MNRAS **479,** 3405–3414 (2018)
Advance Access publication 2018 June 29

doi:10.1093/mnras/sty1719

Machine learning cosmological structure formation

Luisa Lucie-Smith,[1★] Hiranya V. Peiris,[1,2] Andrew Pontzen[1] and Michelle Lochner[1,3,4]

[1] *Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
[2] *The Oskar Klein Centre for Cosmoparticle Physics, Stockholm University, AlbaNova, SE-106 91 Stockholm, Sweden*
[3] *African Institute for Mathematical Sciences, 6 Melrose Road, Muizenberg 7945, South Africa*
[4] *SKA SA, The Park, Park Road, Cape Town 7405, South Africa*

Accepted 2018 June 21. Received 2018 June 12; in original form 2018 February 19

ABSTRACT

We train a machine learning algorithm to learn cosmological structure formation from *N*-body simulations. The algorithm infers the relationship between the initial conditions and the final dark matter haloes, without the need to introduce approximate halo collapse models. We gain insights into the physics driving halo formation by evaluating the predictive performance of the algorithm when provided with different types of information about the local environment around dark matter particles. The algorithm learns to predict whether or not dark matter particles will end up in haloes of a given mass range, based on spherical overdensities. We show that the resulting predictions match those of spherical collapse approximations such as extended Press–Schechter theory. Additional information on the shape of the local gravitational potential is not able to improve halo collapse predictions; the linear density field contains sufficient information for the algorithm to also reproduce ellipsoidal collapse predictions based on the Sheth–Tormen model. We investigate the algorithm's performance in terms of halo mass and radial position and perform blind analyses on independent initial conditions realizations to demonstrate the generality of our results.

Key words: methods: statistical – galaxies: haloes – dark matter – large-scale structure of Universe.

1 INTRODUCTION

Dark matter haloes are the fundamental building blocks of cosmic large-scale structure, and galaxies form by condensing in their cores. Understanding the structure, evolution, and formation of dark matter haloes is an essential step towards understanding how galaxies form and ultimately to test cosmological models. However, this is a difficult problem due to the highly non-linear nature of the haloes' dynamics. Dark matter haloes originate from random perturbations seeded in the early Universe and grow via mass accretion and mergers with smaller structures throughout their assembly history. *N*-body simulations provide the only practical tool to compute non-linear gravitational effects starting from an initial random field (e.g. Springel, Yoshida & White 2001; Springel 2005; Kuhlen, Vogelsberger & Angulo 2012).

Analytic approximations of structure formation yield useful physical interpretations of these detailed numerical studies. Generally, analytic techniques assume dark matter collapse occurs once the smoothed linear density contrast exceeds a threshold value. Combined with excursion set theory, this ansatz provides a tool to analytically predict the final halo mass of an initially overdense region. This can be used to infer useful quantities such as the abundance of dark matter haloes in the Universe, or the halo mass function, based on properties of a Gaussian random field alone (Press & Schechter 1974; Bond et al. 1991; Bond & Myers 1996). The halo mass function is the quantity most often used to assess the accuracy of different analytic frameworks against numerical simulations. The original form of the halo mass function proposed by Press & Schechter (1974), although qualitatively correct, is known to underestimate the abundance of the most massive haloes, and overestimate the abundance of the less massive ones. The need for precision mass functions led to modifications of the original halo mass function in the form of parametric functions calibrated with cosmological simulations (Jenkins et al. 2001; Reed et al. 2003; Tinker et al. 2008). Pure analytic extensions of the excursion set ansatz have also been constructed that yield better agreement with numerical simulations (Sheth, Mo & Tormen 2001; Maggiore & Riotto 2010; Paranjape & Sheth 2012; Farahi & Benson 2013; Borzyszkowski, Ludlow & Porciani 2014). Given these successful predictions, the excursion set description has become an accepted physical interpretation of the process of structure formation itself.

We present a machine learning approach to learn cosmological structure formation directly from *N*-body simulations. The machine learning algorithm is trained to learn the relationship between the initial conditions and final halo population that results from

[★] E-mail: luisa.lucie-smith.15@ucl.ac.uk

Published by Oxford University Press on behalf of the Royal Astronomical Society

non-linear evolution. Using the resulting initial conditions-to-haloes mapping, we aim to provide new physical insights into the process of dark matter halo formation, and compare with existing interpretations gained from widely investigated analytic frameworks. In contrast to existing analytic theories, our approach does not require prior assumptions about the physical process of halo collapse; the haloes' non-linear dynamics is learnt directly from *N*-body simulations rather than approximated by an excursion set model in the presence of a collapse threshold.

We provide the machine learning algorithm with a set of informative properties about the dark matter particles extracted from the initial conditions. Machine learning algorithms are sufficiently flexible to include a wide range of initial conditions properties that may contain relevant information about halo formation, without changing the training process of the algorithm. We choose these properties to be aspects of the initial density field in the local surroundings of the dark matter particles' initial position. By quantifying their impact on the learning accuracy of the algorithm, we can investigate which aspects of the early universe density field contain relevant information on the formation of dark matter haloes. The trained initial conditions-to-haloes mapping can then also be used to predict the mapping for new initial conditions, without the need to run a further simulation.

The highly non-linear nature of dark matter evolution makes it a problem well suited to machine learning. Machine learning is a highly efficient and powerful tool to learn relationships that are too complex for standard statistical techniques (Witten et al. 2016). In the context of structure formation, machine learning techniques have also been shown to be effective, for example, in learning the relationship between dark and baryonic matter from semi-analytic models (Kamdar, Turk & Brunner 2016; Agarwal, Davé & Bassett 2018; Nadler et al. 2018).

We choose *random forests* (Breiman et al. 1984; Breiman 2001), a popular algorithm that has been shown to outperform other classifiers in many problems (Niculescu-Mizil & Caruana 2005; Caruana & Niculescu-Mizil 2006; Douglas et al. 2011; Lochner et al. 2016). Random forests also lend themselves to physical interpretation, as they provide measures that allow the user to infer which of the inputs are predominantly responsible for the learning outcomes of the algorithm. Random forests are ensembles of decision trees, each following a set of simple decision rules to predict the class of a sample (Ball & Brunner 2010). The prediction of the random forest is given by the average of the probabilistic predictions of the individual trees, where the variance of the forest predictions is greatly reduced compared to that of a single tree.

To apply this approach, we must turn the process of dark matter evolution into a supervised classification problem. We chose to focus on the simplest case of a binary classification task to illustrate the approach and allow for a cleaner understanding of the physics behind the learning process of the algorithm. We distinguish between dark matter particles that end up in haloes of mass above a threshold, and those that belong either to lower mass haloes or to no halo at all. This defines two classes: the former set of particles belongs to the *IN haloes* class, while the latter forms the *OUT haloes* class. The machine learning algorithm is trained to predict whether the dark matter particles in the initial conditions will end up in IN class haloes or in the OUT class at $z = 0$. The training is performed on an existing *N*-body simulation where we already know the associated halo for each particle (if any).

The predictive accuracy of the algorithm crucially depends on the choice of features extracted from the initial conditions and used as input to the machine learning algorithm. We first train the random forest with the initial linear density field as features and subsequently add information on the tidal shear field. We are able to quantify the physical relevance of such properties in the halo collapse process, based on their respective impact on the classification performance of the random forest. Our results demonstrate the utility of machine learning in gaining insights into the physics of structure formation, as well as providing a fast and efficient classification tool.

