Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **509,** 5046–5061 (2022)
Advance Access publication 2021 November 11

https://doi.org/10.1093/mnras/stab3221

# A machine learning approach to mapping baryons on to dark matter haloes using the EAGLE and C-EAGLE simulations

Christopher C. Lovell,[1★] Stephen M. Wilkins,[2★] Peter A. Thomas,[2★] Matthieu Schaller,[3,4] Carlton M. Baugh,[5] Giulio Fabbian[6,7] and Yannick Bahé[4]

[1] *Centre for Astrophysics Research, School of Physics, Astronomy & Mathematics, University of Hertfordshire, Hatfield AL10 9AB, UK*
[2] *Astronomy Centre, Department of Physics and Astronomy, University of Sussex, Brighton BN1 9QH, UK*
[3] *Lorentz Institute for Theoretical Physics, Leiden University, PO Box 9506, NL-2300 RA Leiden, the Netherlands*
[4] *Leiden Observatory, Leiden University, PO Box 9513, NL-2300 RA Leiden, the Netherlands*
[5] *Institute for Computational Cosmology, Department of Physics, Durham University, Science Laboratories, South Road, Durham DH1 3LE, UK*
[6] *School of Physics and Astronomy, Cardiff University, The Parade, Cardiff CF24 3AA, UK*
[7] *Center for Computational Astrophysics, Flatiron Institute, New York, NY 10010, USA*

Accepted 2021 November 2. Received 2021 October 26; in original form 2021 June 9

## ABSTRACT

High-resolution cosmological hydrodynamic simulations are currently limited to relatively small volumes due to their computational expense. However, much larger volumes are required to probe rare, overdense environments, and measure clustering statistics of the large-scale structure. Typically, zoom simulations of individual regions are used to study rare environments, and semi-analytic models and halo occupation models applied to dark-matter-only (DMO) simulations are used to study the Universe in the large-volume regime. We propose a new approach, using a machine learning framework, to explore the halo–galaxy relationship in the periodic EAGLE simulations, and zoom C-EAGLE simulations of galaxy clusters. We train a tree-based machine learning method to predict the baryonic properties of galaxies based on their host dark matter halo properties. The trained model successfully reproduces a number of key distribution functions for an infinitesimal fraction of the computational cost of a full hydrodynamic simulation. By training on both periodic simulations and zooms of overdense environments, we learn the bias of galaxy evolution in differing environments. This allows us to apply the trained model to a *larger* DMO volume than would be possible if we only trained on a periodic simulation. We demonstrate this application using the (800 Mpc)$^3$ P-Millennium simulation, and present predictions for key baryonic distribution functions and clustering statistics from the EAGLE model in this large volume.

**Key words:** galaxies: abundances – galaxies: luminosity function, mass function.

## 1 INTRODUCTION

Cosmological hydrodynamic simulations self-consistently model the evolution of baryonic and cold dark matter, and the subsequent hierarchical assembly of galaxies in a $\Lambda$ cold dark matter universe (Benson 2010; Somerville & Davé 2015). A number of projects, such as EAGLE (Schaye et al. 2015), ILLUSTRIS (Vogelsberger et al. 2014), ILLUSTRIS-TNG (Pillepich et al. 2018), MUFASA (Davé, Thompson & Hopkins 2016), and SIMBA (Davé et al. 2019), have had reasonable success at reproducing key galaxy distribution functions in the low-redshift Universe, such as the galaxy stellar mass function (GSMF). They are typically run within large periodic volumes, ∼100 Mpc on a side, and have mass resolutions of order $\sim 10^6\,\mathrm{M}_\odot$. This is sufficient to resolve the internal structure of galaxies, but still coarse enough to necessitate the use of subgrid models for small-scale stellar and black hole processes.

It is currently computationally infeasible to run simulations at this resolution in substantially larger volumes.[1] This is an issue for certain science questions, since the volumes typically simulated, ∼(100 Mpc)$^3$, do not contain large numbers of rare, overdense environments, as well as galaxies with unusual growth histories (e.g. star bursts). For example, EAGLE contains only seven clusters at $z = 0$, and all these are of relatively low mass ($M_{200,\mathrm{crit}}\,/\,\mathrm{M}_\odot < 10^{14.5}$). In order to simulate rare environments that are not represented in smaller scale periodic volumes, another approach is to use ‘zoom’ simulations (Katz & White 1993; Tormen, Bouchet & White 1997). These use initial conditions selected from a much larger dark-matter-only (DMO) simulation, of order ∼(1 Gpc)$^3$ in volume, and then resimulate a smaller region from this volume with full hydrodynamics. Large-scale tidal forces are preserved by simulating the rest of the volume with low-resolution DMO particles. This approach has been used successfully to simulate cluster environments with the EAGLE model (Bahé et al. 2017; Barnes et al. 2017b).

★ E-mail: c.lovell@herts.ac.uk (CCL); s.wilkins@sussex.ac.uk (SMW); p.a.thomas@sussex.ac.uk (PAT)

[1] The BLUETIDES simulation (Feng et al. 2016) is one of the largest high-resolution hydrodynamic simulations, with a volume 400 $h^{-1}$ Mpc on a side, but was only run to $z = 7$.

Published by Oxford University Press on behalf of Royal Astronomical Society

However, since zooms only simulate a small region of interest they have a number of drawbacks compared with periodic simulations. They cannot be used to predict mean distribution functions *directly*, since they are, by construction, biased regions. One means of circumventing this limitation is to use multiple zoom simulations of differing environments, and weight the relative abundance of each simulation based on its relative total matter overdensity. This technique was first demonstrated with the GIMIC simulations (Crain et al. 2009), and recently used in the FLARES simulations to make predictions for the abundance of galaxies during the epoch of reionization (Lovell et al. 2021; Vijayan et al. 2021). Another drawback is that zooms cannot be used to self-consistently predict aspects of the large-scale structure, such as the clustering of galaxies, since they are by construction non-representative, small volume regions of the universe. Large periodic volumes are the only means of studying these kinds of spatial statistics (e.g. the BAHAMAS project; McCarthy et al. 2017), but these large volumes cannot currently be simulated at the high resolution necessary to model internal galaxy structures. This limits what can be achieved with high-resolution hydrodynamic simulations.

*N*-body DMO simulations predict the distribution of matter as a result of gravitational interactions only, and are therefore significantly cheaper computationally than simulations including the gas hydrodynamics. They are therefore less demanding to run accurately in large volumes, allowing them to be used to explore the large-scale structure. There are also a number of approaches to modelling galaxy evolution that are relatively simpler than running a full hydro simulation, using semi-analytic or phenomenological models to populate haloes in DMO simulations. The host halo has a significant impact on the properties of a galaxy; haloes are the cradles within which galaxies form, and continue to influence the evolution of a galaxy throughout its lifetime (Wechsler & Tinker 2018). Understanding the relationship between galaxy properties and the properties of their host haloes is an important factor in understanding galaxy formation and evolution, and in the subsequent building of these kinds of galaxy evolution models.

