Machine learning cosmological structure formation

Luisa Lucie-Smith,1* Hiranya V. Peiris,1,2 Andrew Pontzen,1 Michelle Lochner1,3,4
1Department of Physics & Astronomy, University College London, Gower Street, London WC1E 6BT, UK
2The Oskar Klein Centre for Cosmoparticle Physics, Stockholm University, AlbaNova, Stockholm, SE-106 91, Sweden
3African Institute for Mathematical Sciences, 6 Melrose Road, Muizenberg, 7945, South Africa
4SKA SA, The Park, Park Road, Cape Town 7405, South Africa

Accepted XXX. Received YYY; in original form ZZZ

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 realisations to demonstrate the generality of our results.

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

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 et al. 2001; Springel 2005; Kuhlen et al. 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 which yield better agreement with numerical simulations (Sheth et al. 2001; Maggiore & Riotto 2010; Paranjape & Sheth 2012; Farahi & Benson 2013; Borzyszkowski et al. 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 cos-
mological 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 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 which 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 which 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 et al. 2016; Agarwal et al. 2017; Nadler et al. 2017).

We choose random forests (Breiman et al. 1984; Breiman 2001), a popular algorithm which 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 allows 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 which end up in haloes of mass above a threshold, and those which 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 organised 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 Sec. 2. In Sec. 3 we interpret the classification output and present our results in Sec. 4. We then extend the feature set to include the tidal shear field in Sec. 5 and discuss the resulting implications. We study the algorithm’s performance as a function of halo properties in Sec. 6. We perform two blind tests of our pipeline on independent simulations in Sec. 7, demonstrating the generality of our results, and finally conclude in Sec. 8.

2 METHOD

We trained and tested the random forest with an existing dark-matter-only simulation produced with P-GADGET-3 (Springel 2005; Springel et al. 2001) and a WMAP5 ΛCDM cosmological model (Dunkley et al. 2009); $\Omega_M = 0.721$, $\Omega_b = 0.279$, $\Omega_m = 0.045$, $\sigma_8 = 0.817$, $h = 0.701$, $n_s = 0.96$. The comoving softening length of the simulation is $\epsilon = 25.6 \, \text{kpc}$. The simulations evolve $256^3$ dark-matter particles, each of mass $M_{\text{particle}} = 8.24 \times 10^8 \, M_\odot$, in a box of comoving size $L = 50 \, h^{-1} \, \text{Mpc}$ from $z = 99$ to $z = 0$.\footnote{We make use of the Python package pynbody (Pontzen et al. 2013) to analyse the information contained in the simulation snapshots.}

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 $\sim 10^9 \, M_\odot$ to $\sim 10^{14} \, M_\odot$.

We used the 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 $M \geq 1.8 \times 10^{12} \, M_\odot$ at $z = 0$ (401 haloes), and...
the OUT class to contain all remaining particles, including those in haloes of mass $M < 1.8 \times 10^{12}$ $\odot$ and those that do not belong to any halo. 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 which would ultimately allow us to extend the binary classification to a multi-class 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 true class label. The robustness of the algorithm was tested further on independent N-body simulations (Sec. 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 $\delta_L \geq \delta_c$. 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 $\delta(x) = [\rho(x) - \bar{\rho}] / \bar{\rho}$, where $\bar{\rho}$ is the mean matter density of the universe, on a smoothing scale $R$,

$$\delta(x; R) = \int \delta(x') W_{\text{TH}}(x - x'; R) \, d^3x',$$

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

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

The convolution (1) was carried out in Fourier space, which naturally accounts for the periodicity of simulations. A window function $W(x, R)$ of characteristic radius $R$ corresponds to a mass scale $M_{\text{smoothing}} = \bar{\rho} V(R)$, where in the case of a top-hat window function $V_{\text{TH}}(R) = 4/3 \pi R^3$. The feature for machine learning then consists of the density contrast smoothed with a top-hat window function of mass scale $M_{\text{smoothing}}$ (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., $3 \times 10^{10} \leq M_{\text{smoothing}} / M_\odot \leq 1 \times 10^{15}$, 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](https://example.com/fig1.png)

