

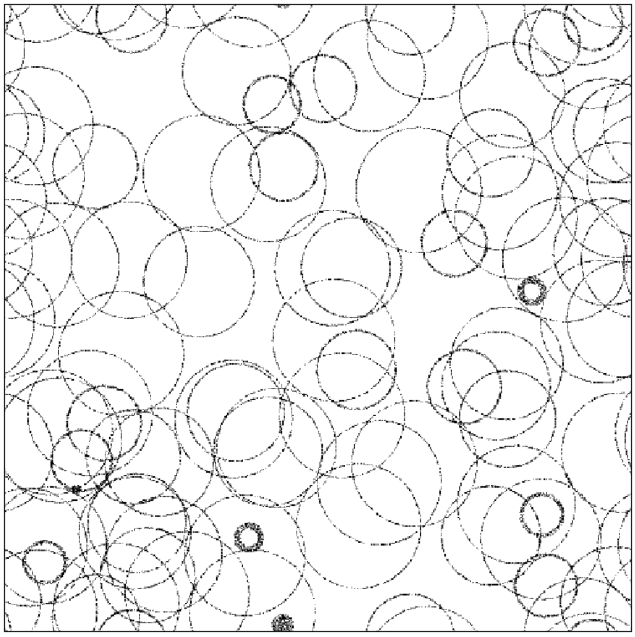

Figure 2. Slice through an example realization of points-on-spheres data. This scenario uses spheres with $R = 10$ and $N_s = 200$. Each sphere appears as a circular annulus as it has been horizontally sliced for this figure. Some annuli appear thicker than others because slicing a thick spherical shell near its pole gives a wider region when viewed from above.

Generating a realization 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 is chosen to be the centres of the spheres. Points are then placed randomly on to the surface of each sphere. Ensuring that the points are indeed uniformly distributed across the spheres surface is most easily done using the method from Muller (1959): sample three random variables x, y, z from the normal distribution $\mathcal{N}(0, 1)$ and divide by the Euclidean norm of these three coordinates. The distribution of the normalized vector

$$\boldsymbol{r} = \frac{R}{\sqrt{x^2 + y^2 + z^2}} \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \tag{6}$$

is then uniform across the surface of a sphere with radius R. After storing the locations of all points on all spheres, the final pixelized realization of the scalar field is generated by rounding the point coordinates to the nearest integer. Fig. 2 shows a slice through an example realization of the testing distribution. All data in this testing section have pixel size of around 0.2 arbitrary physical units.

2.2.1 *Test results*

We test our code by generating points-on-spheres realizations for many R and N_s scenarios. We compare the outputs of our code to that of the true theoretical 3PCF using equation (5). Fig. 3 shows the

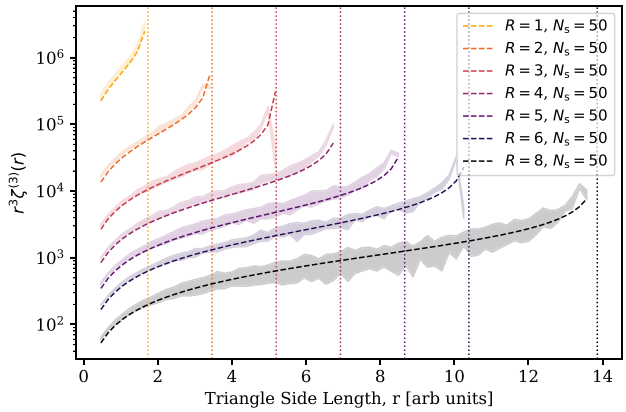

Figure 3. Equilateral 3PCFs for points-on-spheres scenarios with varying sphere radius R. The dashed lines show the theoretical 3PCF for each scenario whose parameters are shown in the legend. The shaded regions show the standard deviation of the measured 3PCF across five realizations, using our code 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. Vertical dotted lines show the theory asymptotes at $R\sqrt{3}$ for each line.

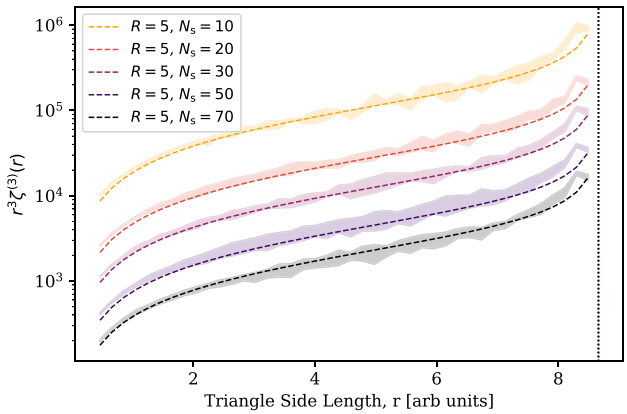

Figure 4. Equilateral 3PCFs for points-on-spheres data with varying n_s between 1×10^{-5} and 7×10^{-5}, using the LS estimator. The dashed theory lines again lie within the measured shaded regions for most triangle side lengths. Vertical dotted line shows the theory asymptote at $R\sqrt{3} = 5\sqrt{3}$.

theoretical and measured equilateral 3PCFs for seven scenarios with a range of R values and fixed $n_s = 5 \times 10^{-5}$, using the Landy–Szalay estimator in equation (4). We plot the dimensionless 3PCF defined as $r^3\xi^{(3)}$ (see e.g. Hoffmann et al. 2019). The theoretical 3PCF is shown in each case as the dashed line. The measured 3PCF estimates are subject to sample variance, meaning that the output from the code depends on the randomly seeded initial conditions. We use five realizations with different random seeds to determine whether the theoretical 3PCF lies inside the spread of the five measured code outputs. The shaded regions in Fig. 3 show the standard deviation of the measured 3PCF across these five realizations. Fig. 4 similarly shows the theoretical and measured 3PCFs for scenarios with fixed $R = 5$ and various N_s values.

The measured and theoretical 3PCFs match closely across most of the triangle side lengths. The theoretical 3PCF in equation (5) has a vertical asymptote at the maximum allowed radius $R\sqrt{3}$. This can be seen in Fig. 3 as a slight upturn near the right-hand sides of each dashed line. Our code slightly overpredicts 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 2.1.1. Averaging the 3PCF over these differently sized triangles (some of which are larger than the valid maximum radius) causes a discrepancy between measured and theoretical 3PCF.

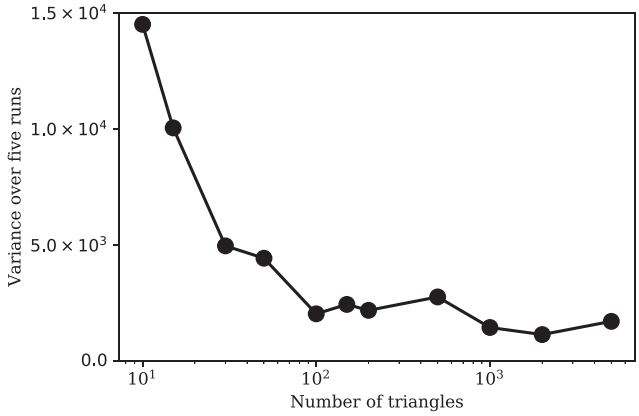

Figure 5. Effect of subsampling triangles in the 3PCF algorithm. When few triangles are used, the outputs from the code show a larger variance than when more triangles are used. For more than around 2000 triangles, the variance plateaus indicate that adding more triangles provides minimal extra information.

2.3 Optimization

A number of steps were taken to optimize and improve the code. First, we added multithreading to make better use of available computational resources. Second, we allow for subsampling of triangle configurations and jackknifing to allow for calculation of errors. The number of triangles found by the matching algorithm in Section 2.1.1 can quickly exceed hundreds of thousands for side lengths larger than around 10 pixels. An accurate measurement of the 3PCF can be obtained more efficiently by subsampling a small number of triangles from all valid matches. We test how this subsampling affects the robustness of the final 3PCF estimate. The 3PCF is calculated on $x_{\rm HII}(\boldsymbol{r})$ data from five randomly seeded 21CMFAST realizations using as input the canonical input parameters $T_{\rm vir} = 10^4$K, $\zeta_{\rm ion} = 30.0$, $R_{\rm max} = 15.0$ Mpc, and $E_0 = 200$ eV. The variance in our code's output over the five realizations is plotted in Fig. 5, as a function of the number of triangles used in the 3PCF algorithm. The variance is large when only a few triangles are used but decreases with a larger number of triangles. For more than around 2000 triangles, the scatter plateaus indicate that adding more triangles is unlikely to result in improved final 3PCF estimates. The remaining variance across the five runs is likely due to sample variance. We use a conservative value of 5000 triangles in all the 3PCF estimates from hereafter.