Semi Analytic Models (SAMs) explicitly assume a close relationship between a galaxy and its host-halo. They treat the complicated physics of galaxy formation with approximate, physically motivated analytical models, applied *ex post facto* to *N*-body DMO simulations (for a review, see Baugh 2006). The halo properties, and their merging history, provide the input parameters for such models, which have successfully reproduced a number of distribution functions simultaneously (e.g. Gonzalez-Perez et al. 2014; Henriques et al. 2015, 2020). Subhalo abundance matching (SHAM) models also rely explicitly on the galaxy–halo relationship, populating dark matter haloes from simulations with rank-ordered galaxies from observations. Such models have been used to constrain the stellar mass–halo mass relation (e.g. Behroozi, Conroy & Wechsler 2010; Moster et al. 2010; Moster, Naab & White 2013; Legrand et al. 2019), though it has been noted that the efficacy of such methods is highly dependent on the observational selection function (Stiskalek et al. 2021). Both these approaches are capable of modelling galaxy evolution over very large volumes, allowing predictions for the clustering of galaxies as well as their evolution in rare, overdense environments. They have also been used in combination with hydrodynamic simulations in order to highlight potential issues (e.g. for satellites where mergers lead to mass-loss; Simha et al. 2012), and SAMs have even been explicitly calibrated to reproduce hydrodynamic simulations (Neistein et al. 2012; Mitchell & Schaye 2021), allowing an investigation into the effects of changes to specific coefficients in the model.

Machine learning methods continue to grow in popularity in all areas of astronomy (see Ball & Brunner 2010; Fluke & Jacobs 2020), and a number of recent papers have explored how they can be used in combination with simulations to emulate galaxy properties, analogous to an SHAM or SAM model. In a pioneering paper, Xu et al. (2013) used the Millennium simulation, coupled with an SAM, to predict the number of galaxies in a given halo using support vector machines and *k*-nearest neighbour algorithms. Later, Kamdar, Turk & Brunner (2016a) showed how tree-based methods can be trained to learn additional properties of the the baryon–halo relationship directly from an existing SAM. They used dark matter properties from each halo as features, and baryonic properties as predictors, and trained the machine to learn the mapping between the two. They then followed this up by applying the same technique to the ILLUSTRIS hydrodynamic simulation (Kamdar, Turk & Brunner 2016b). Agarwal, Davé & Bassett (2018) presented a similar model applied to the MUFASA simulation. Using the more recent ILLUSTRIS-TNG simulation, Jo & Kim (2019) presented a similar model, and then applied this trained model to the much larger DMO MultiDark-Planck simulation. A novel addition to their model was historical halo features (extracted from the halo merger tree), which allowed the model to broadly reproduce key distribution functions, though we note that they do not present tests in the high halo mass regime ($>10^{14}$ M$_\odot$). Sullivan, Iliev & Dixon (2018) used artificial neural networks to better predict the baryon fraction of haloes at high redshift using both dark matter and baryonic properties from their RAMSES-RT radiative transfer simulations. Most recently, a number of hybrid approaches have been presented: Moews et al. (2020) combined the results of an equilibrium model with machine learning on the SIMBA simulations, and Hearin et al. (2020) combined empirical modelling with simulation outputs from an SAM to populate large DMO volumes with galaxies. Icaza-Lizaola et al. (2021) demonstrated, using a sparse regression approach, that halo angular momentum has little impact on the stellar–halo mass relation. Finally, a number of approaches have demonstrated predictions for baryonic properties of the cosmic web not necessarily linked to discrete subhaloes (e.g. Sinigaglia et al. 2021).

In this paper, we build on these previous works, by combining the results of both periodic and zoom cosmological simulations from the EAGLE project to train a machine learning model to learn the relationship between galaxy baryonic properties and their host dark matter haloes. Our approach is unique in two ways. First, we match subhaloes from each hydrodynamic simulation with those in a DMO counterpart (simulated from the same initial conditions), in order to avoid the effect of baryons on the host dark matter halo (Schaller et al. 2015). This allows the model to be directly applied to an independent DMO simulation, without leading to biases in the predictions due to differences in the dark matter features.

Secondly, we address the issue of *generalization error*. Machine learning methods are a powerful set of techniques for making predictions on data that look similar to the data on which they are trained, but fail when presented with new data that lie outside of the bounds of the original training data. This presents a problem for models trained on smaller periodic volumes, since such volumes will not contain the massive clusters present in larger DMO simulations, and hence any model trained on these volumes would not provide good predictions for galaxies in overdense environments. We avoid this by including clusters from the C-EAGLE project (C-EAGLE; Bahé et al. 2017; Barnes et al. 2017b) in our training set. This allows us to apply the trained model to the much larger volume (800 Mpc)$^3$ P-MILLENNIUM simulation (Baugh et al. 2019), and predict distribution functions of key baryonic properties within this enormous volume, extending the dynamic range, as well as allowing

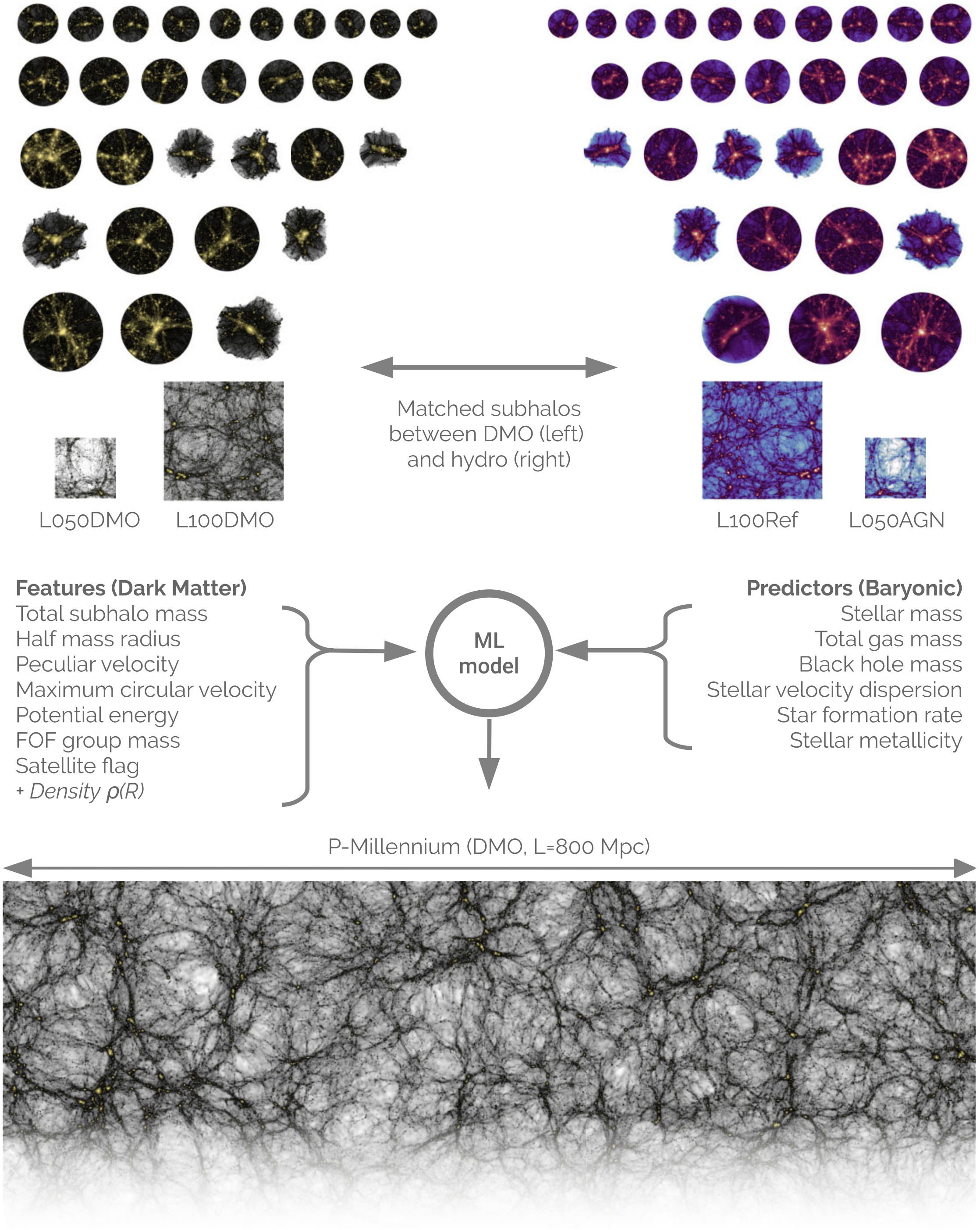

**Figure 1.** Diagram showing the simulations (approximately to scale) used throughout this work, and the features and predictors used for training the machine learning model. At the top are the C-EAGLE zoom simulations; each image shows the distribution of dark matter (left) in the DMO simulations, and the gas (right) in the full hydro simulation, centred on the centre of potential of the most massive FOF group in each simulation, within a radius $r = 15 \times R_{\mathrm{crit},\,200}$. Below these are the cubic periodic `L100Ref` and `L050AGN` simulations, again showing the dark matter (left) and gas (right). In the centre are tables detailing the features from the DMO simulations (left) and the predictors from the hydro simulations (right). At the bottom is a cropped image of the dark matter distribution in the P-Millennium simulation, to which the trained machine learning model is applied to predict the baryonic properties of its haloes.