The paper is organized as follows. We present an overview of the classification pipeline and describe how we extract features from the linear density field and train the machine learning algorithm in Section 2. In Section 3, we interpret the classification output and present our results in Section 4. We then extend the feature set to include the tidal shear field in Section 5 and discuss the resulting implications. We study the algorithm's performance as a function of halo properties in Section 6. We perform two blind tests of our pipeline on independent simulations in Section 7, demonstrating the generality of our results, and finally conclude in Section 8.

2 METHOD

We trained and tested the random forest with an existing dark-matter-only simulation produced with P-GADGET-3 (Springel et al. 2001; Springel 2005) and a 5-year *Wilkinson Microwave Anisotropy Probe* (*WMAP*5) Λ cold dark matter (ΛCDM) cosmological model (Dunkley et al. 2009): $\Omega_\Lambda = 0.721$, $\Omega_\mathrm{m} = 0.279$, $\Omega_\mathrm{b} = 0.045$, $\sigma_8 = 0.817$, $h = 0.701$, and $n_\mathrm{s} = 0.96$. The comoving softening length of the simulation is $\epsilon = 25.6\,\mathrm{kpc}$. The simulations evolve 256^3 dark matter particles, each of mass $M_\mathrm{particle} = 8.24 \times 10^8\,\mathrm{M}_\odot$, in a box of comoving size $L = 50\,h^{-1}\,\mathrm{Mpc}$ from $z = 99$ to 0.[1]

The haloes were identified using the SUBFIND halo finder (Springel et al. 2001), a friends-of-friends method with a linking length of 0.2, with the additional requirement that particles in a halo be gravitationally bound. While SUBFIND also identifies substructure within haloes, we consider the entire set of bound particles to make up a halo and do not subdivide them further. The simulation contains 18 801 haloes at $z = 0$, ranging from masses of $\sim 10^9$ to $\sim 10^{14}\,\mathrm{M}_\odot$.

We used the final snapshot ($z = 0$) to label each particle with its corresponding class. At $z = 0$, we split the dark matter particles between two classes: *IN haloes* and *OUT haloes*. We chose the IN class to contain all particles in haloes of mass $M \geq 1.8 \times 10^{12}\,\mathrm{M}_\odot$ at $z = 0$ (401 haloes), and the OUT class to contain all remaining particles, including those in haloes of mass $M < 1.8 \times 10^{12}\,\mathrm{M}_\odot$ and those that do not belong to any halo.[2] This choice was made in order to split the haloes into the two classes at an intermediate scale within the mass range probed by the simulation. Our pipeline allows the selection of any mass threshold that would ultimately allow us to extend the binary classification to a multiclass one.

Each particle, with its associated class label, was traced back to the initial conditions ($z = 99$) where we extracted features to be used as input for the random forest as described below. The random forest was trained based on these input features and the known output class for a training subset of particles. We tested the algorithm using the remaining dark matter particles, where the random forest's class prediction was compared to their respective

[1] We make use of the PYTHON package PYNBODY (Pontzen et al. 2013) to analyse the information contained in the simulation snapshots.

[2] The mass scale $M = 1.8 \times 10^{12}\,\mathrm{M}_\odot$ corresponds to the mass of a particular halo of the simulation and was chosen as the class boundary for convenience.

true class label. The robustness of the algorithm was tested further on independent *N*-body simulations (Section 7).

2.1 Density field features

Most machine learning algorithms, including random forests, require a *feature extraction* process to extract key properties of the dark matter particles. The classification performance crucially depends on whether or not the chosen features provide meaningful information to allow for a clean separation between the IN and OUT classes.

We extracted machine learning features from the linear density field. This choice was motivated by the work of Press & Schechter (1974, PS) who developed a model to predict the (comoving) number density of dark matter haloes as a function of mass based on properties of the linear density field. The ansatz is that a Lagrangian patch will collapse to form a halo of mass M at redshift z if its linear density contrast exceeds a critical value $\delta_c(z)$. An improved theoretical footing for PS theory was developed by Bond et al. (1991) based on the excursion-set formalism, known as extended Press–Schechter (EPS). The crucial assumption is that the final halo mass corresponds to the matter enclosed in the *largest* possible spherical region with density contrast $\delta_L = \delta_c$. This method yields a halo mass function qualitatively consistent with numerical simulations, suggesting that a useful mapping between Lagrangian regions and final collapsed haloes can be obtained from spherical overdensities. This motivates our choice of machine learning features from the initial linear density field as follows.

We smoothed the density contrast $\delta(\boldsymbol{x}) = [\rho(\boldsymbol{x}) - \bar{\rho}]/\bar{\rho}$, where $\bar{\rho}$ is the mean matter density of the universe, on a smoothing scale R,

$$\delta(\boldsymbol{x}; R) = \int \delta\left(\boldsymbol{x}'\right) W_{\mathrm{TH}}\left(\boldsymbol{x} - \boldsymbol{x}'; R\right) \mathrm{d}^3x', \tag{1}$$

where $W_{\mathrm{TH}}(\boldsymbol{x}, R)$ is a real space top-hat window function:

$$W_{\mathrm{TH}}(\boldsymbol{x}, R) = \begin{cases} \dfrac{3}{4\pi R^3} & \text{for } |\boldsymbol{x}| \leq R, \\ 0 & \text{for } |\boldsymbol{x}| > R. \end{cases} \tag{2}$$

The convolution (1) was carried out in Fourier space, which naturally accounts for the periodicity of simulations. A window function $W(\boldsymbol{x}, R)$ of characteristic radius R corresponds to a mass scale $M_{\mathrm{smoothing}} = \bar{\rho}V(R)$, where in the case of a top-hat window function $V_{\mathrm{TH}}(R) = 4/3\pi R^3$. The feature for machine learning then consists of the density contrast smoothed with a top-hat window function of mass scale $M_{\mathrm{smoothing}}$ (or, smoothing scale R) centred on the particle's position in the initial conditions.

We repeated the smoothing for 50 mass scales evenly spaced in $\log M$ within the range allowed by the volume and resolution of the simulation box, i.e. $3 \times 10^{10} \leq M_{\mathrm{smoothing}}/\mathrm{M}_{\odot} \leq 1 \times 10^{15}$, yielding a set of 50 features per particle. We found that using a larger number of smoothing scales did not yield improvement in the classification performance, meaning that 50 smoothing scales were sufficient to capture the relevant information carried by the density field.

In the context of excursion set theory, the density contrast of a particle as a function of smoothing scale is known as a *density trajectory*. Fig. 1 shows examples of density trajectories of particles belonging to the true IN and OUT classes. The trajectories describe whether particles are found in overdense or underdense regions as a function of increasing mass scale. As one approaches the largest mass scales probed by the simulation box, the trajectories start to converge to $\delta(x, \infty) = 0$, where the density coincides with the mean density of the Universe. The ensemble of trajectories constitutes the full feature set we used to first train then test the random forest.