3 MODELS OF REIONIZATION

The 21 cm differential brightness temperature $\delta T_{\rm b}$ is defined as the difference between the measured 21 cm brightness temperature and the uniform background CMB brightness temperature. By removing the background CMB temperature, the value of $\delta T_{\rm b}(\boldsymbol{r})$ then specifies the extent of 21 cm emission ($\delta T_{\rm b} > 0$) or absorption ($\delta T_{\rm b} < 0$). The actual observable for radio interferometers is $\delta T_{\rm b} - \langle \delta T_{\rm b} \rangle$, where

$\langle \delta T_{\rm b} \rangle$ is the global reionization signal averaged across the whole sky. Furlanetto et al. (2006) give an approximate relationship for the 21 cm brightness temperature $\delta T_{\rm b}(\boldsymbol{r})$ as

$$\delta T_{\rm b}(\boldsymbol{r}) = 27\, x_{\rm H\,I}(\boldsymbol{r}) \left[1+\delta(\boldsymbol{r})\right] \left(\frac{\Omega_{\rm b}h^2}{0.023}\right) \left(\frac{0.15}{\Omega_{\rm m}h^2}\right)^{1/2} \times \left(1-\frac{T_\gamma}{T_{\rm spin}}\right) \left(\frac{1+z}{10}\right)^{1/2} \left(\frac{H(z)}{H(z)+\delta_r v_r(\boldsymbol{r})}\right) {\rm mK}. \quad (7)$$

This approximation includes the effects of neutral hydrogen fraction $x_{\rm H\,I}(\boldsymbol{r})$; total matter density contrast $\delta(\boldsymbol{r})$; 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 $T_{\rm S}$ which quantifies the relative populations of electrons in the higher and lower energy states of the 21cm transition; the Hubble parameter $H(z)$; and $\delta_r v_r(\boldsymbol{r})$, the radial velocity gradient.

The spin temperature can be written (Furlanetto et al. 2006) as a sum of three parts

$$T_{\rm spin}^{-1} = \frac{T_\gamma + x_\alpha T_\alpha^{-1} + x_{\rm c} T_{\rm K}^{-1}}{1 + x_\alpha + x_{\rm c}} \quad (8)$$

with T_γ the background CMB temperature, $T_{\rm K}$ the kinetic gas temperature, T_α the Lyman alpha colour temperature (closely linked to the gas temperature for all redshifts of interest), and the coupling coefficients for collisions ($x_{\rm c}$) and the Wouthysen–Field coefficient (x_α) from Wouthuysen (1952). In particular, the kinetic gas temperature and the Lyman alpha background radiation have a strong effect on the global and local evolution of the spin temperature. These two features change throughout the EoR as different physical processes interact with the growth of structure in the Universe.

3.1 21CMFAST

We use the publicly available seminumerical code 21CMFAST to generate our data. We briefly describe the algorithm in this subsection. The simulation begins by seeding an initial linear density field on to a three-dimensional grid at very high redshift. This linear density field is evolved using first-order perturbation theory (see Zeldovich 1970) to approximate gravitational collapse, giving an approximate gravitationally evolved density field $\delta(\boldsymbol{r})$.

The simulation then finds the highest density regions where the matter will collapse to form luminous structures and thus contribute ionizing photons towards the reionization process. The extent of collapse is calculated directly from the non-linear density field following the model of spherical collapse (Press & Schechter 1974).[8] If the mean enclosed density in a region exceeds a theoretical critical value then the region is assumed to collapse. The collapse fraction $f_{\rm coll}(\boldsymbol{r}, R)$ on decreasing scales R is then found from the contributions of both resolved and unresolved haloes. The default 21CMFAST implementation has a minimum halo mass $M_{\rm min}$ for a halo to host star-forming galaxies that evolves with redshift, corresponding to a minimum virial temperature $T_{\rm vir}$ for ionizing photons.

The ionization fraction field $x_{\rm H\,II}(\boldsymbol{r})$ is found by determining whether the collapsed matter in a region generates enough ionizing photons to ionize the enclosed hydrogen atoms. An ionizing efficiency parameter $\zeta_{\rm ion}$ specifies how many ionizing photons are sourced per unit of collapsed matter. If $f_{\rm coll}(\boldsymbol{r}, R) \geq \zeta_{\rm ion}^{-1}$ for any particular region, then the central pixel is painted as fully ionized using the method in Zahn et al. (2006). This differs from the default method of another common seminumerical simulation SIMFAST21, which instead paints the full spherical region as ionized if there are enough photons using the method in Mesinger & Furlanetto (2007). See Hutter (2018) for a discussion of these two methods.

Fluctuations in the spin temperature are calculated by considering the kinetic gas temperature and the Lyman α background temperature. The kinetic gas temperature $T_{\rm K}$ is 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. The rate of emitted X-ray photons is assumed to be proportional to the growth rate of collapsed matter in the dark matter haloes. 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 intergalactic 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. See Mesinger et al. (2011) for a full derivation of the calculations and assumptions that 21CMFAST makes for the spin temperature fluctuations.

The final step is to use equation (7) and calculate the 21 cm brightness temperature field $\delta T_{\rm b}(\boldsymbol{r})$ using the non-linear density field $\delta(\boldsymbol{r})$, the neutral fraction field $x_{\rm H\,I}(\boldsymbol{r}) = 1 - x_{\rm H\,II}(\boldsymbol{r})$, and the spin temperature fluctuation field $T_{\rm spin}(\boldsymbol{r})$.

In this paper we consider different reionization scenarios by changing three of these simulation parameters:

(i) The ionization efficiency $\zeta_{\rm ion}$, specifying how many ionizing photons are sourced per unit of collapsed matter;
(ii) The E_0 parameter which controls the minimum energy (or frequency) of X-ray photons which are able to escape into the IGM;
(iii) The minimum virial temperature $T_{\rm vir}$ which specifies a lower mass limit $M_{\rm min}$ of collapsed matter which produces ionizing photons and X-rays.

Fixing the other simulation parameters involves setting the efficiency of X-rays to a constant value. We use $\zeta_{\rm X} = 10^{-57}\,{\rm M}_\odot^{-1}$ 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 IGM X-ray properties are then parametrized by E_0.

3.2 Training and testing data details

We run 1000 21CMFAST simulations in total for our data. Each simulation generates three-dimensional realizations of the $\delta T_{\rm b}$ field in a cube of size 250 Mpc resolved into 256^3 pixels (smoothed from density fields resolved into 768^3 pixels). Each simulation uses a different random seed for the initial conditions. The resulting redshifts from this algorithm are between $z = 5$ and $z = 26.6$ (see Mesinger et al. 2011 for a description of the iterative algorithm that generates these steps). These redshifts 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. We ignore simulated results for higher redshifts, because the mean ionization fraction is extremely small and the mean bubble size is generally smaller than the resolution of our simulations. For each simulation, we calculate the statistics of interest: the 3PCF using 3PCF-FAST described in Section 2; the bubble size distribution, described in this section;

[8] In order to match the more accurate ellipsoidal collapse model (Sheth, Mo & Tormen 2001), 21CMFAST afterwards normalizes the spherical collapse fractions so that their average value matches that expected from ellipsoidal collapse.

and the global ionization fraction, found by trivially averaging the ionization fraction field $x_{\rm H\,II}(\boldsymbol{r})$ for each redshift.

In order to sample a range of different reionization scenarios, we use a Latin Hypercube (McKay, Beckman & Conover 1979) approach.[9] 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:

(i) $T_{\rm vir}$ in the logarithmic range $[10^4, 2 \times 10^5]$ K,
(ii) $\zeta_{\rm ion}$ in the linear range [5, 100],
(iii) E_0 in the linear range [100, 1500] eV.

These ranges were chosen to match those by the simulation authors (e.g. Greig & Mesinger 2015). The lower $T_{\rm vir}$ limit comes from a minimum temperature for the cooling of atomic hydrogen accreting on to haloes. The upper limit arises from observations of high-redshift Lyman break galaxies (Greig & Mesinger 2015). The $\zeta_{\rm ion}$ upper and lower limits correspond roughly to escape fractions of 5 per cent to 100 per cent for ionizing photons for standard values of the other controlling factors in Greig & Mesinger (2015) such as the number of ionizing photons produced per stellar baryon. The range for E_0 was chosen in a similar way to Park et al. (2019), 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 $\langle x_{\rm H\,II}(z)\rangle = 0.5$.

3.2.1 *Three-point correlation function measurements*

We use the code described in Section 2 to calculate the 3PCF. We calculate $\xi^{(3)}$ of both the ionization fraction field $x_{\rm H\,II}(\boldsymbol{r})$ and of the 21 cm differential brightness temperature field $\delta T_{\rm b}(\boldsymbol{r})$. In the code, we use 28 equilateral triangle bin configurations with side lengths spaced in bins between 5 and 109 Mpc. These bins are spaced linearly for radii less than 20 Mpc, with logarithmically spaced bins for higher radii.[10] Increasing the number of r-vector configurations beyond these equilateral triangles would almost certainly improve our ability to predict the mean bubble size or ionization fraction history. Further work would be needed, however, to investigate what size and shape of triangle configurations encode the most information about the topology of the EoR.

3.2.2 *Mean free path measurements for $x_{\rm H\,II}(\boldsymbol{r})$*

In order to measure the mean bubble size, we use our own implementation of the mean free path method described in Mesinger & Furlanetto (2007). The mean free path method simulates the emission of photons from random locations within the transparent regions. The distance travelled by each photon before it hits a phase change (from ionized to neutral) is measured and the resulting number of rays in a range of radius bins is calculated as dP/dR. We use 10^5 simulated photons in our measurements, and the resulting distances are rounded to the nearest pixel size ($L/N = 250.0$ Mpc$/256 = 0.98$ Mpc). Fig. 6 shows the resulting distributions for RdP/dR from a simulation with canonical parameter values $T_{\rm vir} = 10^4$K, $\zeta_{\rm ion} = 30$, and $E_0 = 200$ eV. We use the mean of these mean free path distributions (hereafter written $R_{\rm bubble}$) as a statistic to trace the mean bubble size.

[9]Using the implementation from Agarwal et al. (2014).
[10]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.