predictions of clustering statistics on larger scales for higher mass haloes. The method is shown diagrammatically in Fig. 1.

While often negatively perceived as a 'black box', many machine learning methods in fact provide a wealth of insights into the form of their predictive model, and the weight given to their input parameters. This presents an opportunity to learn, in an unbiased manner, what parameters best explain the galaxy–halo connection. We train the model with a range of dark matter properties, and

**Table 1.** Details on each simulation set. The columns provide (1) the name or description of the simulation set, (2) the prefix used throughout this paper, (3) the total volume, (4) the number of subhaloes with mass > $10^{10}\,\mathrm{M}_\odot$, (5) the number of those haloes matched between the hydro and DMO simulations (see Section 2.2), (6) the number of subhaloes in the training set, (7) the number of subhaloes in the test set, (8) the value of the viscosity parameter, and (9) the value of the $\Delta T$ parameter.

| Simulation | Prefix | Volume (Mpc$^3$) | $N_{\rm halo}(> 10^{10}\,\mathrm{M}_\odot)$ | $N_{\rm matched}$ | $N_{\rm train}$ | $N_{\rm test}$ | $C_{\rm visc}$ | $\Delta T$ |
|---|---|---|---|---|---|---|---|---|
| Reference L0100N1504 | `L100Ref` | $100^3$ | 88 173 | 86 861 | 69 615 | 17 246 | $2\pi$ | $10^{8.5}$ |
| AGNdT9 L0050N0752 | `L050AGN` | $50^3$ | 11 423 | 11 265 | 9031 | 2231 | $2\pi \times 10^2$ | $10^9$ |
| C-EAGLE | `ZoomAGN` | $202.7^3$ | 373 275 | 364 408 | – | – | $2\pi \times 10^2$ | $10^9$ |
| C-EAGLE + L050AGN | `L050AGN` + `ZoomAGN` | $203.7^3$ | 384 698 | 375 673 | 300 770 | 74 903 | $2\pi \times 10^2$ | $10^9$ |

explore the relative predictive power of each one on the baryonic properties. Hydrodynamic simulations represent the cutting edge of cosmological modelling; machine learning methods could provide a practical way of extracting quantitative information on the modelled relationships. All of these insights can be used to inform future analytic, semi-analytic, and hydrodynamic model development.

This paper is laid out as follows. In Section 2, we present the simulations used to train the model, as well as our algorithm for matching subhaloes between the hydro and DMO runs. Section 3 details the machine learning methods used, as well as our choice of features and predictors. Section 4 details our results on test sets, including the effect of including density information. Section 5 presents our results on independent DMO simulations, including the P-Millennium simulation, and Section 6 shows our feature exploration analysis. Finally, in Section 7 we discuss our results and summarize our conclusions. Throughout, we assume a (flat) Planck year 1 cosmology ($\Omega_{\rm m} = 0.307$, $\Omega_\Lambda = 0.693$, $h = 0.6777$, Planck Collaboration I 2014) and a Chabrier stellar initial mass function (IMF; Chabrier 2003).

## 2 SIMULATIONS

### 2.1 The EAGLE and C-EAGLE simulations

The EAGLE project is a suite of cosmological hydrodynamic simulations (Crain et al. 2015; Schaye et al. 2015) employing subgrid models for feedback from stars and active galactic nuclei (AGN). EAGLE has been shown to accurately reproduce many observed relations, including the GSMF, galaxy sizes, quenched fractions, gas content, and black hole masses (Lagos et al. 2015; Trayford et al. 2015, 2017; Bahé et al. 2016; Crain et al. 2017; Furlong et al. 2017; McAlpine et al. 2017) at a range of redshifts (e.g. Furlong et al. 2015). A number of different resolutions and volumes make up the EAGLE simulation suite. In this work, we use the 'fiducial' resolution simulations, with gas particle mass $m_g = 1.8 \times 10^6\,\mathrm{M}_\odot$, dark matter particle mass $9.7 \times 10^6\,\mathrm{M}_\odot$, and a physical softening length of 0.7 kpc. Haloes in the simulation are identified first through a Friends-Of-Friends (FOF) halo finder, and then split into child self-bound objects with SUBFIND (Dolag et al. 2009). Cluster-Eagle (or C-EAGLE, Bahé et al. 2017; Barnes et al. 2017b) uses the EAGLE model to simulate cluster environments using the 'zoom' resimulation technique (Katz & White 1993; Tormen et al. 1997). 30 clusters at $z = 0$ (shown in Fig. 1), with a range of halo masses ($14 < \log_{10}(\mathrm{M}_{200}\,/\,\mathrm{M}_\odot) < 15.51$), are selected from a (3.2 Gpc)$^3$ 'parent' DMO simulation (Barnes et al. 2017a). The clusters are resimulated at an identical resolution to the fiducial periodic EAGLE simulation. Full details on the selected clusters are provided in Barnes et al. (2017b).

C-EAGLE uses the AGNdT9 calibration of the EAGLE model (Schaye et al. 2015), which, compared with the fiducial Reference model, uses a higher value for $C_{\rm visc}$, which controls the sensitivity of the BH accretion rate to the angular momentum of the gas, and a higher gas temperature increase from AGN feedback, $\Delta T$. A larger $\Delta T$ leads to fewer, more energetic feedback events, whereas a lower $\Delta T$ leads to more continual heating. Schaye et al. (2015) show that AGNdT9 predicts X-ray luminosities and hot gas fractions in galaxy groups in better agreement with observational constraints, though with some discrepancies on cluster scales (Barnes et al. 2017b).

Table 1 details the simulations used in this work, and any combinations. `L100Ref` is a (100 Mpc)$^3$ periodic volume (shown in Fig. 1) run with the Reference model parameters; the hydro simulation contains $1504^3$ dark matter and $1504^3$ gas particles. `L050AGN` is a smaller (50 Mpc)$^3$ periodic volume (shown in Fig. 1) run at the same resolution as `L100Ref` but with the AGNdT9 model parameters; it contains $752^3$ dark matter and $752^3$ gas particles. `L050AGN` + `ZoomAGN` is a combination of `L050AGN` with the zoom cluster regions from C-EAGLE. We also match with DMO counterparts to each of these simulations, run using the same initial conditions; the match is described in Section 2.2. We use the snapshot corresponding to $z = 0.101$ in all simulations.

Throughout the rest of this text, whenever we refer to a *model* we are referring to a *machine learning* model (unless otherwise stated) trained on the matched hydro-DMO simulations indicated in the name. The simulations are all referred to explicitly as *simulations* to distinguish them from the machine learning models.