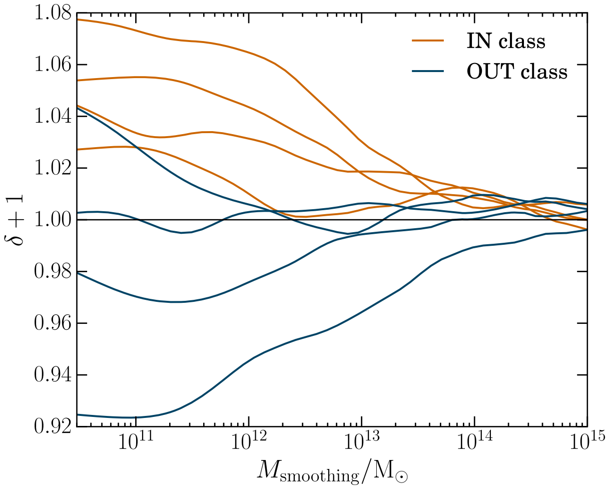

Figure 1. Examples of density trajectories corresponding to particles belonging to the IN and OUT classes. The linear density field is smoothed with a real space top-hat filter centred on each particle's initial position. We calculate the smoothed overdensity δ as the smoothing mass scale M is increased.

2.2 Training the random forest

We make use of the random forest implementation in the SCIKIT-LEARN (Pedregosa et al. 2011) PYTHON package. The random forest was trained using a set of 50 000 randomly selected particles from the simulation, each carrying its own set of density features and corresponding IN or OUT class label. The size of the training set was chosen to form a subset of particles representative of the full simulation box. To test for representativeness, we checked the performance of the algorithm for training sets of different sizes and found no improvement for training sets larger than 50 000 particles. Therefore, we concluded that 50 000 randomly selected particles are sufficient to form a training set representative of the full simulation box. The remaining particles in the simulation were used as a test set; the trained random forest predicts the class label of the particles in the test set, which is then compared to the particles' true labels to assess the algorithm's performance. Note also that random forests are robust to correlated features (Breiman 2001), meaning that the high correlation present in our density features does not affect the predictive performance of the algorithm.

Like most machine learning algorithms, random forests have hyperparameters that need to be optimized for a given training set. These include the number of trees and the maximum depth of the forest, the maximum number of particles at the end node of a tree, and the size of the subset of features to select at a node split. We used a grid search algorithm combined with k-fold cross-validation (Kohavi 1995) to optimize the random forest's hyperparameters. In k-fold cross-validation, the training set is divided into k equally sized sets where $k - 1$ sets are used for training and one is used as a validation set, on which the algorithm is tested. This procedure is repeated k times so that each set is used as a validation set once. For each validation set we evaluate a score based on a chosen scoring metric [here we use the area under the receiver operating characteristic (ROC) curve, see Section 3] and average scores over all k validation sets to obtain the final score of a training set. Here,

Table 1. Confusion matrix for two classes: positives and negatives. We use this to quantify the performance of the machine learning algorithm, where the positives are particles of the IN class and the negatives are particles of the OUT class.

		True class	
		P	N
Predicted class	P	True positive (TP)	False positive (FP)
	N	False negative (FN)	True negative (TN)

we performed a fivefold cross-validation for all combinations of hyperparameters and retained the combination that achieved the best score.

3 INTERPRETING THE CLASSIFICATION OUTPUT

A random forest (like most machine learning algorithms) outputs a probabilistic measure of belonging to a class for every particle. For practical use this must be mapped onto a concrete class for each particle. Many approaches exist for such a mapping; we choose to consider different probability thresholds at which a particle is considered to belong to a class. A high probability threshold will contain a very pure sample of particles but also will be incomplete. As the probability threshold decreases, one allows for a more complete set of particles at the expense of including misclassified ones.

Once the probability-to-class mapping is established, we quantify the performance of the algorithm making use of a confusion matrix for binary classification problems as shown in Table 1. Throughout this analysis we always take the positives to be particles of the IN class and negatives to be particles of the OUT class. The perfect classifier consists of true positives and true negatives only. A more realistic classifier will include a number of incorrectly classified particles: misclassified positives fall in the false negative category, yielding a loss of *completeness*, and misclassified negatives fall in the false positive category, yielding an increase in *contamination*. We measure the true positive rate (TPR), the ratio between the number of particles correctly classified as positives and the total number of positives in the data set,

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{3}$$

and the false positive rate (FPR), the ratio between the number of particles incorrectly classified as positives and the total number of negatives in the data set,

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{4}$$

ROC curves (Green & Swets 1966; Hilden 1991; Fawcett 2006) are a tool to graphically represent the balance between completeness and contamination at various probability thresholds. A ROC curve compares the true positive rate to the false positive rate as a function of decreasing probability threshold. As one lowers the probability threshold, one allows for a more complete set of IN particles (increase in true positive rate) at the expense of a larger contamination of misclassified particles (increase in false positive rate). The area under the curve (AUC) of a ROC curve is a useful quantity to compare classifiers. The perfect classifier would have an AUC of 1, whereas a random assignment of classes would obtain an AUC of 0.5. Typically, algorithms are considered to be performing well if AUC ≥0.8.

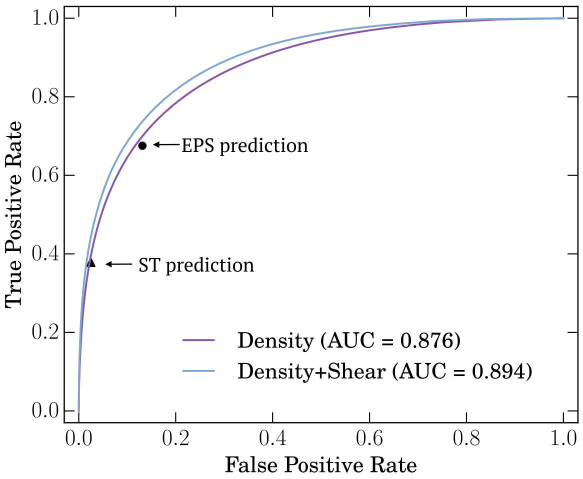

Figure 2. ROC curves for the density feature set and the combined shear and density feature set. The machine learning algorithm is able to learn the information contained in the density trajectories to match the EPS prediction. The Sheth–Tormen (ST) prediction represents an extension of standard excursion set developed by Sheth & Tormen (1999), which adopts a moving collapse barrier motivated by tidal shear effects. The comparison between the two ROC curves shows little improvement in the test set classification once information on the shear field is added. The ST analytic prediction also does not provide an overall improvement compared to the EPS prediction; the false positive rate (or, contamination) decreases at the expense of decreasing the true positive rate (or, completeness). The machine learning algorithm is able to recover the ST analytic prediction when presented with information on the density field alone by altering the probability threshold.

We use ROC curves and AUCs to evaluate and compare the performance of the random forest for different feature sets (Sections 4 and 5), different halo mass and radial position ranges (Section 6), and different simulations (Section 7).

4 DENSITY FIELD CLASSIFICATION

Fig. 2 shows the ROC curve for the density feature set resulting from classifying all particles in the simulation that were not used for training the random forest. The random forest achieves an AUC score of 0.876.

In order to assess whether machine learning can learn as much as human-constructed models, we wish to compare its performance to existing theories. In particular, the EPS formalism motivated our choice of density features and has been demonstrated to infer approximately correct number densities of collapsed haloes from a Gaussian random field (Bond et al. 1991). Although EPS is commonly used to predict the dark matter halo mass function, we make use of it to predict an independent set of class labels for the test set particles and compare their accuracy to that of the machine learning predictions.