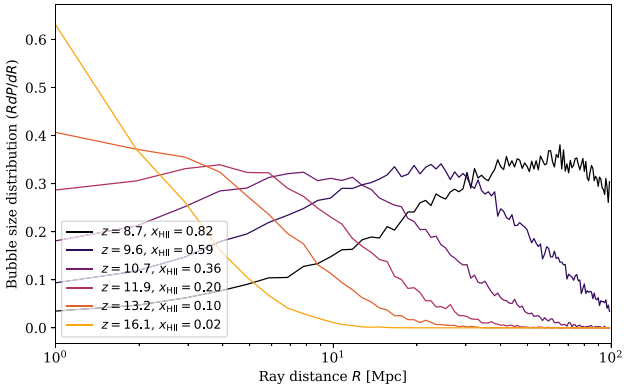

Figure 6. Example mean free path measurements of RdP/dR using ionization fraction field data $x_{\rm H\,II}(\boldsymbol{r})$. Each line shows RdP/dR for a single redshift taken from a simulation with $T_{\rm vir} = 10^4$K, $\zeta_{\rm ion} = 30.0$, and $E_0 = 200$ev.

4 MACHINE LEARNING TECHNIQUES

The 3PCF of 21 cm data likely encodes much information about the underlying reionization processes. We can uncover the relationships between $\xi^{(3)}$ and these physical processes by looking at how the 3PCF changes over a range of physical scenarios. We use machine learning techniques to learn these relationships. Our models extract relationships between physical processes and $\xi^{(3)}$ measurements by using simulated data. Our trained models can then use unseen measurements of 3PCF data to make predictions about the physical status of reionization in the data. In particular, we train models that predict the global ionization fraction and the mean bubble size from $\xi^{(3)}$. Training these models is effectively a form of high-dimensional curve fitting: learning a best-fitting functional form $f(\boldsymbol{x})$ that maps from a set of input values $\boldsymbol{x} = \xi^{(3)}$ to a set of output values ($R_{\rm bubble}$ or $\langle x_{\rm H\,II}(z)\rangle$). After training, our models can make predictions for new unseen data. For instance, the mean bubble size model can take measurements of $\xi^{(3)}$ and predict the mean bubble size. In this section, we describe the machine learning techniques we use along with a theoretical description of how they are trained. All models are trained on the same architecture, each on a single node using 16 Xeon E5-2650 cores and 128GB RAM. 700 of our simulations are used for training and validation, and 300 simulations are held back for testing.

4.1 Artificial neural networks

Artificial neural networks (ANNs) are a common regression technique for learning a complex non-linear relationship between two sets of variables: the 'inputs' and 'outputs'. An ANN represents the relationship in functional form $y_i = f(\boldsymbol{x_i})$ by manipulating the inputs $\boldsymbol{x_i}$ through a series of weighted summations and function evaluations. For ANNs, this series of repeated operations occurs in a series of distinct layers. The values in the first layer $\boldsymbol{h^{(0)}}$ are the input variables $\boldsymbol{x_i}$. The values from one layer $h_j^{(l-1)}$ affect the values in the following layer $h_j^{(l)}$ according to

$$\boldsymbol{h^{(l)}} = h_j^{(l)} = \phi_\theta \left(\sum_{i=1}^{N_i} W_{ij}^{(l)} h_j^{(l-1)} \right). \tag{9}$$

The values in each layer are thus a sum over the values in the previous layer, weighted using a set of trainable values $W_{ij}^{(l)}$. The summations into each neuron are passed through an activation function $\phi_\theta(x)$, which determines the resulting output values that are passed on to

the next layer of neurons. At the end of this process, the final layer contains the network's fitted evaluations of the function, $y_i = f(\boldsymbol{x_i})$. Training these models involves choosing the set of weights $W_{ij}^{(l)}$ which most closely mimic the function's behaviour. The 'closeness' with which the model mimics the relationship in the training data is quantified using an objective function

$$\text{Objective} = \frac{1}{2N}\sum_{n=1}^{N}\left[f(\boldsymbol{x_n}) - y_n\right]^2 - \frac{\alpha}{2}\sum_{i,j,l}\left(W_{ij}^{(l)}\right)^2 \tag{10}$$

so that training is then done by finding the values $W_{ij}^{(l)}$ which minimize this objective function for some training data $(\boldsymbol{x_n}, y_n)$. The regularization parameter α in this equation allows finer control over the complexity of the model. A high value of α encourages the training towards simpler models, with more of the weight values $W_{ij}^{(l)}$ being near zero. Three common activation functions $\phi_\theta(x)$ are the hyperbolic tangent function $\phi_\theta(x) = \tanh(x)$; the logistic function $\phi_\theta(x) = 1/(1+\exp(-x))$; and the rectified linear unit ('relu') function $\phi_\theta(x) = \max(0, x)$. All three activation functions are used during our hyperparameter search method.

The weights are initialized randomly and are then updated iteratively in order to improve the objective function. Each iteration is known as a single 'epoch'. In each epoch, the weights are updated using the current gradient of the objective. By using this gradient, the weights are moved towards a value that should cause the objective function to improve. Our ANNs use the backpropagation algorithm (Werbos 1974), a common technique for efficiently calculating the gradient of the objective function (see Rumelhart, Hinton & Williams 1986 for a more detailed description of this algorithm). 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 the learning. An adaptive learning rate is usually set to decrease 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. We discuss this and other choices made in our models in Section 4.2 later.

Multilayer perceptrons (MLPs) are a subclass of ANNs, with the restrictions that they contain at least one hidden layer and have a non-linear activation function. Fig. 7 shows a typical MLP's layer structure, with lines representing the weighted connections between values. Circles represent the neurons which hold the values $h_j^{(l)}$ and pass the weighted inputs through the activation function. Our MLPs are implemented using SCIKIT-LEARN from Pedregosa et al. (2011). We use the 'adam' optimization method (Kingma & Ba 2014), which terminates either when the maximum number of epochs has been reached or when the objective function falls below a tolerance of 10^{-10} for at least two consecutive iterations.

We measure the goodness of fit between predicted output values $y^*(k, z)$ and measured output values $y(k, z)$ using a mean squared error function

$$\text{MSE}\left[y, y^*\right] = \frac{1}{N}\sum_{i}^{N}\left(\frac{y_i - y_i^*}{y_i}\right)^2, \tag{11}$$

also making use of the root mean squared error $\text{RMSE} = \sqrt{\text{MSE}}$ and the percentage mean squared error $=100.0 \times$ rmse. A percentage rmse of 100 indicates that the predicted and measured output values are wrong by an average factor of 2.

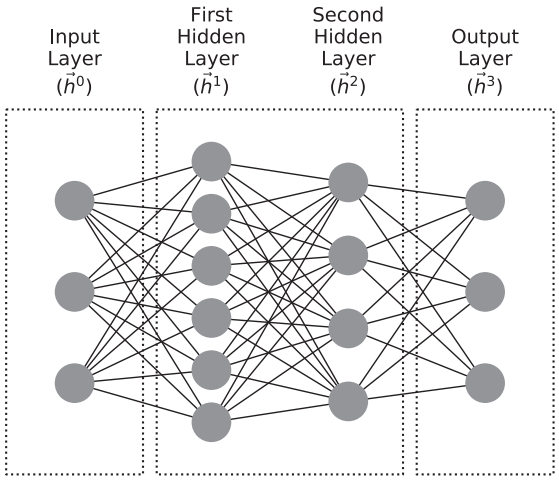

Figure 7. Visualization of an MLP with two hidden layers. Lines are weighted connections from left to right. Circles are neurons which hold the values and pass them to the following layer.

4.2 Hyperparameter search

The weighted-connection values W_{ij} of an MLP are updated during the training process in order to find the best match between the input and outputs in the training data. Several aspects of models must also be fixed before even starting to train the model. We refer to these values as hyperparameters. The hyperparameters can have a strong effect on the final accuracy of predictions but it is rarely obvious what hyperparameter values will result in the most accurate model. We use a random search method with cross-validation to find the best hyperparmeters for each of our model applications. This process is described here.

In order to determine the best hyperparameter values, we train and compare a large number of models with a range of initial hyperparameter values. Each model is trained using a set of randomized hyperparameters and the model with highest prediction accuracy is selected as the best model. 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 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 3PCF measurements as the inputs, there are around 30 input dimensions to the model. We 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 a similar order of magnitude to the input dimensionality while also remaining computationally feasible. The full set of parameters which were randomly varied for each model in the hyperparmaeter search are:

(i) Number of hidden layers uniformly in the linear range [1, 3],
(ii) Size of each layer uniformly in the linear range [0, 500],
(iii) Training batch size uniformly in the linear range [30, 500],
(iv) Number of training epochs uniformly in the range [50, 500],

(v) Initial learning rate uniformly in the log range $[10^{-4}, 10^{-2}]$,
(vi) Learning rate either constant and adaptive with equal chance,
(vii) Activation from 'relu', 'tanh', or 'logistic' with equal chance,
(viii) Regularization parameter α from equation equation (10) uniformly in the log range $[10^{-4}, 10^{-2}]$.

These ranges match those suggested by the SCIKIT-LEARN website (Pedregosa et al. 2011). We use fixed default values for the 'adam' parameters BETA_1 $= 0.9$, BETA_2 $= 0.999$, EPSILON $= 1e-08$, and TOL $= 0.0001$. For all models, the weight values are initialized using the Xavier initialization strategy (Glorot & Bengio 2010). This method sets the weights in the ith layer by sampling uniform values in the range $[-U_i, U_i]$. The normalizing 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'). Note that there are seven hyperparameters being varied in this random search. Given that we choose only 1000 different random sets from the above ranges, it is unlikely that we have identified the optimal model.