### 2.2 Matching between hydrodynamic and DMO simulations

Including baryons can lead to significant alterations to the underlying dark matter haloes (Weinberg et al. 2008). For example, Schaller et al. (2015) demonstrate that, in the EAGLE simulation, the halo centres are more 'cuspy' in the presence of stars. In order to apply our trained model to DMO simulations it is necessary to avoid these effects, as they will bias any predictions based on the dark matter features. We achieve this by matching subhaloes in each hydrodynamic simulation to their counterparts in DMO simulations, and use the properties of the matched haloes in the DMO simulation as our features. The galaxy properties that a given halo would have if hydrodynamics had been included are then predicted. Each DMO simulation is run from the same initial conditions, but is not split into baryonic and dark-matter species. Aside from this, all cosmological and numerical parameters are identical.

We perform the match using the approach of Schaller et al. (2015). We first find the 50 most bound dark matter particles in a subhalo in the hydro simulation, and search for haloes in the DMO simulation that have 50 per cent or more of these same particles (matched on particle ID). We then perform the same match in reverse (subhaloes in the DMO matched with subhaloes in the hydro simulation). Those haloes that match bijectively are linked.

Fig. 2 shows the fraction of haloes matched from the DMO simulation at a given DMO halo mass for the two periodic simulations

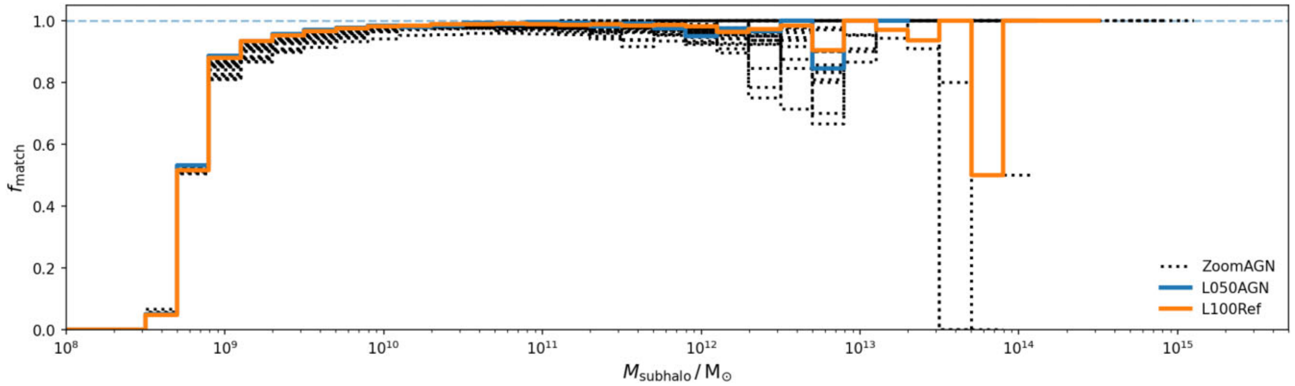

**Figure 2.** Fraction of subhaloes from each DMO simulation matched with a counterpart in the hydro simulation, binned by total subhalo mass. `L050AGN` and `L100Ref` are shown in blue and orange, respectively, and each zoom from `ZoomAGN` is shown as a black dashed line.

(`L100Ref` and `L050AGN`) as well as each of the C-EAGLE clusters. We also detail the total number of haloes and the number of matched haloes for each simulation set in Table 1. More than 95 per cent of subhaloes with $M_{\rm subhalo} > 10^{10}\,{\rm M}_\odot$ are matched bijectively across all simulations. We hence choose to train our model only on subhaloes with masses above this threshold (see Section 3.3 for details). By using a threshold dependent only on the DMO properties, we can use a similar threshold in any target DMO simulation (subject to the existing resolution constraints of that simulation).

It is noticeable that there are a larger fraction of subhaloes at the high-mass end ($M_{\rm subhalo}\,/\,M_\odot > 5 \times 10^{12}$) that are not matched, in both the periodic and zoom simulations. We looked at these cases individually, and found, where a single halo was identified in the DMO simulation, the halo finder splits this halo into multiple individual haloes in the baryonic simulation. Missing these haloes reduces the size of our training set, which is particularly disappointing at the high-mass end where the number of haloes is already low, however we do not expect this to lead to biases in our predictions due to the already heterogenous nature of our training set.

### 2.3 The P-Millennium simulation

P-Millennium is a large DMO simulation (800 cMpc on a side; particle mass $1.06 \times 10^8\,{\rm M}_\odot$) using the same Planck Collaboration I (2014) cosmology as EAGLE. Baugh et al. (2019) first presented the simulation, and demonstrated its use as a parent volume for the Galform model, in order to predict the atomic hydrogen content of galaxies. Safonova, Norberg & Cole (2021) also used P-Millennium as a parent simulation for an SHAM model, generating mock catalogues. P-Millennium uses the same FOF and Subfind structure finders as the EAGLE simulation project, which means the features can be used directly for any model trained on EAGLE. We present our predictions using P-Millennium in Section 5.

## 3 MACHINE LEARNING METHODS

### 3.1 Extremely randomized trees

We used the Scikit-Learn (Pedregosa et al. 2011) implementation of Extremely Randomized Trees (ERT; Geurts, Ernst & Wehenkel 2006), a tree-based ensemble method. ERT is demonstrably effective in this domain compared with other popular machine learning methods (Kamdar et al. 2016a; Jo & Kim 2019).

To understand what makes ERT such an effective learner, first consider a single decision tree. Decision trees are typically constructed top down, numerically evaluating all splits for each feature using a cost function. The best split (lowest cost) is chosen at each level. Some of the issues seen with Decision Trees, particularly overfitting, can be alleviated by ensembling many different trees trained on subsets of the data. Random Forests extend this idea by, at each split, randomly limiting the feature space from which splits can be made (within individual trees not all of the data are used, but over the whole ensemble they are). This increases the variance by stopping strong features from dominating each tree. ERT also introduces another layer of randomness; each split is chosen at random from the range of values available for each feature. Bad splits are still rejected, but the extra layer of randomness encourages exploration of the full feature space, creating more ‘weak’ learners for use in the ensemble. At each iteration, only the best split from the subset of features is chosen, and the iterative procedure continues until a leaf node condition is reached.

Within ERT the mean squared error (MSE) is used to evaluate each split. To quantify the effective fit of each model, and to discriminate between models, we used both MSE and the Pearson correlation coefficient ($\rho$), defined below:

$$\rho = \frac{{\rm cov}(X_{\rm predicted} X_{\rm test})}{\sigma_{X_{\rm predicted}} \sigma_{X_{\rm test}}}. \quad (1)$$

### 3.2 Features and predictors

We chose our features from the properties of the DMO haloes and their host FOF haloes. Some features are expected to be of greater importance for predicting certain baryonic properties; we explore this in Section 6. The selected subhalo features are as follows: total subhalo mass ($M_{\rm sub}$), half-mass radius ($R_{1/2}$), peculiar velocity ($v$), maximum circular velocity ($v_{\rm max}$), radius of maximum circular velocity ($R_{v_{\rm max}}$), potential energy ($E_{\rm p}$), FOF group mass ($M_{\rm crit,\,200}$), and finally a boolean feature that specifies whether the subhalo is a satellite or a central.

Since we wish to evaluate the impact of environment we also include additional features to quantify this. As a simple measure of environment we calculated the density of dark matter within spheres centred on a given subhalo in the DMO simulation. We ran a periodic KD-tree search for neighbouring particles, then calculated the density on different scales, $R = [1, 2, 4, 8]$ Mpc, to quantify both the small- and large-scale environment. We indicate in the text where these additional features are included in a given training set.