Following EPS, the fraction of haloes of mass M are equivalent to the fraction of density trajectories with a first upcrossing of the density threshold barrier δ_{th} at mass scale M. We take the density threshold to be the spherical collapse threshold adopted by Bond et al. (1991): $\delta_{\mathrm{th}}(z) = (D(z)/D(0))\delta_{\mathrm{sc}}$, where $\delta_{\mathrm{sc}} \approx 1.686$. The predicted halo mass of each particle is given by the smoothing mass scale of the particle's first upcrossing. We then assign to each particle an IN or OUT label depending on whether its predicted halo mass falls in the mass range of the IN or OUT class. We emphasize that the labels inferred from the EPS framework are independent from the predictions of the random forest.

We plot in Fig. 2 the resulting true positive rate and false positive rate inferred from the EPS predicted labels and find that the EPS prediction lies on the ROC curve of the random forest. In other words, the random forest is able to 'learn' EPS and the EPS results correspond to a ∼42 per cent probability threshold on the ROC curve. Machine learning adds the flexibility to trade contamination for completeness along the ROC curve as we vary the probability threshold. Instead, EPS results in a single point in true positive rate–false positive rate space since it gives a single prediction for each particle rather than a probability associated with a class.

4.1 Physical interpretation

The algorithm's performance depends on whether or not the input features contain relevant information to separate particles between classes. For example, the ideal feature would split a set of particles into two pure sets, each containing only particles of one class. By contrast, irrelevant features are not able to distinguish between classes, yielding a poor class separation in the two resulting sets. Therefore, we can determine which features contain the most information in mapping particles into the correct halo mass range, based on their ability to separate classes when training the random forest.

There are many metrics designed to measure the relevance of the inputs to a machine learning algorithm; here we use *feature importances* (Louppe et al. 2013). The importance of a feature X is a weighted sum of the impurity decrease[3] at all nodes t where the feature is used, averaged over all trees T in the forest:

$$\mathrm{Imp}(X) = \frac{1}{N_\mathrm{T}} \sum_{T} \sum_{t \in T} p(t) \Delta i(t), \tag{5}$$

where N_T is the number of trees, $p(t)$ is the fraction of particles reaching node t, and $\Delta i(t)$ is the impurity decrease, i.e. the difference in entropy between the parent node and the child nodes.

We calculate the relative importances in the density feature set to find the most relevant features in distinguishing between the IN and OUT classes. Fig. 3 shows the relative importance of each density feature as a function of its smoothing mass scale. The importances are normalized such that the sum of all importances is 1 and the errors are computed by training the random forest multiple times, each with a randomly drawn set of training particles. The largest halo mass in the simulation is marked by a grey line. We find that most of the information lies in mass ranges of 10^{12}–10^{13} $\mathrm{M}_{\odot}$, just above the boundary between the IN and OUT classes.

5 ADDING THE TIDAL SHEAR TENSOR

Peaks in Gaussian random fields are inherently triaxial (Doroshkevich 1970; Bardeen et al. 1986). Therefore, extensions of the standard spherical model were made in order to incorporate the dynamics of ellipsoidal collapse. The impact of the tidal shear on properties of collapsed regions has been extensively studied (Bond & Myers 1996; Sheth & Tormen 1999; Sheth et al. 2001). Sheth & Tormen (1999, ST) have studied how ellipsoidal collapse modifies the mass function of dark matter haloes in the excursion set formalism. Spheres are distorted into an ellipsoid due to tidal shear effects and the collapse time of a halo therefore depends explicitly on the ellipticity and prolateness of the tidal shear field.

We extended the original density feature set to incorporate additional information on the local tidal shear field around particles. We studied the impact on the halo classification performance and quantified the shear's relevance in the training process via the feature importances. The advantage of studying tidal shear effects with machine learning is that these can be straightforwardly translated into features and used as input to the same machine learning algorithm. On the other hand, analytic models usually require incorporating approximations to the tidal shear within the excursion set formalism. In general, any potentially relevant physical property can be added in the form of a feature without adding complexity to the algorithm.

We will first describe how we constructed features from the tidal shear field, then present the classification results of the full density and shear feature sets.

5.1 Tidal shear features

The deformation tensor is given by the Hessian of the gravitational potential:

$$D_{ij} = \frac{\partial^2 \Phi}{\partial x_i \partial x_j}, \tag{6}$$

where $\Phi(\mathbf{x})$ is the peculiar gravitational potential at position $\mathbf{x}$ and is related to the density contrast via Poisson's equation $\nabla^2 \Phi = \delta$.

The ordered eigenvalues of D_{ij}, $\lambda_1 \geq \lambda_2 \geq \lambda_3$, can be reparametrized in terms of the ellipticity, e, and prolateness, p (Bond & Myers 1996):

$$e = \frac{\lambda_1 - \lambda_3}{2\delta}, \tag{7}$$

$$p = \frac{\lambda_1 - 2\lambda_2 + \lambda_3}{2\delta}, \tag{8}$$

where $\lambda_1 + \lambda_2 + \lambda_3 = \delta$ and δ is the smoothed overdensity used as a density feature. In order to minimize redundancy between the features, we removed the density dependence from the ellipticity and prolateness. We computed the eigenvalues of the traceless deformation tensor, known as the tidal shear tensor, $t_i = \lambda_i - \delta/3$, now satisfying $t_1 + t_2 + t_3 = 0$. The ellipticity and prolateness in terms of the traceless eigenvalues t_i take the form

$$e_\mathrm{t} = t_1 - t_3, \tag{9}$$

$$p_\mathrm{t} = 3\,(t_1 + t_3)\,. \tag{10}$$

For each particle we assigned two new features e_t and p_t evaluated at each smoothing mass scale. Therefore, the original 50-dimensional feature set of density contrasts was augmented to a 150-dimensional feature set given by the density contrast, ellipticity, and prolateness. To test the robustness of random forests to a high-dimensional feature space, we used principal component analysis (PCA) to reduce the 150-dimensional feature set to a 10-dimensional space retaining 98 per cent of the information contained in the original feature set. We found identical predictive performance, meaning that random forests are robust to a 150-dimensional feature set.

5.2 Results

The ROC curve of the density and shear feature set is overplotted in Fig. 2. We find that adding information on the tidal shear tensor

[3] We use Shannon entropy to measure the impurity at a node $i_\mathrm{E}(t) = -\sum_{i=1}^{c} p(j, t) \log_2 p(j, t)$, where $p(j, t)$ is the proportion of particles that belong to class j at node t and c is the total number of classes.