4.3 Cross-validation

By trying a range of different hyperparameter values as described, we can usually find a model with better prediction accuracy. However, this process is sensitive to overfitting. In order to determine which model has the highest accuracy while reducing the chance of overfitting, we use five-fold cross-validation approach. Five models are trained with the same fixed hyperparameters, where each model is provided with data from only four of the five folds. In each case, the fifth excluded fold is used to calculate the prediction performance, using equation (11). The performances are thus measured on unseen data, so that the 'best' model with highest performance is one which performs well on the unseen validation data. The overall accuracy score is taken as the mean of the validation scores. This cross-validation approach is used to compare the performance of every combination of hyperparameter values. After finding the best hyperparameter values, the model is trained for a final time using all the training data. Standard practice for machine learning tasks is to retain a final segment of the data to check the final performance of the best model. If the model performs well on this testing data, then we are more confident that it makes good predictions for completely unseen data.

4.4 Input and output scaling

Data scaling can be used to improve the efficiency of ANNs during training, and also to improve the quality of the final predictions. The weight values in our neural networks are initialized at small values as described in Section 4.2. 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 normalizing or standardizing each input feature. Normalizing an input feature forces all values to lie in the range [0, 1] by linearly scaling the minimum and maximum feature values. Standardizing an input feature scales the feature to have a mean of zero and a standard deviation of 1.0. Scaling the model output value(s) also has a beneficial effect on the final prediction accuracy. Our neural networks use the SCIKIT-LEARN objective function in equation (10) to quantify the goodness of fit during training. Scaling the output values using normalization or standardization can help mitigate the relative importance of output values with different magnitudes.

The input features to our models are the 3PCF measurements $\xi^{(3)}(r)$ for a range of different triangle sizes r. These 3PCF values span a wide range of magnitudes. We use the MinMaxScaler method from SCIKIT-LEARN to normalize separately each 3PCF bin. We also compare the effect of scaling the 3PCF values by four different powers of the binned radius values: the raw 3PCF $\xi^{(3)}$; the dimensionless 3PCF $r^3\xi^{(3)}(r)$ used for more natural visualizations (see for instance Hoffmann et al. 2019); and two other powers of the radius for completeness: $r\xi^{(3)}(r)$ and $r^2\xi^{(3)}(r)$. The output features to our models are either the bubble sizes $R_{\rm bubble}$ or the global ionization fraction $\langle x_{\rm H\,II}(z)\rangle$. We scale the $R_{\rm bubble}$ function using the $\sinh^{-1}$ function as described by Lupton, Gunn & Szalay (1999).

5 LEARNING TYPICAL BUBBLE SIZES FROM THE 3PCF

The progress of the EoR can be tracked by measuring the mean size of ionized regions. Ionized regions are initially small and isolated around the earliest ionizing sources. The regions continually grow throughout the EoR, and the precise details of this continued growth depends on the physical interactions between ionizing sources and the surrounding neutral regions. The sources themselves are seeded from the clustered non-linear density field and so show significant clustering (Kramer, Haiman & Oh 2006), but the details of reionization also affect the clustering of the resulting ionization fraction field $x_{\rm H\,II}(\boldsymbol{r})$ and 21 cm brightness temperature field $\delta T_{\rm b}(\boldsymbol{r})$. Throughout the EoR, the mean bubble size $R_{\rm bubble}$ will likely boost the 3PCF at characteristic triangle sizes. Thus, the 3PCF contains information about the physics of reionization (Mcquinn et al. 2007). Similarly, higher order clustering statistics contain information about the physical reionization parameters (see for instance Shimabukuro et al. 2016) which affect the morphology of the $x_{\rm H\,II}(\boldsymbol{r})$ and $\delta T_{\rm b}(\boldsymbol{r})$ fields.

In this section, we train models to predict the mean bubble size $R_{\rm bubble}$ using the 3PCF from simulated data. First, we use 3PCF measurements of the ionization fraction field $x_{\rm H\,II}(\boldsymbol{r})$ to train our models. The resulting model is a useful means of determining whether $\xi^{(3)}$ does indeed contain information about the mean bubble size. In practice, however, the ionization fraction field $x_{\rm H\,II}(\boldsymbol{r})$ is difficult to disentangle from the actual results of interferometer experiments. In the second half of this section, we train models to predict the mean bubble size using simulated $\delta T_{\rm b}(\boldsymbol{r})$ data, which would be directly available from interferometer observations. As well as the data cleaning steps in Sections 3.2.1 and 3.2.2, we also exclude data with global ionization fraction outside the range $0.01 \leq x_{\rm H\,II} \leq 0.95$.

5.1 Results training on $x_{\rm H\,II}(\boldsymbol{r})$ data

In this subsection, we train a model to learn how the mean bubble size $R_{\rm bubble}$ is related to the 3PCF of ionization fraction data $x_{\rm H\,II}(\boldsymbol{r})$. Our training and testing data are from the range of simulated reionization scenarios described in Section 3.2, and we use the MLP model described in Section 4. Fig. 8 shows the measured $x_{\rm H\,II}(\boldsymbol{r})$ 3PCF for a range of redshifts, showing the true mean 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 amplitude of the dimensionless 3PCF seen in Fig. 8 reaches a peak at intermediate scales. Either side of the peak the amplitude decreases, although at larger scales above 20 Mpc the amplitude near the start and end of

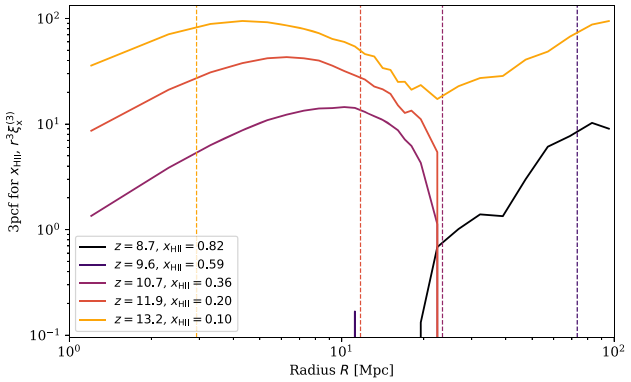

Figure 8. Example measurements of $r^3\xi^{(3)}$ for ionization fraction field data $x_{\rm H\,II}(\boldsymbol{r})$. Each line shows the measured statistic for a single redshift, all taken from a simulation with $\zeta_{\rm ion} = 30.0$, $T_{\rm vir} = 10^4$ K, and $E_0 = 200$ eV. The redshifts and corresponding global ionization fraction are shown for each line in the legend.

Table 1. Rmse performance on unseen testing data using the four different input scaling types in Section 4.4. 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 5.1, but this indicates that the relationship between $r^2\xi^{(3)}$ and the mean bubble size is easier to learn than the other inputs.

Input scaling	Rmse
$\xi^{(3)}$	6.49
$r\xi^{(3)}$	3.59
$r^2\xi^{(3)}$	2.02
$r^3\xi^{(3)}$	2.22

the EoR ($x_{\rm H\,II} = 0.82$ and $x_{\rm H\,II} = 0.10$) show a second rise in the amplitude.

Before running a full hyperparameter search, we first compare four options for different input-scaling types. We train one model for each of the four possible input scaling types, namely $\xi^{(3)}$, $r\xi^{(3)}$, $r^2\xi^{(3)}$, and $r^3\xi^{(3)}$. The MLP models in this section all have the same architecture, namely two hidden layers both containing 100 nodes. The following values are used for the 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 mid-points of the allowed random search ranges or, for categorical choices, as the default parameters suggested by the code authors (Pedregosa et al. 2011). Table 1 shows the resulting overall rmse values for models using each of the four different scaling types. Our results indicate that scaling the 3PCF by r^2 or r^3 generates more accurate models than scaling by r or not scaling at all. Using $\xi^{(3)}$ or $r\xi^{(3)}$ as inputs makes it harder for our MLP models to uncover a relationship between the 3PCF and the mean bubble size. We use $r^2\xi^{(3)}(r)$ as inputs to our models hereafter, as these have the best overall rmse value.

5.1.1 *Best final model*

We now find the best MLP model to predict the mean bubble sizes from the 3PCF of ionization fraction field data $x_{\rm H\,II}(\boldsymbol{r})$. We use the full hyperparameter search method described in Section 4.2, comparing

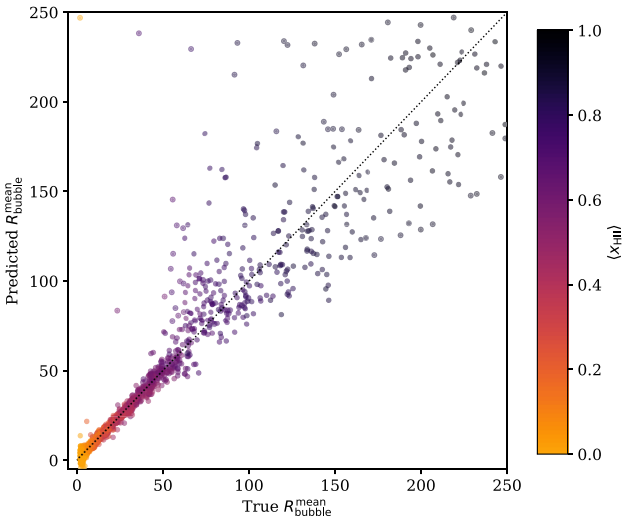

Figure 9. Predicted bubble size versus true bubble size for the best model in Section 5.1.1. These predictions are made on unseen testing data, using only the 3PCF of ionization fraction field data as inputs to the model. The predicted values and true values generally lie along the diagonal for values of $R_{\rm bubble} < 70$Mpc. Larger mean bubble sizes are harder to model and show much lager scatter away from the diagonal, as discussed in the text.