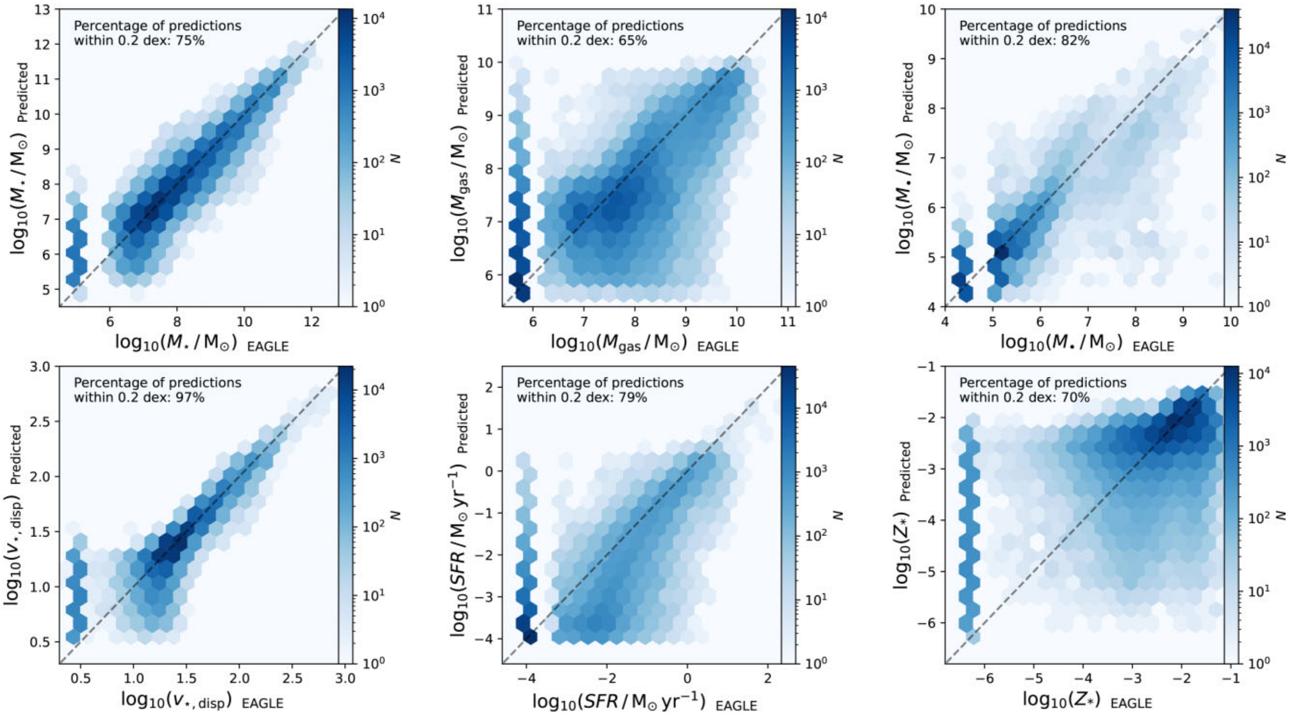

**Figure 3.** Predicted (from the machine learning model) against the true baryonic properties on the test set from the `L050AGN + Zoom` simulation set. Clockwise from top left: stellar mass, gas mass, black hole mass, star formation rate, stellar metallicity, and stellar velocity dispersion. The vertical bar separated from the rest of the distribution to the left in each panel corresponds to galaxies with a true value of zero for that corresponding predictor (see Section 3.2). The fraction of galaxies whose predicted property is within 0.2 dex of the true value is quoted at the top left of each panel.

More dark matter features are available in the subfind catalogues, and additional features could be calculated from the particle information (such as the large-scale tidal torque), but we limited our chosen features to those above as they are present in both the EAGLE and P-MILLENNIUM catalogues. Combinations of features may also lead to better predictive accuracy; we will explore this systematically in future work.

We predict six baryonic properties: the stellar mass, gas mass, black hole mass, stellar velocity dispersion, star formation rate, and stellar metallicity. The stellar mass and gas mass are taken from the central 30 kpc of each subhalo to allow better comparison with observations. We transform all of these predictors into log space, which has been shown to improve the prediction accuracy due to the typically large dynamic range of cosmological properties (Jo & Kim 2019). If the value is zero, we set it to some small value, determined by the resolution limit where appropriate,

$$
\begin{aligned}
M_\star\,/\,\mathrm{M}_\odot &\geqslant 1\times10^5\\
M_{\mathrm{gas}}\,/\,\mathrm{M}_\odot &\geqslant 5\times10^5\\
M_\bullet\,/\,\mathrm{M}_\odot &\geqslant 2\times10^4\\
\mathrm{SFR}\,/\,\mathrm{M}_\odot\ \mathrm{yr}^{-1} &\geqslant 1\times10^{-4}\\
\mathrm{Z}_* &\geqslant 5\times10^{-7}\\
v_{\star,\mathrm{disp}}\,/\,\mathrm{km\ s}^{-1} &\geqslant 3.
\end{aligned}
$$

### 3.3 Training

We train our model on all haloes with a dark matter mass (as measured in the DMO simulation) $M_{\mathrm{sub}}\,/\,\mathrm{M}_\odot \geqslant 1\times10^{10}$. The completeness of our selection with respect to stellar mass is shown in detail in Appendix A. By applying our selection to the dark matter properties we can use the same thresholds when applying the model to independent DMO simulations. We split our data into training and test sets, 80-20 per cent, respectively. All hyperparameter optimization, parameter scaling, and training is done on the training set, and only final model assessment is performed on the test set. For each feature set, the hyperparameters of the ERT instance are chosen through an exhaustive grid search. For each set of hyperparameters, $k$-fold cross-validation is performed (Stone 1974) with $k = 10$ folds, and the coefficient of determination, $\mathrm{R}^2$, is used to discriminate,

$$
R^2 = 1 - \frac{\sum_i (X^i_{\mathrm{test}} - X^i_{\mathrm{predicted}})^2}{\sum_i (X^i_{\mathrm{test}} - X_{\mathrm{mean,train}})^2}\quad . \tag{2}
$$

We standardize all of our features and predictors by subtracting the mean and scaling to unit variance.

## 4 PREDICTING BARYONIC PROPERTIES FROM DARK MATTER PROPERTIES

We first present results for the `L050AGN + ZoomAGN` model, with the fiducial feature set (excluding environmental features). Fig. 3 shows the predicted against the true value for the six baryonic properties in the test set. Figs 4–6 compare the predicted and true distribution of these properties in the test set as violin plots.[2] Together, these figures show how accurately the model predictions are, and how well the cosmic distribution is reproduced. We also quote the fraction of galaxies where the predicted value is within 0.2 dex of the true value; for the stellar velocity dispersion this is as high as 97 per cent, but even for the gas mass, which has the

[2] Bin width of the kernel density estimate is calculated using Scott's rule (Scott 1979).