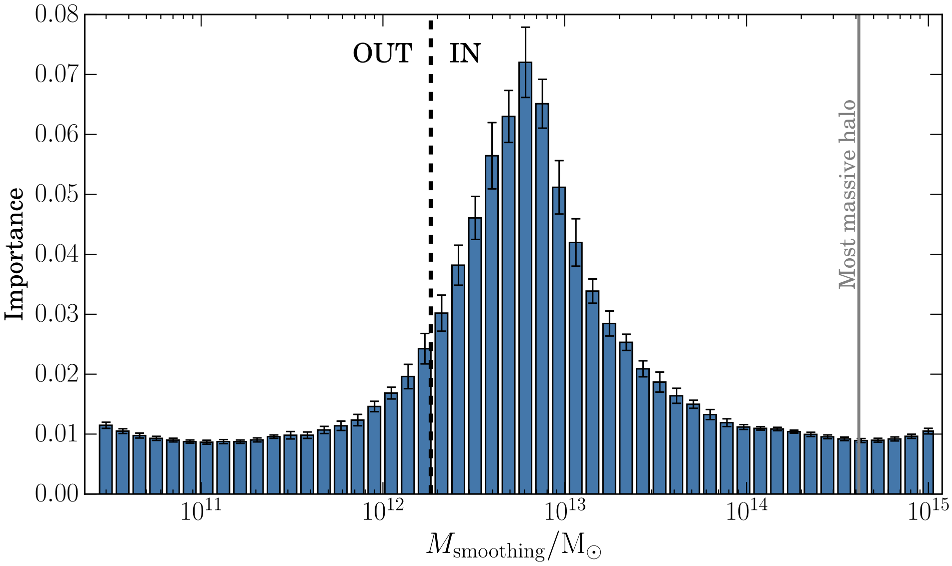

Figure 3. The importance ranking of the density features, shown as a function of their smoothing mass scales. The most relevant information in the training of the random forest comes from the density contrast smoothed at mass scales 10^{12}–10^{13} $\mathrm{M}_\odot$ scales, within the mass range of the IN class haloes. The largest halo mass in the simulation is marked by a grey line.

shows little improvement compared to the case of the density-only feature set. We find an improvement of only 2 per cent in the AUC of the ROC curve. Fig. 4 demonstrates the low impact of the shear features in the classification process. The three panels show the relative importance in the training process of the random forest of the density, ellipticity, and prolateness features as a function of smoothing mass scales. The most relevant features are the density contrasts smoothed on mass scales in the range 10^{12}–10^{13} $\mathrm{M}_\odot$, similar to what was found in the case of the density-only feature set (Fig. 3). The distributions of the density importances in the two feature sets are consistent despite minor variations in the peak and variance of the distributions. The changes are due to the change in the range of hyperparameters when increasing the dimensionality of the feature set from 50 to 150 features. The ellipticity and prolateness have low feature importance scores confirming that the information they contain is irrelevant to the training process of the machine learning algorithm compared with that of the density field.

As with the density feature set, we can compare the machine learning predictions to existing analytic predictions based on the same set of properties of the initial conditions. The ST formalism provides a prescription to predict the final halo mass of a particle based on the density field and the shear field, which we can use to compare to the machine learning output.

ST accounts for the effect of the shear field in the context of the excursion set formalism by adopting a moving collapse barrier rather than the spherical collapse barrier adopted by Bond et al. (1991). The ST collapse barrier $b(z)$ varies as a function of the mass variance $\sigma^2(M)$ and is given by

$$b(z) = \sqrt{a}\delta_{\rm sc}(z)\left[1 + \left(\beta\frac{\sigma^2(M)}{a\delta^2_{\rm sc}(z)}\right)^\gamma\right], \qquad (11)$$

where $\delta_{\rm sc}(0) \approx 1.686$, the parameters $\beta = 0.485$ and $\gamma = 0.615$ incorporate an approximation to ellipsoidal dynamics, and $a = 0.707$ is a normalization constant. These values are the best-fitting parameters found in Sheth et al. (2001). The predicted halo mass of each particle follows the excursion-set framework as for the EPS case; the largest mass scale at which the particle's trajectory upcrosses the collapse barrier in equation (11) gives the predicted halo mass.

The triangle labelled 'ST prediction' in Fig. 2 shows the true and false positive rates predicted by ST. In our study, the ST formalism does not yield an absolute improvement to EPS theory; the false positive rate decreases at the expense of a decrease in the true positive rate. Therefore ST predicts a less contaminated but more incomplete set of IN class particles compared to EPS, corresponding to a probability threshold of 73 per cent on the ROC curve. We find that the random forest is able to reproduce the ST result with both the density-only feature set and the shear and density feature set. This shows that there is sufficient information in the density field for the random forest to match the analytic ST prediction.

Overall, we find that shear effects do not contain additional physical information to improve the classification output of the random forest. The learning process of the algorithm is predominantly driven by the local overdensity around dark matter particles and unaffected by the surrounding tidal shear. The analytic ST prediction, interpreted as an improvement to standard EPS due to the inclusion of tidal shear effects, can be reproduced by the random forest when trained on the density field only. In conclusion, these results show that the physical processes leading to dark matter halo formation for our choice of mass scale splitting the two classes are insensitive to tidal shear effects in the initial conditions.

6 CLASSIFICATION DEPENDENCE ON HALO MASS AND RADIAL POSITION

We now investigate how properties of particles such as the position within a halo and the halo mass affect the accuracy of classification when the algorithm is trained on density features only. To do this we split the test particles into categories based on their radial and halo mass properties to study their respective classification performance.

First, we subdivided particles of the IN class into three mass ranges: particles in *cluster*-sized haloes ($1 \times 10^{14} \leq M_{\rm halo}/{\rm M}_\odot \leq 4 \times 10^{14}$), particles in *group*-sized haloes ($1 \times 10^{13} \leq M_{\rm halo}/{\rm M}_\odot < 1 \times 10^{14}$), and particles in *galaxy*-sized haloes ($1.2 \times 10^{12} \leq$

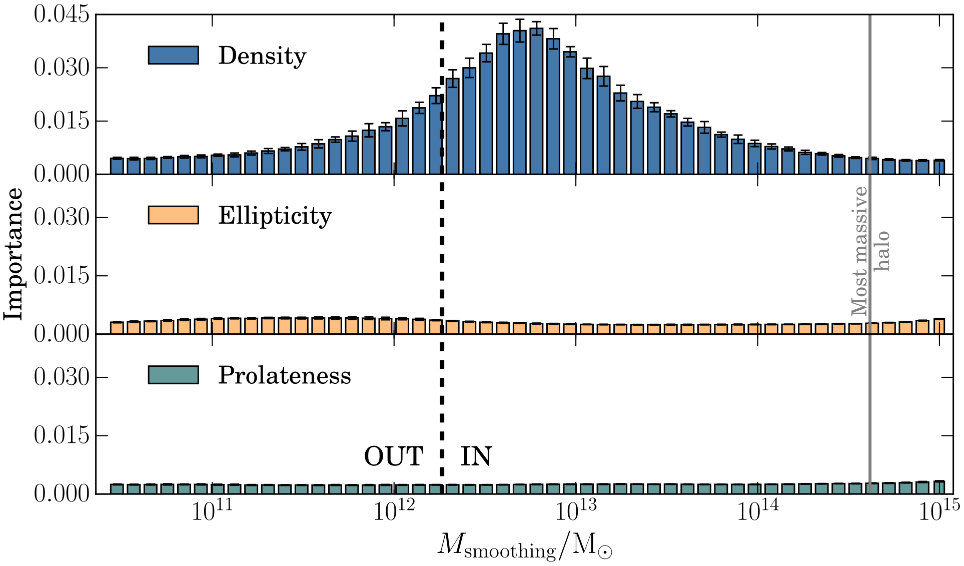

Figure 4. Relative importance of the density features (upper panel), ellipticity features (middle panel), and prolateness features (lower panel) in the full shear and density feature set. The density features are more relevant than the ellipticity and prolateness features. This confirms that the shear field adds little information in distinguishing whether particles will collapse in haloes of mass above the class boundary mass scale or not, compared with the density field.