1000 randomly chosen models and selecting the one with best cross-validated performance. The resulting best MLP model uses three hidden layers with sizes [148, 142, 93]; training batch size of 296; a maximum of 563 epochs (of which the model used all epochs before terminating); adaptive learning rate starting at 4.7×10^{-3}; the 'relu' activation function; and L2 regularization parameter 3.9×10^{-4}. Fig. 9 shows the accuracy of the best MLP model's predictions for unseen testing data. We plot all predicted $R_{\rm bubble}$ values as a function of the true values. Marker colours are used to indicate the value of $\langle x_{\rm H\,II}(z) \rangle$ 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. Fig. 12 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 in this subsection has a median prediction error of 10.1 per cent.

The accuracy of the model depends strongly on the magnitude of the true bubble size. The model struggles to make accurate predictions for mean bubble sizes that are larger than 70 Mpc: predictions for $R_{\rm bubble} < 70$ Mpc lie close to the diagonal, but predictions for $R_{\rm bubble} > 70$ Mpc show much larger scatter. This can be understood in terms of the relationship between the 3PCF and the mean bubble size. Near the end of the EoR, the widespread overlap of ionized bubbles gives rise to a larger average mean free path of ionizing photons, but also blurs the definition of a mean bubble size. Many bubbles have merged, and thus the 'mean' bubble size is a less clear feature. The model's ability to learn the mean bubble size from 3PCF measurements reflects this.

5.2 Results training on $\delta T_{\rm b}(\boldsymbol{r})$ data

The situation is more complicated when using measurements of the 21 cm differential brightness temperature field $\delta T_{\rm b}(\boldsymbol{r})$ instead of the ionization fraction field $x_{\rm H\,II}(\boldsymbol{r})$. The relationship between $\delta T_{\rm b}$ and the ionization fraction $x_{\rm H\,II}$ given in equation (7) is assumed to be linear, but the other terms in this equation also impact the morphology

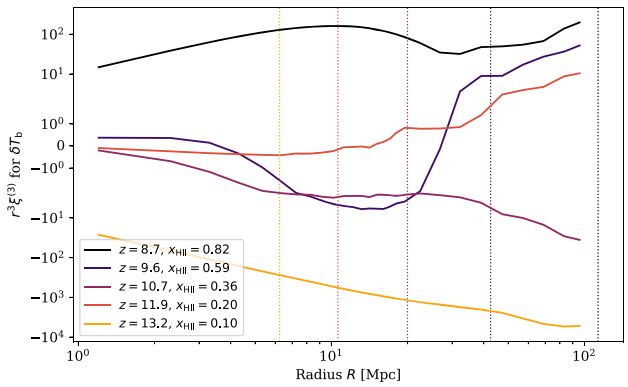

Figure 10. Example measurements of $r^3\xi^{(3)}$ for 21 cm differential brightness temperature field data $\delta T_b(\boldsymbol{r})$, using the same simulation as Fig. 8. These data have also been processed using the radius bins in Section 3.2.1.

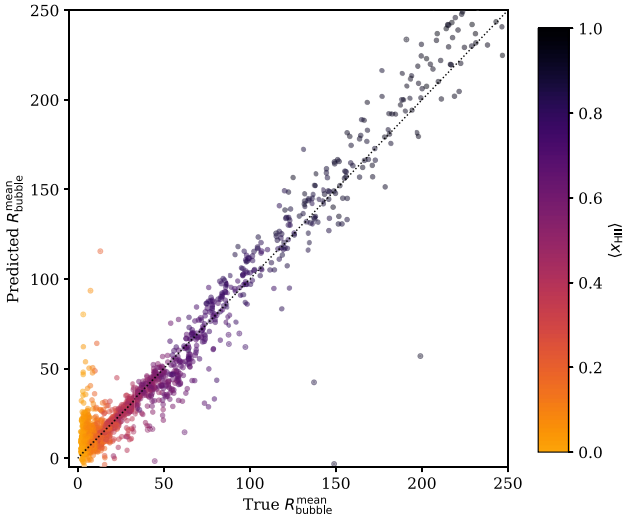

Figure 11. Predicted bubble size versus true bubble size for unseen testing data, using the best model in Section 5.2. This model uses the 3PCF of $\delta T_b(\boldsymbol{r})$ data to predict the mean bubble size. The predicted values and true values generally lie along the main diagonal for middling values of R_{bubble} between 25 and 100 Mpc. The model can accurately predict the mean 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.

of the 21 cm brightness temperature field. Most notably, local spin temperature fluctuations $T_{spin}(\boldsymbol{r})$ and local density fluctuations $\delta(\boldsymbol{r})$ can both change the local values of $\delta T_b(\boldsymbol{r})$. Fluctuations in these values confuse the otherwise simple relationship between the 3PCF and the mean bubble size. Fig. 10 shows the measured $\delta T_b(\boldsymbol{r})$ 3PCF from a simulation with parameters $\zeta_{ion} = 30.0$, $T_{vir} = 10^4$ K, and $E_0 = 200$ eV. The true mean bubble sizes (calculated as the mean of the mean free path distributions) are shown as vertical lines. The 3PCF of the brightness temperature data has a more complex evolution over the EoR than the 3PCF of ionization fraction data shown in Fig. 8. In general, the amplitude of the $\delta T_b(\boldsymbol{r})$ 3PCF decreases until around $\langle x_{HII}(z) \rangle = 0.25$, before increasing to a maximum near the end of the EoR. The complex evolution of other features is less obvious and justifies the need for machine learning models here.

Using the same method as for the ionization fraction field model, we train a model to predict the mean bubble sizes using the 3PCF of simulated $\delta T_b(\boldsymbol{r})$ data. The resulting best model uses three hidden layers with sizes [158, 188, 187]; training batch size of 169; a maximum of 864 epochs; adaptive learning rate starting at 1.3×10^{-3}; the 'relu' activation function; and L2 regularlization parameter 4.3×10^{-3}.

This $\delta T_b(\boldsymbol{r})$ MLP model has a median prediction error of 13.4 per cent. This performance is slightly worse than using $x_{HII}(\boldsymbol{r})$ data, indicating that the extra complexities of including local spin temperature fluctuations and local density field fluctuations do indeed contaminate the relationship between the mean bubble size and the data field correlations. The model cannot distinguish between correlations of ionized regions and correlations of low-density contrast regions ('underdense' regions), because both of these scenarios give rise to lower values for δT_b. Similarly, regions with low local values for the spin temperature T_{spin} can mimic ionized regions.

We plot the $\delta T_b(\boldsymbol{r})$ model's predicted mean bubble sizes for unseen testing data in Fig. 11, as a function of the true mean bubble size. Two features are worth nothing in comparison to the previous $x_{HII}(\boldsymbol{r})$ MLP model. First, although the average performance of the δT_b model is worse, the performance at larger bubble sizes is better. Whereas the $x_{HII}(\boldsymbol{r})$ model's predictions showed a large scatter around the diagonal for $R_{bubble} > 70$ Mpc, the $\delta T_b(\boldsymbol{r})$ model's predictions show a more consistent relationship with the mean bubble size: all predictions with $R_{bubble} > 25$ Mpc are made with a roughly consistent accuracy. In particular for larger bubble sizes, the 3PCF of brightness temperature data appears to encode more information about bubble sizes than does the ionization fraction field. A likely reason for this is the effect of neutral regions. Whereas neutral regions in the ionization fraction field have a uniform value of $x_{HI} = 1.0$, these regions can have different values in the brightness temperature field owing to the other terms in equation (7). This relationship could encode information in the brightness temperature field correlations that does not exist in the ionization fraction field, thus allowing our MLP model to learn the mean bubble size more easily. The second interesting feature is the $\delta T_b(\boldsymbol{r})$ model's poorer performance at low bubble sizes, seen as the large solid cluster of markers in the bottom left of Fig. 11. It is not immediately obvious why this occurs. Including spin temperature fluctuations certainly causes a more complex relationship between the ionization fraction field and the 3PCF of the brightness temperature field. It is possible that this effect is worse at earlier times, when the mean bubble sizes are generally smaller.

Fig. 12 shows the histograms of prediction errors for both final best MLP models: one using $x_{HII}(\boldsymbol{r})$ data, and one using $\delta T_b(\boldsymbol{r})$ data. Ideally, all predictions would be near zero percentage rmse. The distribution of errors for these two models does not depend strongly on which data are used ($x_{HII}(\boldsymbol{r})$ or $\delta T_b(\boldsymbol{r})$) although, as mentioned above, each model has different prediction accuracies for different mean bubble size regimes.