**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 $\delta$ 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...
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.

| Predicted Class | True Class | P | N |
|-----------------|------------|---|---|
| P               | True Positive (TP) | False Positive (FP) |
| N               | False Negative (FN) | True Negative (TN) |

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 which need to be optimised 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 optimise 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 curve, see Sec. 3) and average scores over all k validation sets to obtain the final score of a training set. Here, we performed a five-fold cross validation for all combinations of hyperparameters and retained the combination which 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,

\[ TPR = \frac{TP}{TP + FN}. \]  

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,

\[ FPR = \frac{FP}{FP + TN}. \]  

Receiver Operating Characteristic (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

![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 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.](attachment:image.png)
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.

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

4 DENSITY FIELD CLASSIFICATION

Figure 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 $M$ is equivalent to the fraction of density trajectories with a first upcrossing of the density threshold barrier $\delta_{th}$ at mass scale $M$. We take the density threshold to be the spherical collapse threshold adopted by Bond et al. (1991): $\delta_{th}(z) = (D(z)/D(0))\delta_{sc}$, where $\delta_{sc} \approx 1.686$. 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 emphasise 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% 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 at all nodes $t$ where the feature is used, averaged

\[\text{Importance} = \sum_t w_t \cdot \Delta I(t)\]

where $w_t$ is the weight of node $t$, and $\Delta I(t)$ is the decrease in impurity at node $t$. The impurity of a node is typically measured by Shannon entropy.
over all trees $T$ in the forest:

$$\text{Imp}(X) = \frac{1}{N_T} \sum_T \sum_{i \in T} p(t) \Delta i(t),$$

where $N_T$ is the number of trees, $p(t)$ is the fraction of particles reaching node $i$ 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 normalised 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 $10^{12} - 10^{13} \, M_{\odot}$, 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.

$$i_E(t) = - \sum_{j=1}^{c} p(j, t) \log_2 p(j, t),$$

where $p(j, t)$ is the proportion of particles that belong to class $j$ at node $i$ and $c$ is the total number of classes.

5.1 Tidal shear features

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

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

where $\Phi(x)$ is the peculiar gravitational potential at position $x$ 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 re-parametrised in terms of the ellipticity, $e$, and prolateness, $p$ (Bond & Myers 1996):

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

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

where $\lambda_1 + \lambda_2 + \lambda_3 = \delta$ is the smoothed overdensity used as a density feature. In order to minimise 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

$$e_t = t_1 - t_3,$$

$$p_t = 3(t_1 + t_3).$$

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 PCA to reduce the 150-dimensional feature set to a 10-dimensional space retaining 98% 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 shows little improvement compared to the case of the density-only feature set. We find an improvement of only 2% 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 $10^{12} - 10^{13} \, M_{\odot}$, 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

MNRA 000, 1–11 (2017)
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.

The ST formalism provides a prescription to predict the final 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 \( (1 \times 10^{14} \leq M_{\text{halo}}/M_\odot \leq 4 \times 10^{14}) \), particles in group-sized haloes \( (1 \times 10^{13} \leq M_{\text{halo}}/M_\odot < 1 \times 10^{14}) \) and particles in galaxy-sized haloes \( (1.2 \times 10^{12} \leq M_{\text{halo}}/M_\odot < 1 \times 10^{13}) \). 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 panel of Fig. 5, where the
The IN class particles are split into inner \((r/r_{\text{vir}} \leq 0.3)\), mid \((0.3 < r/r_{\text{vir}} \leq 0.6)\) and outer \((0.6 < r/r_{\text{vir}} \leq 1)\) 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 panel: The IN class particles are split into cluster-sized \((1 \times 10^{12} \leq M_{\text{halo}}/M_\odot \leq 4 \times 10^{13})\), group-sized \((1 \times 10^{13} \leq M_{\text{halo}}/M_\odot < 1 \times 10^{14})\) and galaxy-sized \((1.2 \times 10^{14} \leq M_{\text{halo}}/M_\odot < 1 \times 10^{15})\) haloes and the ROC curves show the random forest’s performance in classifying each category. 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.