$M_{\rm halo}/{\rm M}_{\odot} < 1 \times 10^{13}$). We combined each of these subsets in turn with all the OUT particles to form three distinct test sets.

The ROC curves for the three mass range categories of haloes are shown in the right-hand panel of Fig. 5, where the ROC curve of the full original test set is shown for comparison (dashed line). We find that particles in cluster-sized haloes reach an AUC of 0.913, whilst particles in group-sized haloes and galaxy-sized haloes are increasingly more difficult to classify. We overplotted the ST (triangles) and EPS (dots) predictions for each halo mass category of particles, again showing results consistent with those of the machine learning algorithm.

It is likely that the decrease in performance as a function of halo mass is a result of the choice of mass scale used to split haloes into classes, $M = 1.8 \times 10^{12}\,{\rm M}_{\odot}$. This was a necessary step in order to define the two classes of the binary classification problem. Haloes of mass just above and below the IN/OUT mass boundary belong to different classes although they originate from Lagrangian regions with similar properties reflecting their similarity in mass. Therefore, the closer the haloes of different classes are in mass, the harder it is for the random forest to distinguish whether their particles belong to one class or the other. Fig. 6 further demonstrates that haloes of mass approaching the IN/OUT mass boundary from above and below contain a larger fraction of misclassified particles. In the upper (lower) panel, we show the false positive (negative) rate, i.e. the ratio of misclassified OUT (IN) particles over all particles contained in each halo mass bin, for four different probability thresholds. The true halo mass of each particle is shown on the horizontal axis in terms of its distance from the IN/OUT mass boundary. We find that the false positive and negative rates increase for particles in haloes of mass approaching the IN/OUT mass boundary.

We next investigated possible correlations between the particles' position within the haloes and the random forest's classification performance. Here, we subdivided particles of the true IN class into three radial ranges, subject to their radial position in the halo with respect to the halo's virial radius $r_{\rm vir}$. We defined particles in the *inner radial* range ($r/r_{\rm vir} \leq 0.3$), particles in the *mid radial* range ($0.3 < r/r_{\rm vir} \leq 0.6$), and particles in the *outer radial* range ($0.6 < r/r_{\rm vir} \leq 1$). Similar to the mass range study, each subset of haloes was combined with all the OUT class particles from the original set to form three distinct sets.

The left-hand panel of Fig. 5 shows the ROC curves for the three radial categories, together with that of the original test set again shown for comparison (dashed line). Particles in the innermost regions of haloes are the best classified by the random forest, achieving an AUC of 0.937 that is greater than that obtained when classifying *all* particles in the simulation. The classification performance of the random forest decreases as we move from the halo's centre-of-mass towards the virial radius.

We first tested whether the decrease in performance when classifying particles of the outer radial range was due to underrepresentativeness in the training set. Indeed, if the training particles of the outer radial range are not representative of the entire simulation, the classifier's performance on the outer radial range test set would be strongly affected. To test this, we retrained the machine learning algorithm with a training set containing equal number of particles for each radial range category. We found identical ROC curves and AUCs as in the left-hand panel of Fig. 5, therefore excluding the possibility that the higher misclassification rate of outer radial range particles is due to non-representativeness in the training set.

One other possible reason may be that particles living in outer regions of haloes are more likely to have been affected by late-time halo mergers, tidal stripping, or accretion events. Therefore, the final halo mass prediction for such particles is the result of a more complicated dynamical history involving these late-time effects. Conversely, particles near the halo's centre-of-mass are less sensitive to the halo's assembly history and their final halo mass prediction correlates more strongly with the local overdensity in the initial conditions. This hypothesis could be verified by adding features sensitive to the particles' dynamical history (for instance a particle's initial distance to the nearest density peak) and testing whether this information improves the classification of particles located at the boundary of the halo's virial region. In addition to this, the further particles are from the centre of haloes, the closer they are to the boundary between the IN and OUT classes, where particles

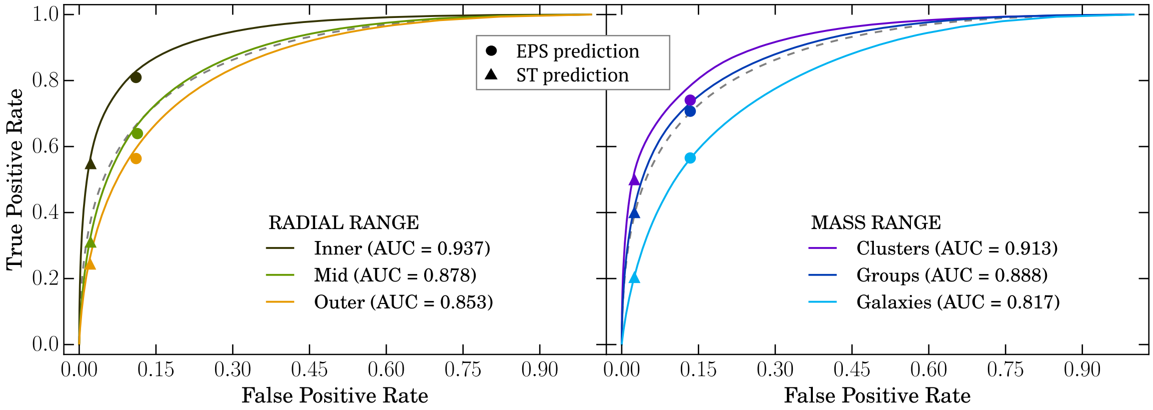

Figure 5. Left-hand panel: the IN class particles are split into inner ($r/r_{\rm vir} \leq 0.3$), mid ($0.3 < r/r_{\rm vir} \leq 0.6$), and outer ($0.6 < r/r_{\rm vir} \leq 1$) radial ranges according to their distance from the centre of the halo. The ROC curves for each category show that the classification performance improves for particles closer to the halo's centre of mass. Right-hand panel: the IN class particles are split into cluster-sized ($1 \times 10^{14} \leq M_{\rm halo}/{\rm M}_{\odot} \leq 4 \times 10^{14}$), group-sized ($1 \times 10^{13} \leq M_{\rm halo}/{\rm M}_{\odot} < 1 \times 10^{14}$), and galaxy-sized ($1.2 \times 10^{12} \leq M_{\rm halo}/{\rm M}_{\odot} < 1 \times 10^{13}$) haloes, and the ROC curves show the random forest's performance in classifying each category. Particles in higher mass haloes are increasingly better classified by the random forest. The ROC curve of the full test set of particles is shown as a dashed line in both panels for comparison. The EPS and ST predictions, labelled by dots and triangles respectively, are also overplotted for each halo mass and radial position category.