5.3 Effect of modelling weights

Training an MLP model involves finding the optimal 'weight' parameters. These weights are usually initialized to random values as discussed in Section 4.1. Different initial weight values will result in different final weight values at the end of training. Thus, the performance of an MLP model depends on the choice of initial weight values. It is interesting to determine the impact that the choice of initial weight values has on model performance. The black line in Fig. 13 shows the distribution of median prediction errors for a set of 500 models, each of which has different randomized

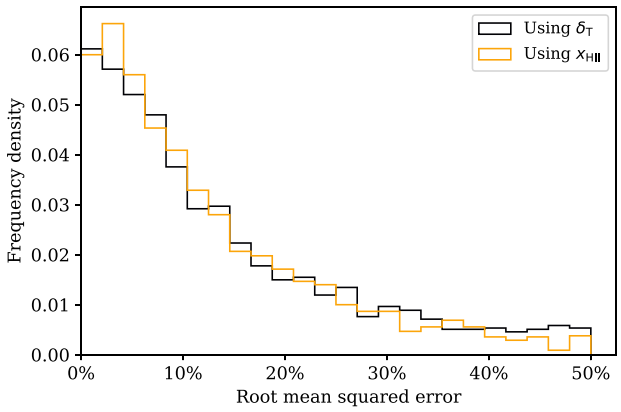

Figure 12. Histogram of prediction errors for mean bubble size models. Visibly, the overall distribution of errors does not depend strongly on which data are used. However, it can be seen from Figs 9 and 11 that the prediction accuracies of these two MLP models depend strongly on the mean bubble size: the $\delta T_b(\boldsymbol{r})$ model makes better predictions at higher bubble sizes, and the $x_{HII}(\boldsymbol{r})$ model makes better predictions at lower bubble sizes.

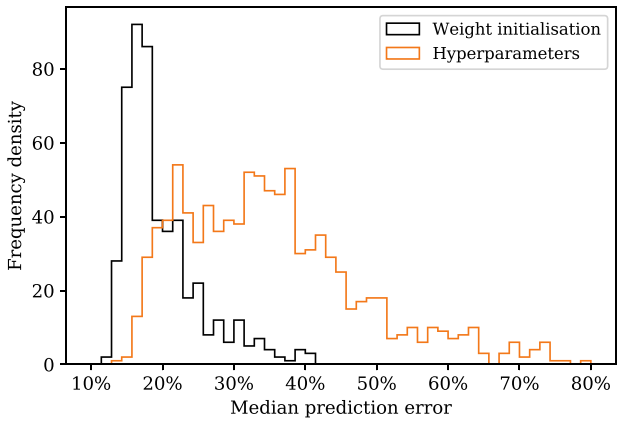

Figure 13. Histogram showing the spread of rmse model performance, either for different weight initializations of the MLP models or from all 1000 models in the hyperparameter. For the 'weight initializations' line, all models have the best hyperparameters as determined in Section 5.2.

initial weights but identical hyperparameters to our best model in Section 5.2. The lighter orange line shows the distribution of median prediction errors for all 1000 MLP models in the full hyperparameter search. For our best model, weight initialization clearly has a strong effect on the model performance: the median prediction error can vary between 10 per cent and 30 per cent. Although, it is possible that our best model is particularly susceptible to different weight initializations, it is likely that other MLP models would also have a similar magnitude of spread in performances. This likely puts an upper limit on the possible performance from any MLP model, even if a deeper hyperparameter search were performed. Note that the best model's rmse value of roughly 13 per cent lies close to the best performance found by varying random initial weights. Our best model has almost certainly benefited somewhat from a 'lucky' weight initialization, in the sense that retraining the MLP model with different initial weights would likely lead to a worse rmse performance. Our model is a good one – it has an acceptable rmse performance on unseen testing data – but a broader investigation

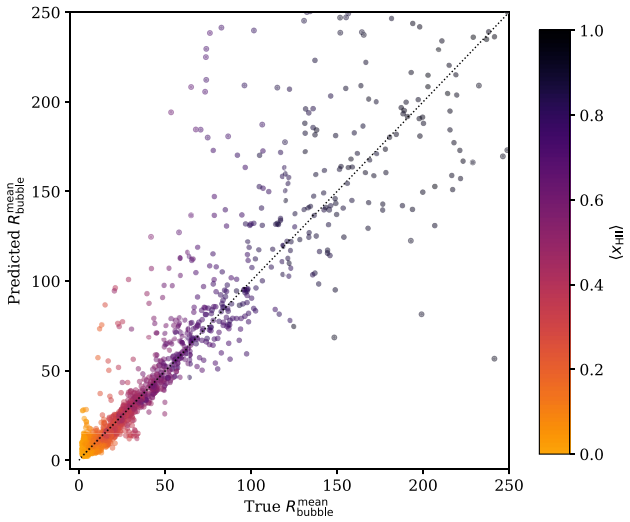

Figure 14. Predicted bubble size versus true bubble size for unseen testing data, using the best model in Section 5.4. This model uses the power spectrum of $\delta T_b(\boldsymbol{r})$ data to predict the mean bubble size. The predicted values and true values generally lie along the main diagonal, although in comparison to the equivalent 3PCF model in Fig. 11 this model makes significantly worse predictions.

into the full hyperparameter space could potentially lead to a higher accuracy model.

5.4 Comparison to power spectrum

The 21 cm line is subject to many sources of noise. In particular, thermal noise in the raw observed data affects our ability to make inferences from 21 cm maps. In order to reduce the effect of noise, statistical quantities such as clustering statistics can be used. These metrics are less affected by noise since they are calculated as averages across the entire map. The 3PCF is a higher order clustering statistic and so should reduce the effect of noise. However, it is also interesting to check whether using a lower order statistic such as the power spectrum provides equally good results. In this section, we use the full hyperparameter search method described in Section 4.2 to find an MLP model that predicts the mean bubble sizes from the power spectrum of the differential brightness temperature field $\delta T_b(\boldsymbol{r})$. As in the previous sections, we compare 1000 randomly chosen models and select the one with best cross-validated performance. The resulting best MLP model uses three hidden layers with sizes [191, 110, 76]; training batch size of 194; a maximum of 596 epochs (of which the model used all epochs before terminating); constant learning rate starting at 4.8×10^{-3}; the 'relu' activation function; and L2 regularization parameter 4.5×10^{-4}.

Fig. 14 shows the accuracy of the best $P(k)$ model's predictions for unseen testing data. This model has a median prediction error of 18.3 per cent, somewhat worse than the equivalent model in Section 5.2 which uses 3PCF measurements instead of power spectrum measurements as inputs to the MLP model. The information encoded in the power spectrum appears to be less strongly related to the mean bubble size than the information encoded in the 3PCF. It is worth noting that if noise was added to the underlying simulated δT_b maps then the performances of the 3PCF models would likely be impacted. More investigation would be needed to determine whether this impact would be greater for the 3PCF model than for the power spectrum models.

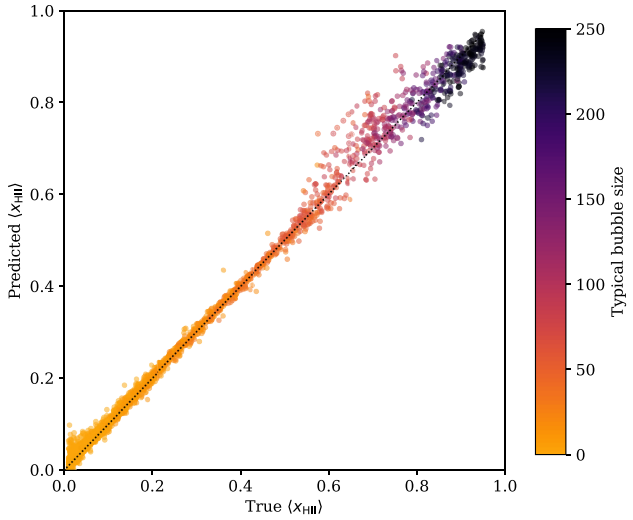

Figure 15. Predicted global ionization fraction versus true global ionization fraction for unseen testing data, using the ionization fraction 3PCF as inputs. The predicted and true values lie very closely along the diagonal, particularly for values $\langle x_{\rm H\,II}(z)\rangle < 0.6$. Predictions for $\langle x_{\rm H\,II}(z)\rangle > 0.6$ are slightly worse as discussed in the text.

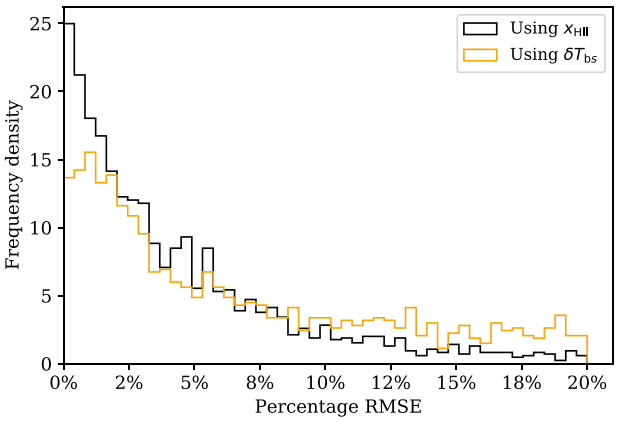

Figure 16. Histogram of prediction errors for predicting the global ionization fraction. Each line shows the histogram of errors for a single model. The model using $x_{\rm H\,II}(\boldsymbol{r})$ 3PCF data has a much more accurate median prediction error (3.6 per cent) than the model using $\delta T_{\rm b}(\boldsymbol{r})$ data (16.0 per cent).