The left 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 which 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 under-representativeness 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 re-trained 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 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
Consider probability thresholds of 70% that overall these particles have very low misclassification rate of OUT particles that do not belong to any halo. We find the range category. The machine learning algorithm again shows its ability to match the excursion set close to the halo’s centre-of-mass. The machine learning algorithm shows a high completeness-to-contamination trade-off.

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% to assign particles to the IN class (as in the upper panel of Fig. 6), the fraction of misclassified over all particles that don’t belong to haloes is 2.45%, 4.3%, 6.58% and 10.11%, 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 which 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 halos and in galaxy-sized halos. We find no significant improvement in the classification performance of such particles, other than the 2% 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 re-training, 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 realisation of the same WMAP5 ΛCDM cosmology adopted in the training simulation, for a box of also same size and resolution (see Sec. 2). The second independent test simulation (P-Test) is a realisation of a different cosmological model, a Planck ΛCDM cosmology(Planck Collaboration et al. 2016) in a box of comoving size $L = 50 \, \text{Mpc}$ containing $N = 512^3$ particles. Moreover, in the P-Test simulation we identify haloes at $z = 0$ using the Amiga Halo Finder (AHF)(Gill et al. 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-Test 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% for the density-only feature set and 1.1% for the density and shear feature set. For the W-Test simulation, the AUC difference with the training simulation is of 1.3% for the density-only feature set and 1.6% for the density and shear feature set. Such differences

4 The cosmological parameters are $\Omega_m = 0.6914$, $\Omega_b = 0.3086$, $\Omega_k = 0.045$, $\sigma_8 = 0.831$, $h = 0.6727$, $n_s = 0.96$.
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 particles. These blind tests demonstrate the robustness of the results from a machine learning algorithm trained on one simulation, and applied to different realisations of the same cosmology or realisations of different cosmologies.

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 Sheth-Tormen 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 realisations and variations in cosmological parameters. We conclude that the linear density field contains sufficient information to predict the formation of dark 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 multi-class 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 which features are the most relevant to cosmological structure for-
Machine learning structure formation

ACKNOWLEDGEMENTS

LLS thanks Nina Roth for providing one of the simulations used in this work and for useful discussions. LLS 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., Dave R., Bassett B. A., 2017, preprint, (arXiv:1712.03255)
Ball N. M., Brunner R. J., 2010, International Journal of Modern Physics 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
Caruana R., Niculescu-Mizil A., 2006, in Proceedings of the 23rd International Conference on Machine Learning. ICML ’06. ACM, New York, NY, USA, pp 161–168, doi:10.1145/1143844.1143865, http://doi.acm.org/10.1145/1143844.1143865
Doroshkevich A. G., 1970, Astrofizika, 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 Recogn. Lett., 27, 861
Gill S. P. D., Knebe A., Gibson B. K., 2004, MNRAS, 351, 399
Green D., Swets J., 1966, Signal detection theory and psychophysics
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 Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI’95. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 1137–1143
Kuhlen M., Vogelsberger M., Angulo R., 2012, Physics of the Dark Universe, 1, 50
Lochner H. M., Turk M. J., Brunner R. J., 2016, MNRAS, 455, 642
Maggiore M., Riotto A., 2010, ApJ, 711, 907
Mokler E. O., Mao Y.-Y., Wechsler R. H., Garrison-Kimmel S., Wetzel A., 2017, preprint, (arXiv:1712.04467)
Nadler E., Caruana R., 2005, in Proceedings of the 22Nd International Conference on Machine Learning. ICML ’05. ACM, New York, NY, USA, pp 625–632, doi:10.1145/1102351.1102430, http://doi.acm.org/10.1145/1102351.1102430
Planck Collaboration et al., 2016, A&A, 594, A13
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
Sip咸 G., Mo H. J., Tormen G., 2001, MNRAS, 323, 1
Springel V., 2005, MNRAS, 364, 1105
Witten I. H., Frank E., Hall M. A., Pal C. J., 2016, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann

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