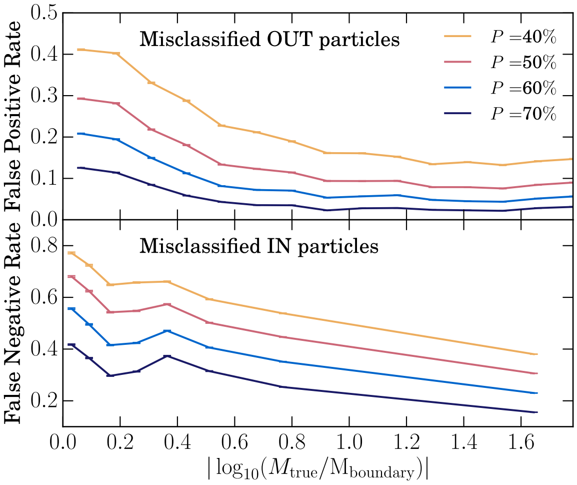

Figure 6. Fraction of misclassified particles in haloes of each mass bin range, where the halo mass bins are labelled as a function of their distance from the IN/OUT boundary mass scale. The upper (lower) panel shows the fraction of misclassified OUT (IN) particles, i.e. the false positive (negative) rate in each mass bin. We consider four distinct probability thresholds for assigning a particle's (IN or OUT) class, where higher thresholds imply lower contamination. The misclassification rate increases as the true mass approaches the classification boundary for all choices of the completeness-to-contamination trade-off.

are harder to classify for the machine learning algorithm. This also translates into a larger uncertainty in the halo mass prediction for particles at the edge of haloes compared to those in the innermost regions of haloes. As a result, the overall uncertainty in the halo mass predictions of centre-of-mass particles is smaller than for particles in the outskirts of haloes. This result is also consistent with excursion set predictions, where ST demonstrated that centre-of-mass particles provide a better estimate of the final halo mass compared to inferences made from the full ensemble of particles in the simulation. To confirm this, we overplotted the EPS (dots) and ST (triangles) predictions for the three radial test sets in the left-hand panel of Fig. 5, demonstrating that analytic formalisms also perform increasingly well for particles that are close to the halo's centre-of-mass. The machine learning algorithm again shows its ability to match the excursion set predictions at fixed probability thresholds for each radial range category.

For completeness, we also explored the misclassification rate of OUT particles that do not belong to any halo. We find that overall these particles have very low misclassification rates compared to particles in haloes. For example, if we consider probability thresholds of 70, 60, 50, and 40 per cent to assign particles to the IN class (as in the upper panel of Fig. 6), the fraction of misclassified over all particles that do not belong to haloes are 2.45, 4.3, 6.58, and 10.11 per cent, respectively. Therefore, the OUT particles predicted by the random forest form a highly pure and complete set.

In conclusion, we find that the best classified categories of particles are those that are further away from the classification boundary, both in terms of mass and radius: particles in the most massive and least massive haloes in the simulation; particles in the innermost regions of haloes; and those furthest away in voids. We further tested whether the addition of the tidal shear information could improve the classification performance of poorly classified particles, such as those in the outskirts of haloes and in galaxy-sized haloes. We find no significant improvement in the classification performance of such particles, other than the 2 per cent improvement found for the whole ensemble and reflected in each mass and radial category.

7 BLIND TESTS ON INDEPENDENT SIMULATIONS

Up to this point we have trained and tested the machine learning algorithm on a single dark-matter-only simulation. To test whether the machine learning algorithm trained on one simulation also gives robust results for different N-body simulations without retraining, we performed blind tests of our pipeline on two independent simulations from the one used for training.

The first independent test simulation (W test) is a different realization of the same *WMAP5* ΛCDM cosmology adopted in the training simulation, for a box of also same size and resolution (see Section 2). The second independent test simulation (P test) is a realization of a different cosmological model, a *Planck* ΛCDM

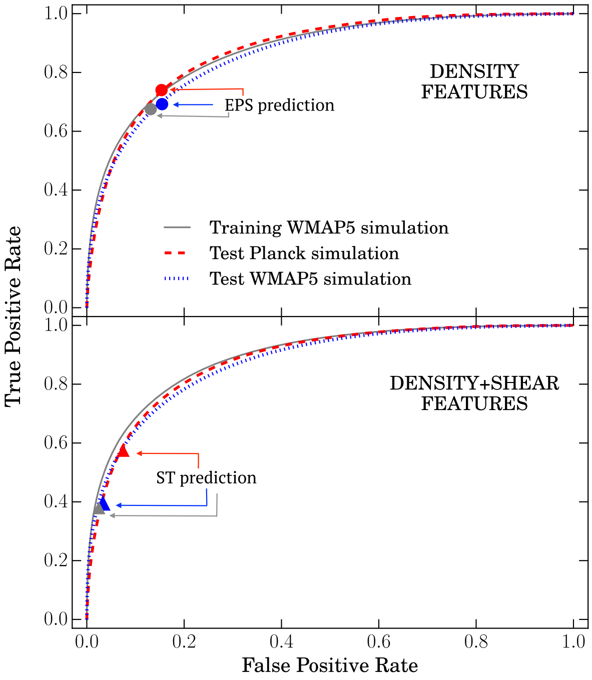

Figure 7. We perform a blind test of the trained machine learning algorithm on two independent *N*-body simulations: a different realization of the *WMAP*5 cosmology used in the training simulation, and a realization of a *Planck* cosmological model. The ROC curves are consistent in all three simulations for both the density feature set and the density and shear feature set, with differences in the AUCs of order ∼1 per cent. The EPS and ST predictions in each simulation match the machine learning performance at different probability thresholds, such that the ST formalism always predicts a less contaminated but more incomplete set of IN particles. These blind tests demonstrate the robustness of the results from a machine learning algorithm trained on one simulation, and applied to different realizations of the same cosmology or realizations of different cosmologies.

cosmology[4] (Planck Collaboration XIII 2016) in a box of comoving size $L = 50$ Mpc containing $N = 512^3$ particles. Moreover, in the *P*-test simulation we identify haloes at $z = 0$ using the AMIGA Halo Finder (AHF; Gill, Knebe & Gibson 2004; Knollmann & Knebe 2009), instead of the SUBFIND halo finder used in both the training simulation and the *W*-test simulation. This allows us to simultaneously test the sensitivity of the machine learning algorithm to the choice of halo finder. For each test simulation, we extracted the input features from the initial conditions and used the pre-trained machine learning algorithm to predict the class labels of the simulations' dark matter particles.

In Fig. 7, we compare the performance of the machine learning algorithm for the independent *W*- and *P*-test simulations with that of the test set of particles in the training simulation. The upper panel shows the ROC curves obtained from predictions based on the density features only, whilst the lower panel shows the case of density and shear features. The machine learning algorithm produces consistent ROC curves in all three simulations for both feature sets. The *P*-test simulation yields a difference in AUC with the training simulation of 0.2 per cent for the density-only feature set and 1.1 per cent for the density and shear feature set. For the *W*-test simulation, the AUC difference with the training simulation is of 1.3 per cent for the density-only feature set and 1.6 per cent for the density and shear feature set. Such differences between the test and training simulations are consistent with uncertainties in the AUC due to statistical noise.

The EPS and ST predicted labels are calculated from the first upcrossings of each simulation's respective particles' trajectories. In all three simulations, the machine learning algorithm is able to match the analytic predictions at different probability thresholds, such that the ST formalism consistently predicts a less contaminated but more incomplete set of IN class particles. For the *W*-test simulation, the EPS and ST predictions match the machine learning predictions at probability thresholds of 41.5 and 74.5 per cent, respectively, differing only slightly to the 42.8 and 74.7 per cent probability thresholds of the training simulation. For the *P*-test simulation, the match to the EPS and ST predictions is found at the lower probability thresholds of 40 and 56 per cent, respectively. This is because the change in cosmological parameters in the *Planck* simulation results in a slightly lower EPS collapse barrier and a significantly lower ST collapse barrier compared to those in a *WMAP*5 cosmological setting. Therefore, trajectories in the *P*-test simulation upcross the collapse barriers at larger smoothing mass scales, resulting in more complete but also less pure sets of predicted IN particles. The change in completeness and contamination is such that both the ST and EPS predictions still match the machine learning ROC curves of the *P*-test simulation, but for lower probability thresholds than the *WMAP*5 simulations.