6 LEARNING THE GLOBAL IONIZATION FRACTION FROM THE 3PCF

In the previous section, we investigate using 3PCF measurements to predict the mean bubble size. The mean bubble size is a useful metric for tracking the growth of ionizing regions, but the global ionization fraction $\langle x_{\rm H\,II}(z)\rangle$ is a more direct measurement for the overall progress of the EoR. The redshift history of $\langle x_{\rm H\,II}(z)\rangle$ can be strongly affected by the reionization parameters: different ionizing efficiency $\zeta_{\rm ion}$ scenarios have a different abundance of ionizing photons, which affects the EoR duration; different $T_{\rm vir}$ scenarios have different halo mass function distributions, leading to more or fewer ionizing sources and also affect the EoR duration.

In this section, we train a model to predict the value of $\langle x_{\rm H\,II}(z)\rangle$ from 3PCF measurements. Our models learn the relationship between the 3PCF and global ionization fraction by using the same simulated data in Section 5. Measurements of the 3PCF and bubble size distribution use the methods described in Sections 3.2.1 and 3.2.2, respectively. The data are cleaned using the same ionization fraction filters, namely $0.01 \leq \langle x_{\rm H\,II}(z)\rangle \leq 0.95$.

6.1 Results training on $x_{\rm H\,II}(\boldsymbol{r})$ data

We use the same search strategy as in the previous section. The best 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 per cent. Fig. 15 indicates the performance from this model, showing the predicted values of $\langle x_{\rm H\,II}(z)\rangle$ as a function of the true $\langle x_{\rm H\,II}(z)\rangle$ values in the testing data. Marker colours show the mean bubble size. All markers lie close to the perfect-model diagonal in Fig. 15, confirming that this model makes extremely accurate predictions. As in the previous section, the model accuracy is higher for $\langle x_{\rm H\,II}(z)\rangle < 0.6$ than for $\langle x_{\rm H\,II}(z)\rangle > 0.6$.

Ionization fraction 3PCF measurements have a very strong relationship with the global ionization fraction. Ionization fraction field data contain a range of bubble sizes. 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 ionization fraction. The predictions begin to worsen near the end of the EoR for $\langle x_{\rm H\,II}(z)\rangle > 0.6$, when overlap causes a more complex bubble size distribution. However, the predictions are still visibly good and still have a low rmse value.

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

In this subsection, we train a model to predict the global ionization fraction $\langle x_{\rm H\,II}(z)\rangle$ from $\delta T_{\rm b}(\boldsymbol{r})$ 3PCF data. We use the same search strategy as the previous subsections. The best 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}.

As seen in Fig. 17, predicting the global ionization fraction using $\delta T_{\rm b}(\boldsymbol{r})$ 3PCF data gives less accurate results than using $x_{\rm H\,II}(\boldsymbol{r})$ data. The $\delta T_{\rm b}(\boldsymbol{r})$ model's median prediction error is 16.0 per cent, much worse than the error of 3.6 per cent for the $x_{\rm H\,II}(\boldsymbol{r})$ model. Fig. 16 gives the final prediction histograms for the two global ionization fraction models, using either ionization fraction data $x_{\rm H\,II}(\boldsymbol{r})$ or brightness temperature field data $\delta T_{\rm b}(\boldsymbol{r})$. Predictions of the global ionization fraction depend strongly on which data are used: the prediction errors for the model using $x_{\rm H\,II}(\boldsymbol{r})$ data are much lower than those for the $\delta T_{\rm b}(\boldsymbol{r})$ model.

The model predictions shown in Fig. 17 deviate more widely from the perfect diagonal than the predictions in Fig. 15. Interestingly, this model's accuracy *increases* for the later stages of the EoR with $\langle x_{\rm H\,II}(z)\rangle > 0.6$, as opposed to decreasing the accuracy of the model using $x_{\rm H\,II}(\boldsymbol{r})$ 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 $\delta T_{\rm b}(\boldsymbol{r})$ field at early times than at later times. Thus, the morphology of the $\delta T_{\rm b}(\boldsymbol{r})$ field is more closely linked to that of the $x_{\rm H\,II}(\boldsymbol{r})$ field at later times.

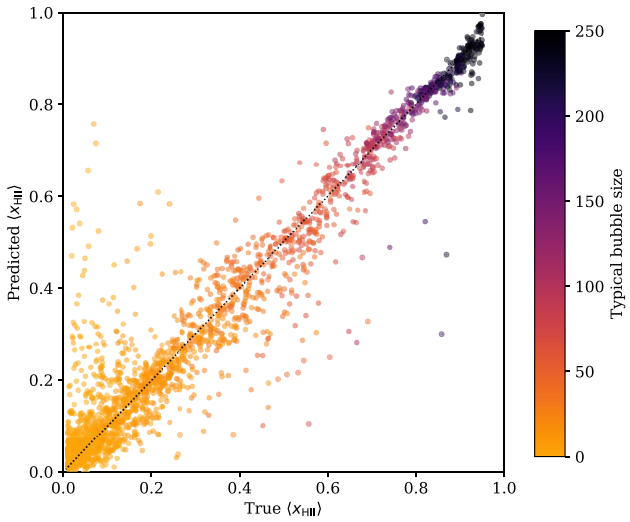

Figure 17. Predicted global ionization fraction versus true global ionization fraction for unseen testing data, using $\delta T_{\rm b}(\boldsymbol{r})$ as inputs. The predictions generally lie along the diagonal, but with larger scatter than using $x_{\rm H\,II}(\boldsymbol{r})$ as model inputs.

7 CONCLUSIONS

The 3PCF of the 21 cm signal contains valuable information about the morphology and history of the EoR. We present an optimized code for estimating the 3PCF of 3D pixelized data, such as the outputs from seminumerical simulations. The code includes jackknifing for error estimates and user-changeable parameters for choosing a level of approximate sampling. We test the code on a testing distribution with known analytical 3PCF, finding that the estimates from our code match the true 3PCF closely. After testing, we use our code to calculate the 3PCF for a range of simulated reionization scenarios using 21CMFAST. Throughout, we assume an idealized case where instrumental noise is negligible, and 21 cm foregrounds have been perfectly removed.

We use machine learning techniques and train models to recover both the typical bubble size and the global ionization fraction from measured 3PCF outputs of seminumerical simulations. We first train models to recover the typical bubble size, from the 3PCF of either ionization fraction data or 21 cm 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. The model using $x_{\rm H\,II}(\boldsymbol{r})$ 3PCF data has better performance at small bubble sizes ($1 < R_{\rm bubble} < 70$ Mpc, whereas the model using $\delta T_{\rm b}(\boldsymbol{r})$ has better performance for larger bubble sizes ($R_{\rm bubble} > 25$ Mpc). Both features can be understood in terms of how the data field morphologies evolve over the EoR. We compare the performances of predict the typical bubble size using either the 3PCF or the power spectrum. We find that using the 3PCF instead of the power spectrum leads a noticeable improvement in the final MLP model's prediction accuracy, with median prediction accuracies of around 10 per cent and 14 per cent, respectively.

We then train a model to recover the global ionization fraction from ionization fraction 3PCF data. The resulting model has extremely accurate predictions and shows the three-point clustering of $x_{\rm H\,II}(\boldsymbol{r})$ data is strongly related to the evolution of the global ionization fraction. Our model is able to uncover this relationship with median prediction accuracy of 4 per cent, although the predictions are slightly less accurate for the later stages of the EoR with $\langle x_{\rm H\,II}(z) \rangle > 0.6$. Unfortunately, this model would practically not be useful in EoR analysis because the ionization fraction field is difficult to probe directly. Instead, observations are made in terms of the differential brightness temperature. We train a fourth and final model to predict the global ionization fraction from the 3PCF of the differential brightness temperature field. This MLP model has a median prediction accuracy of 16 per cent. The resulting model makes accurate predictions for the late stages of the EoR ($\langle x_{\rm H\,II}(z) \rangle > 0.6$), but struggles with the early stages.

As with all machine learning projects, our models to predict the typical bubble size and global ionization fraction could likely be improved by gathering more data from a wider range of reionizaion scenarios. This would allow the models to learn more general connections between the 3PCF measurements and characteristic reionization features. Providing other brightness temperature field summary statistics could also improve our models, for instance the distribution of pixel brightnesses (Ichikawa et al. 2009) or the size distribution of bright regions (Kakiichi et al. 2017). We also note that our 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. Further studies will be necessary to evaluate the effectiveness of such an approach in the presence of instrumental effects and noise, as well as foreground residuals.

The techniques in this paper are tested on simulated data. We have assumed that instrumental noise is negligible at our scales and lower redshifts of operation, as is expected during the EoR upcoming experiments such as the SKA (Koopmans et al. 2015). Instrumental smoothing will predominantly affect smaller scale features on the same scale as the instrument's point spread function, and the effect on larger scale features would be minimal. While as noted by Watkinson et al. (2019), the bispectrum of Gaussian noise is zero, there will be noise and possibly bias on both the 3PCF and the bispectrum due to sample variance, instrumental systematic effects, ionispheric effects, finite number of baselines, restricted field of view, and radio frequency interference. All of these will need to be considered in future studies. Furthermore, we have assumed a best-case scenario where 21 cm foregrounds have been perfectly removed. This assumption is not uncommon in recent literature (see e.g. Shimabukuro & Semelin 2017; Gillet et al. 2018; Jennings et al. 2018) but remains the subject of much discussion. Several studies (Chapman et al. 2014; Mertens, Ghosh & Koopmans 2018; Li et al. 2019) have claimed that foreground removal can be effective for the power spectrum. Watkinson, Trott & Hothi (2020) show that foregrounds could be a problem for recovering the 21 cm bispectrum. More work would be needed to understand the impact of foreground residuals on the 3PCF signal.