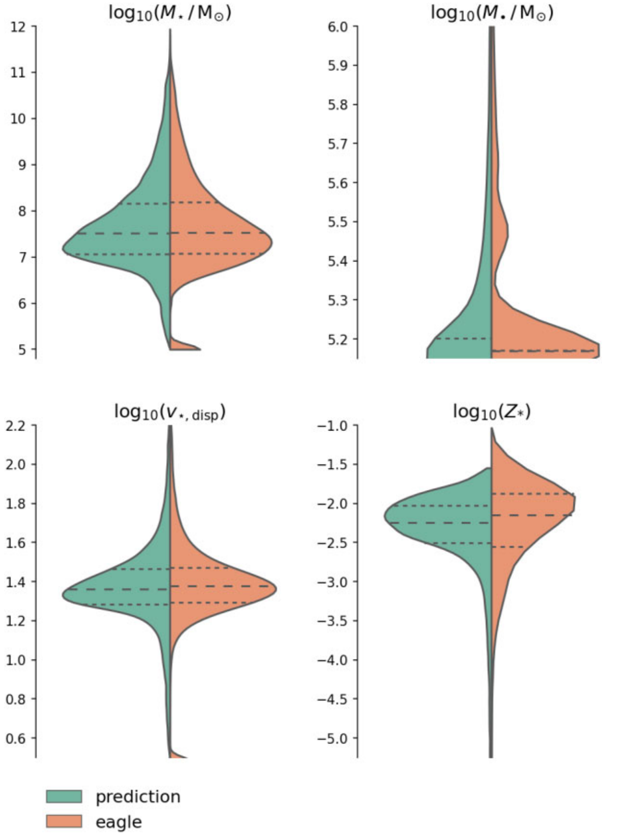

**Figure 4.** Violin plots showing the distribution of predicted baryonic properties (green) from the machine learning model against the true values (orange) in the L050AGN + Zoom simulation set. The dashed and dotted lines show the median and upper/lower quartiles of each distribution, respectively. Each distribution is a kernel density estimate of the true underlying distribution, which may smooth some features, particularly where the distribution is discontinuous (e.g. galaxies with zero gas mass). Clockwise from top left: stellar mass, black hole mass, stellar velocity dispersion, and stellar metallicity.

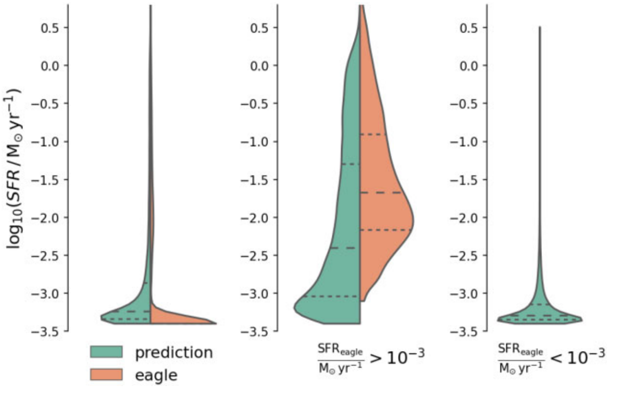

**Figure 5.** Violin plots showing the distribution of predicted SFR (green) from the machine learning model against the true SFR (orange) in the L050AGN + Zoom simulation set. The left plot shows the total distribution, which is heavily skewed towards quiescent galaxies, since the sample is dominated by low-mass galaxies that are artificially quenched. The central plot shows the distribution ignoring those galaxies with zero SFR in the test set. The right plot shows only the *predicted* SFR for all galaxies with zero SFR in the test set (note that this violin is symmetric as only a single property is plotted).

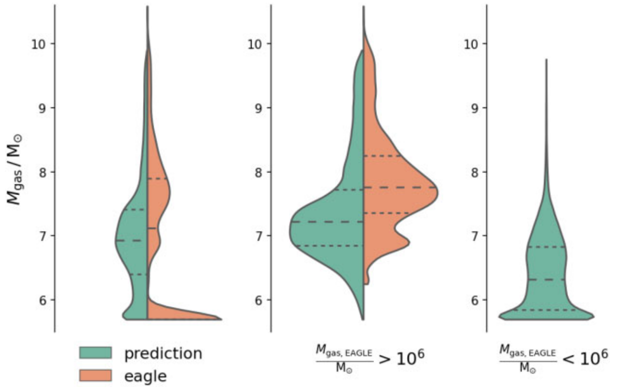

**Figure 6.** As for Fig. 5, but showing the distribution of total gas mass.

lowest prediction accuracy, this is still close to two-thirds of the sample (65 per cent). This is comparable to the accuracy achieved in Neistein et al. (2012) in their SAM trained on a hydro simulation, though we push our predictions to lower stellar masses.

To qualitatively demonstrate the accuracy of the ERT model, we compare to predictions utilizing a single feature (subhalo mass or $V_{\mathrm{max}}$), analogous to an SHAM approach. For this single feature we fit an isotonic regression model[3] between the feature and each predictor (for the whole data set, not just training). This model ensures monotonicity, and broadly fits each predictor well considering the simplicity of the model. We again quote the fraction of galaxies where the predicted value from this simple relation is within 0.2 dex of the true value. The ERT model shows greater accuracy for all predictors compared with the Isotonic model, whether subhalo mass or $V_{\mathrm{max}}$ are used. This is particularly the case for the gas mass (49 per cent where $V_{\mathrm{max}}$ is used, compared with 65 per cent for the ERT model). Full details on the Isotonic fits, and comparison to the predicted GSMF and projected correlation function, are provided in Appendix B.

The model is able to accurately predict both the stellar mass and stellar velocity dispersion remarkably well, however there is more structure in the joint plots for other properties. Predictions for the stellar metallicity show a greater spread than the other values, perhaps unsurprisingly due to its known complex dependence on the star formation history, however the violin plot shows that the overall distribution is recovered. Black hole masses in EAGLE are dominated by newly formed black holes at the seed mass ($10^5\ \mathrm{M}_{\odot}$), as more haloes reach the mass-threshold for black hole seeding. The model is able to capture these, and does a reasonable job of predicting the masses of more massive black holes.

The relations for the total gas mass and SFR are more complicated. There are a large number of galaxies with zero star formation, and the right-hand panel of Fig. 5 shows that the model predicts a range of SFRs for these galaxies, though the majority are limited to $< 3 \times 10^{-3}\ \mathrm{M}_{\odot}\ \mathrm{yr}^{-1}$. To see how well the model predicts the distribution of star-forming galaxies, we show in the middle panel of Fig. 5 the distribution of SFR ignoring quiescent galaxies. It is clear that the model underpredicts the SFR for most galaxies. This may be due to the quiescent galaxies biasing the predictions for other haloes, as well as ERT predicting a smooth distribution of SFRs when a discontinuous distribution would be more appropriate. The SFR is also known to be more strongly dependent on the assembly history;

[3] see here for details on the Isotonic regression model employed.

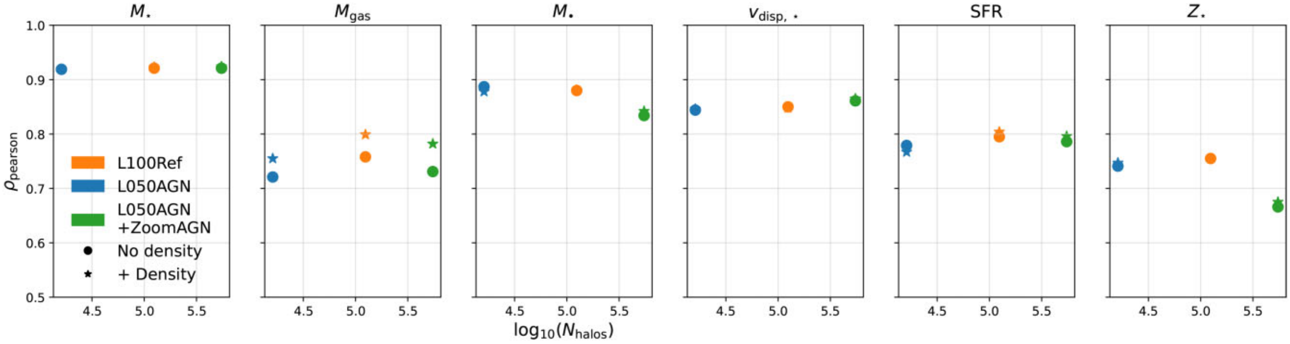

**Figure 7.** Comparison of fit accuracy described by the Pearson correlation coefficient ($\rho_{\rm pearson}$), measured on the test set, against the number of haloes in the training set, for each of the baryonic predictors. The `L100Ref`, `L050AGN`, and `L050AGN + Zoom` simulation sets are shown in orange, blue, and green, respectively, where bullet markers show results with the fiducial feature set, and star markers show result including all local density features (see Section 4.1).

including features that encode this may lead to better predictions, which we discuss in Section 7.