We conclude that the mapping learnt by the algorithm on one simulation can be generalized to different simulations based on the same or different cosmological parameters, without the need for retraining, and that the results are insensitive to simulation settings.

8 CONCLUSIONS

We have presented a machine learning approach to investigate the physics of dark matter halo formation. We trained the algorithm on *N*-body simulations, from which it learns to predict whether regions of an initial density field later collapse into haloes of a given mass range. This generated a mapping between the initial conditions and final haloes that would result from non-linear evolution, without the need to adopt halo collapse approximations. Our approach provided new physical insight into halo collapse, in particular in understanding which aspects of the initial linear density field contain relevant information on the formation of dark matter haloes.

We provided the algorithm with a set of properties describing the local environment around dark matter particles. By studying the performance of the algorithm in response to different inputs, insights can be gained into the physics relevant to dark matter halo formation. When the algorithm was trained on spherical overdensities from the linear density field, we found that it matched predictions based on EPS theory. When providing the algorithm with additional information on the tidal shear field (motivated by ellipsoidal collapse approximations), the classification performance of the machine learning was not enhanced. We showed that, for the mass threshold considered in our classification problem, the ST ellipsoidal collapse model can be recovered from spherical overdensities alone, with predictions that differ from those of EPS theory only in the completeness-to-contamination trade-off. By performing blind analyses of our pipeline, we confirmed the generality of our results for independent initial conditions realizations and variations in cosmological parameters. We conclude that the linear density field contains sufficient information to predict the formation of dark

[4] The cosmological parameters are $\Omega_\Lambda = 0.6914$, $\Omega_\mathrm{m} = 0.3086$, $\Omega_\mathrm{b} = 0.045$, $\sigma_8 = 0.831$, $h = 0.6727$, and $n_\mathrm{s} = 0.96$.

matter haloes at the accuracy of existing spherical and ellipsoidal collapse analytic frameworks.

While the focus of this paper has been on the density field and tidal shear field, any additional property of interest can be extracted from the initial conditions and used as input to the same machine learning algorithm. This allows for straightforward extensions of the present work to investigate the physics of dark matter halo formation further. Future work could also extend the binary classification problem presented in this work into multiclass classification or regression problems. Potential applications of such an extended framework include a new approach to obtaining a halo mass function, which can be directly tested against existing fitting formulae adopted by analytic approaches. More sophisticated machine learning algorithms such as deep learning offer the ability to learn from the training data that features are the most relevant to cosmological structure formation, and future work will investigate their suitability for structure formation studies.

ACKNOWLEDGEMENTS

LL-S thanks Nina Roth for providing one of the simulations used in this work and for useful discussions. LL-S was supported by the Science and Technology Facilities Council. HVP was partially supported by the European Research Council (ERC) under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement number 306478-CosmicDawn. AP was supported by the Royal Society. ML acknowledges support from the SKA, NRF, and AIMS. This work was partially enabled by funding from the UCL Cosmoparticle Initiative.

REFERENCES

Agarwal S., Davé R., Bassett B. A., 2018, MNRAS, 478, 3410

Ball N. M., Brunner R. J., 2010, Int. J. Modern Phys. D, 19, 1049

Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, ApJ, 304, 15

Bond J. R., Myers S. T., 1996, ApJS, 103, 1

Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440

Borzyszkowski M., Ludlow A. D., Porciani C., 2014, MNRAS, 445, 4124

Breiman L., 2001, Machine Learning, 45, 5

Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984, Classification and Regression Trees. Wadsworth, Belmont, CA.

Caruana R., Niculescu-Mizil A., 2006, in ICML '06: Proceedings of the 23rd International Conference on Machine Learning. ACM, New York, p. 161

Doroshkevich A. G., 1970, Afz, 6, 581

Douglas P., Harris S., Yuille A., Cohen M. S., 2011, Neuroimage, 56, 544

Dunkley J. et al., 2009, ApJS, 180, 306

Farahi A., Benson A. J., 2013, MNRAS, 433, 3428

Fawcett T., 2006, Pattern Recognition Lett., 27, 861

Gill S. P. D., Knebe A., Gibson B. K., 2004, MNRAS, 351, 399

Green D. M., Swets J. A., 1966, Signal Detection Theory and Psychophysics. Wiley, New York

Hilden J., 1991, Medical Decision Making, 11, 95

Jenkins A., Frenk C. S., White S. D. M., Colberg J. M., Cole S., Evrard A. E., Couchman H. M. P., Yoshida N., 2001, MNRAS, 321, 372

Kamdar H. M., Turk M. J., Brunner R. J., 2016, MNRAS, 455, 642

Knollmann S. R., Knebe A., 2009, ApJS, 182, 608

Kohavi R., 1995, in IJCAI'95: Proceedings of the 14th International Joint Conference on Artificial Intelligence. Vol. 2. Morgan Kaufmann Publishers Inc., San Francisco, CA, p. 1137

Kuhlen M., Vogelsberger M., Angulo R., 2012, Phys. Dark Universe, 1, 50

Lochner M., McEwen J. D., Peiris H. V., Lahav O., Winter M. K., 2016, ApJS, 225, 31

Louppe G., Wehenkel L., Sutera A., Geurts P., 2013, in Burges C. J. C., Bottou L., Welling M., Ghahramani Z., Weinberger K. Q., eds, Advances in Neural Information Processing Systems. Vol. 26. Curran Associates, Inc., New York, p. 431

Maggiore M., Riotto A., 2010, ApJ, 711, 907

Nadler E. O., Mao Y.-Y., Wechsler R. H., Garrison-Kimmel S., Wetzel A., 2018, ApJ, 859, 129

Niculescu-Mizil A., Caruana R., 2005, in ICML '05: Proceedings of the 22nd International Conference on Machine Learning. ACM, New York, p. 625

Paranjape A., Sheth R. K., 2012, MNRAS, 426, 2789

Pedregosa F. et al., 2011, J. Machine Learning Res., 12, 2825

Planck Collaboration XIII, 2016, A&A, 594, A13

Pontzen A., Roškar R., Stinson G. S., Woods R., Reed D. M., Coles J., Quinn T. R., 2013, pynbody: Astrophysics Simulation Analysis for Python. Astrophysics Source Code Library, record ascl:1305.002

Press W. H., Schechter P., 1974, ApJ, 187, 425

Reed D., Gardner J., Quinn T., Stadel J., Fardal M., Lake G., Governato F., 2003, MNRAS, 346, 565

Sheth R. K., Tormen G., 1999, MNRAS, 308, 119

Sheth R. K., Mo H. J., Tormen G., 2001, MNRAS, 323, 1

Springel V., 2005, MNRAS, 364, 1105

Springel V., Yoshida N., White S. D. M., 2001, New Astron., 6, 79

Tinker J., Kravtsov A. V., Klypin A., Abazajian K., Warren M., Yepes G., Gottlöber S., Holz D. E., 2008, ApJ, 688, 709

Witten I. H., Frank E., Hall M. A., Pal C. J., 2016, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc., Burlington, MA.

This paper has been typeset from a TeX/LaTeX file prepared by the author.