There are several other possible avenues of future work to build on these 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, using a larger selection of triangle configurations (both sizes and shapes) would likely provide more information and make it easier to recover the bubble size statistics. Thirdly, training models to map from 3PCF measurements directly to parameters in a similar way to Shimabukuro & 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. Finally, investigating the effect of realistic experiment conditions would indicate whether the 3PCF of future 21 cm measurements could be used to extract physically meaningful bubble size statistics.

This work presents the first attempt to predict fundamental properties of the EoR using the 3PCF and machine learning techniques. We provide a publicly available code 3PCF-FAST to help the community perform similar analyses in the future.

ACKNOWLEDGEMENTS

WDJ was supported by the Science and Technology Facilities Council (ST/M503873/1) and from the European Community through the DEDALE grant (contract no. 665044) within the H2020 Framework Program of the European Commission. CAW's research is supported by a UK Research and Innovation Future Leaders Fellowship, grant number MR/S016066/1. However, the research presented in this paper was carried out with financial support from the European Research Council under ERC grant number 638743-FIRSTDAWN (held by Jonathan Pritchard). FBA acknowledges support from the DEDALE grant, from the UK Science and Technology Research Council (STFC) grant ST/M001334/1, and from STFC grant ST/P003532/1.

DATA AVAILABILITY STATEMENT

The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

Agarwal S., Abdalla F. B., Feldman H. A., Lahav O., Thomas S. A., 2014, MNRAS, 439, 2102
Ali Z. S. et al., 2015, ApJ, 809, 61
Aristizabal Sierra D., Fong C. S., 2018, Phys. Lett. B, 784, 130
Barkana R., 2018, Nature, 555, 71
Becker G. D., Rauch M., Sargent W. L. W., 2007, ApJ, 662, 72
Bowman J. D., Rogers A. E. E., Monsalve R. A., Mozdzen T. J., Mahesh N., 2018, Nature, 555, 67
Bromm V., Yoshida N., Hernquist L., McKee C. F., 2009, Nature, 459, 49
Chapman E., Zaroubi S., Abdalla F., Dulwich F., Jelić V., Mort B., 2014, preprint (arXiv:1408.4695)
Das A., Mesinger A., Pallottini A., Ferrara A., Wise J. H., 2017, MNRAS, 469, 1166
Datta K. K., Jensen H., Majumdar S., Mellema G., Iliev I. T., Mao Y., Shapiro P. R., Ahn K., 2014, MNRAS, 442, 1491
DeBoer D. R. et al., 2017, PASP, 129, 045001
Ewall-Wice A., Chang T.-C., Lazio J., Doré O., Seiffert M., Monsalve R. A., 2018, ApJ, 868, 63
Fialkov A., Barkana R., Cohen A., 2018, Phys. Rev. Lett., 121, 011101
Fraser S. et al., 2018, Phys. Lett. B, 785, 159
Furlanetto S. R., Oh S. P., Briggs F. H., Peng Oh S., Briggs F. H., 2006, Phys. Rep., 433, 181
Gaztañaga E., Scoccimarro R., 2005, MNRAS, 361, 824
Ghara R., Choudhury T. R., Datta K. K., 2015, MNRAS, 447, 1806
Gillet N., Mesinger A., Greig B., Liu A., Ucci G., 2019, MNRAS, 484, 282
Giri S. K., D'Aloisio A., Mellema G., Komatsu E., Ghara R., Majumdar S., 2019, J. Cosmol. Astropart. Phys., 2019, 058
Glorot X., Bengio Y., 2010, in Teh Y. W., Titterington D. M.., eds, Proc. Machine Learning Research, Understanding the Difficulty of Training Deep Feedforward Neural Networks. JMLR.org, p. 249
Gnedin N. Y., 2014, ApJ, 793, 29
Gorce A., Pritchard J. R., 2019, MNRAS, 489, 1321
Greig B., Mesinger A., 2015, MNRAS, 449, 4246
Greig B., Mesinger A., 2017a, MNRAS, 472, 2651
Greig B., Mesinger A., 2017b, MNRAS, 465, 4838
Gunn J. E., Peterson B. A., 1965, ApJ, 142, 1633
Hassan S., Davé R., Finlator K., Santos M. G., 2017, MNRAS, 468, 122
Hoffmann K., Mao Y., Xu J., Mo H., Wandelt B. D., 2019, MNRAS, 487, 3050
Hutter A., 2018, MNRAS, 477, 1549
Hutter A., Watkinson C. A., Seiler J., Dayal P., Sinha M., Croton D. J., 2019, MNRAS, 492, 653
Ichikawa K., Barkana R., Iliev I. T., Mellema G., Shapiro P. R., 2010, MNRAS, 406, 2521
Jennings W. D., Watkinson C. A., Abdalla F. B., McEwen J. D., 2019, MNRAS, 483, 2907
Kakiichi K. et al., 2017, MNRAS, 471, 1936
Kern N. S., Liu A., Parsons A. R., Mesinger A., Greig B., 2017, ApJ, 848, 23
Kingma D. P., Ba J., 2014, preprint (arXiv:1412.6980)
Koopmans L. et al., 2015, Proc. Sci., The Cosmic Dawn and Epoch of Reionisation with SKA. SISSA, Trieste, PoS#1
Kramer R. H., Haiman Z., Oh S. P., 2006, ApJ, 649, 570
Kravtsov A. V., Klypin A. A., Khokhlov A. M., 1997, ApJS, 111, 73
Lambiase G., Mohanty S., 2020, MNRAS, 494, 5961
Landy S. D., Szalay A. S., 1993, ApJ, 412, 64
Lawson K., Zhitnitsky A., 2019, Phys. Dark Universe, 24, 100295
Li W. et al., 2019, MNRAS, 485, 2628
Lupton R. H., Gunn J. E., Szalay A. S., 1999, AJ, 118, 1406
Majumdar S. et al., 2015, MNRAS, 456, 2080
Majumdar S., Pritchard J. R., Mondal R., Watkinson C. A., Bharadwaj S., Mellema G., 2018, MNRAS, 476, 4007
McKay M. D., Beckman R. J., Conover W. J., 1979, Technometrics, 21, 239
Mcquinn M., Lidz A., Zahn O., Dutta S., Hernquist L., Zaldarriaga M., 2007, MNRAS, 377, 1043
Mellema G., Iliev I. T., Alvarez M. A., Shapiro P. R., 2006, New Astron., 11, 374
Mellema G. et al., 2013, Exp. Astron., 36, 235
Mertens F. G., Ghosh A., Koopmans L. V. E., 2018, MNRAS, 478, 3640
Mesinger A., Furlanetto S., 2007, ApJ, 669, 663
Mesinger A., Furlanetto S., Cen R., 2011, MNRAS, 411, 955
Moroi T., Nakayama K., Tang Y., 2018, Phys. Lett. B, 783, 301
Muller M. E. E. M., 1959, Commun. ACM, 2, 19
Muñoz J. B., Loeb A., 2018, Nature, 557, 684
Park J., Mesinger A., Greig B., Gillet N., 2019, MNRAS, 484, 933
Patil A. H. et al., 2017, ApJ, 838, 65
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Planck Collaboration XLVII, 2018, A&A, 596, A108
Pober J. C., Greig B., Mesinger A., 2016, MNRAS, 463, L56
Press W. H., Schechter P., 1974, ApJ, 187, 425
Rumelhart D. E., Hinton G. E., Williams R. J., 1986, Nature, 323, 533
Santos M. G., Ferramacho L., Silva M. B., Amblard A., Cooray A., 2010, MNRAS, 406, 2421
Schmit C. J., Pritchard J. R., 2018, MNRAS, 475, 1213
Semelin B., Combes F., Baek S., 2007, A&A, 474, 365
Sheth R. K., Mo H. J., Tormen G., 2001, MNRAS, 323, 1
Shimabukuro H., Semelin B., 2017, MNRAS, 468, 3869
Shimabukuro H., Yoshiura S., Takahashi K., Yokoyama S., Ichiki K., 2016, MNRAS, 468, 3003
Shimabukuro H., Yoshiura S., Takahashi K., Yokoyama S., Ichiki K., 2017, MNRAS, 568, 1542
Sikivie P., 2018, Phys. Dark Univ., 24, 100289
Sims P. H., Pober J. C., 2019, MNRAS, 492, 22
Tingay S. J. et al., 2013, Publ. Astron. Soc. Aust., 30, e007
Watkinson C. A., Majumdar S., Pritchard J. R., Mondal R., 2017, MNRAS, 472, 2436
Watkinson C. A., Giri S. K., Ross H. E., Dixon K. L., Iliev I. T., Mellama G., Pritchard J. R., 2019, MNRAS, 482, 2653
Watkinson C. A., Trott C. M., Hothi I., 2020, preprint (arXiv:2002.05992)
Werbos P., 1974, PhD thesis, Harvard University
Wouthuysen S. A., 1952, AJ, 57, 31
Yang Y., 2018, Phys. Rev. D, 98, 103503
Yoshiura S., Takahashi K., Takahashi T., 2018, Phys. Rev. D, 98, 063529
Zahn O., Lidz A., McQuinn M., Dutta S., Hernquist L., Zaldarriaga M., Furlanetto S. R., 2006, ApJ, 654, 12
Zeldovich Y., 1970, A&A, 5, 84

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