Fig. 6 shows that, as for the SFR, there is a reasonably tight relation for the total gas mass, except where galaxies have zero gas mass. These galaxies make up a large proportion of all subhaloes, and the model fails to predict low gas masses for these galaxies, instead predicting a wider range of gas masses, as can be seen in the right-hand panel of Fig. 6. This suggests that the physics that causes the evacuation of gas from low-mass haloes is not encoded in the provided dark matter parameters. However, the overall distribution, when renormalized, better reproduces that seen in the test set compared to the SFR.

To demonstrate the impact of adding the C-EAGLE clusters to our training set, we compare the prediction accuracy against models trained only on the periodic volumes. Fig. 7 shows the Pearson correlation coefficient for the `L100Ref`, `L050AGN`, and `L050AGN + zoom` models. Adding the zoom regions leads to a large increase in the training set size, but this has no significant positive effect on the predictive accuracy for any of the features. In fact, for the gas mass, black hole mass, and stellar metallicity the predictive accuracy is actually worse. This may be due to the unique impact of the cluster environment on these three particular baryonic properties of galaxies, for example through the effect of ram pressure stripping and fly-by interactions. So while there is more data for the machine to learn from, the relationship represented is more complicated than that present in the periodic volumes, and therefore more difficult to predict. We stress that in order to make predictions for larger boxes, it is essential to include these environments in the training set, and that a lower predictive accuracy compared to the periodic volumes is not necessarily indicative of a poorer model.

This does not suggest that adding more data does not improve the predictive accuracy – $\rho_{\rm pearson}$ calculated for `L100Ref` is higher than than for `L050AGN` for all baryonic properties, showing the advantage of a larger training set size where the underlying distribution of galaxy properties is broadly similar.

### 4.1 The effect of including local density in the feature set

We add four features for the local density calculated within spheres with radii $R = [1, 2, 4, 8]$ Mpc. Fig. 7 also shows the impact of including these additional features on the predictive accuracy for the `L050AGN`, `L100Ref`, and `L050AGN + Zoom` simulation sets. Including density information has a minor positive impact on the predictive accuracy of all features for almost all simulation sets, though the quantitative impact is small in most cases. The largest impact is seen for the gas mass, with an increase in $\rho_{\rm pearson}$ of approximately +0.05 for the periodic simulation sets, and +0.07 for the `L050AGN + Zoom` simulation set. This fits with the hypothesis suggested above that environmental effects operating in clusters lead to poor predictions for the gas mass when environmental features are not included. Such features are important for accurately predicting specific baryonic properties.

In summary, our model is capable of predicting a range of baryonic properties with reasonable accuracy, and successfully reproduces their cosmic distributions. We now show how the model can be applied to independent, larger DMO volumes, and the impact of including the zoom regions on the predicted relations.

## 5 APPLICATION TO DMO SIMULATIONS

A key aim of the model is to produce predictions for distribution functions and clustering statistics for much larger volumes than can be achieved using periodic hydrodynamic simulations. To this end we test how well our model produces the two-point galaxy correlation function (2PCF), GSMF, star-forming sequence, stellar mass–metallicity relation, and the stellar mass–black hole relation in independent DMO volumes, including the $(800\,{\rm Mpc})^3$ P-Millennium simulation.

### 5.1 The two-point galaxy correlation function

Galaxy clustering measurements provide a powerful means of testing gravity and cosmological parameters, including the contribution of dark energy, as well as the impact of galaxy bias on galaxy formation and evolution. One of the key statistics for measuring clustering is the spherically averaged 2PCF (Peebles 1980), defined as

$$\xi(r) = \frac{1}{\langle n \rangle}\frac{{\rm d}P}{{\rm d}V} - 1, \qquad (3)$$

where $\langle n \rangle$ is the mean comoving number density of galaxies, and ${\rm d}P/{\rm d}V$ is the probability of finding a galaxy in volume ${\rm d}V$ at a comoving distance $r$ from another galaxy. For redshift surveys, where the line-of-sight distance is inaccessible, this is often split into projected and line-of-sight distance components, which can be used to estimate the *projected* correlation function (Davis & Peebles 1983),

$$w_{\rm p}(r_{\rm p}) = 2\int_0^{\pi_{\rm max}} \xi(r_{\rm p}, \pi)\,{\rm d}\pi, \qquad (4)$$

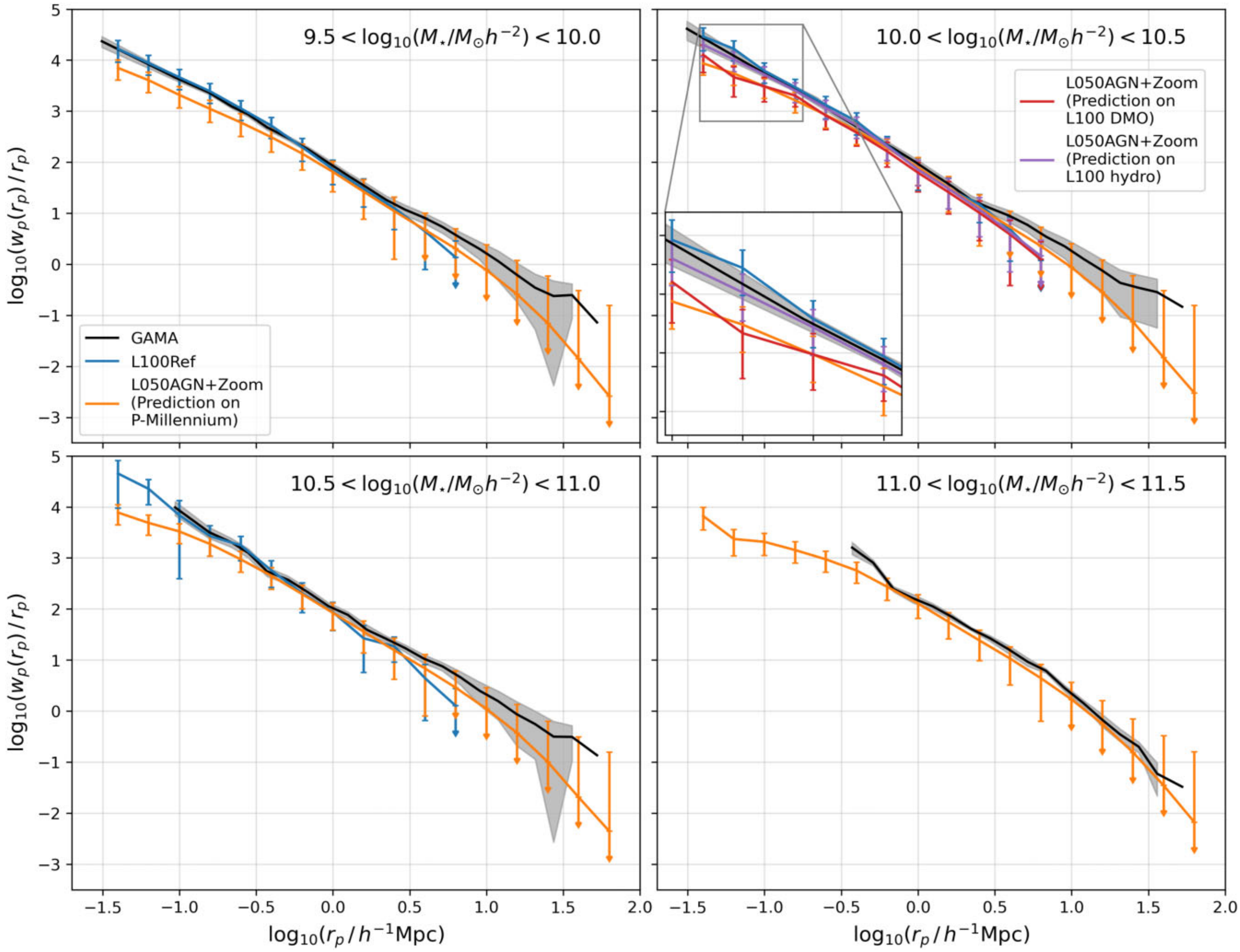

**Figure 8.** Projected correlation function in bins of stellar mass; the mass range is indicated in each column. The results from `L100Ref` are shown in blue, and the `L050AGN + Zoom` machine learning model predictions on P-Millennium are shown in orange. Observational results from GAMA (Farrow et al. 2015) are shown in grey. Errors are estimated using jacknife resampling of each simulation volume.

where $\pi_{\rm max}$ is the maximum distance along the line of sight. Since $w_{\rm p}(r_{\rm p})$ is robust against redshift space distortion effects it is better suited for comparisons with simulations.

Simulation studies of galaxy clustering are typically carried out on large scales with DMO simulations or relatively lower resolution hydrodynamical simulations (e.g. BAHAMAS; McCarthy et al. 2017), and on smaller scales using high-resolution hydrodynamical simulations, which can resolve the baryonic feedback effects on haloes (see van Daalen et al. 2014). We here see how well our machine learning model can provide predictions on both large and small scales *simultaneously* by applying the model to the large-volume P-Millennium simulation. We estimate errors on our clustering statistics using jacknife resampling of each simulation volume (for details, see Artale et al. 2017).

Fig. 8 shows the projected 2PCF measured on the `L100Ref` simulation, the `L050AGN + Zoom` model applied to the P-Millennium simulation, and compared to observational results from GAMA (Farrow et al. 2015) in different stellar mass bins. As shown in Artale et al. (2017), the `L100Ref` simulation is in good agreement with the observational constraints on small scales up to stellar masses of $10^{11}\,{\rm M}_\odot$. However, on larger scales ($r_{\rm p} > 3h^{-1}$Mpc) there is a deficit in the normalization, attributed to finite-volume effects; the smaller periodic boxes do not sample the largest modes in the power spectrum. There are also too few galaxies above a stellar mass of $10^{11}\,{\rm M}_\odot$ in `L100Ref` to obtain robust clustering statistics.

The `L050AGN + Zoom` model, applied to the much larger volume P-Millennium simulation, shows no such deficit at the largest scales. We are in fact able to make predictions out to scales of $100\,h^{-1}$ Mpc, a factor of 10 larger than achievable with the periodic simulations. The model is also able to make predictions for the clustering of the most massive galaxies, $> 10^{11}\,{\rm M}_\odot$, since there are sufficient numbers of these galaxies to produce reliable statistics.

There is, however, a small deficit in the normalization at the smallest scales in the lower mass bins for the `L050AGN + Zoom` model (outside the estimated errors). This may be due to a number of effects, one being the lower resolution of the P-Millennium simulation, which may lead to substructures on small scales being smoothed out. To test the impact of this we applied the `L050AGN + Zoom` model to the DMO 100 Mpc box (using the same initial conditions as the `L100Ref` simulation), which has a mass resolution ~10× higher. This is shown in Fig. 8; at the largest scales the model shows the same deficit as the `L100Ref` simulation, due to the smaller box size. However, at small scales there is the same deficit as in the `L050AGN + Zoom` model applied to P-Millennium. This confirms that it is not resolution effects leading to the lower amplitude.

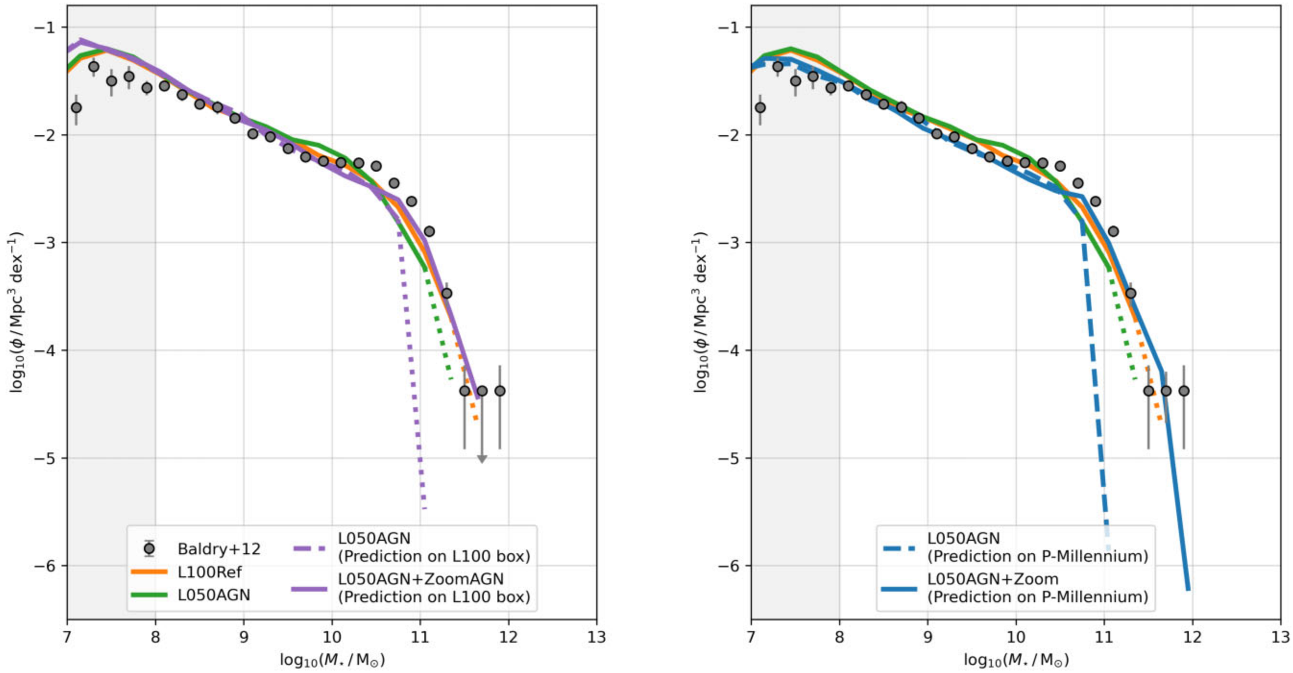

**Figure 9.** The GSMF. Both panels show the GSMF from the `L100Ref` (orange) and `L050AGN` (green) simulation sets for comparison, as well as observational constraints from Baldry et al. (2012). Lines are dotted where there are fewer than 10 galaxies per bin. Left-hand panel: the predicted GSMF on the (100 Mpc)$^3$ DMO volume from machine learning models trained on the `L050AGN` (purple, dashed) and `L050AGN + Zoom` (purple, solid) simulation sets. Right-hand panel: the predicted GSMF on the (800 Mpc)$^3$ P-Millennium DMO simulation, from machine learning models trained on the `L050AGN` (blue, dashed) and `L050AGN + Zoom` (blue, solid) simulation sets.

An alternative explanation is the well-known effect of baryons on their host dark matter haloes (e.g. Schaller et al. 2015; Velliscig et al. 2015). This may not only affect the masses of haloes, but also their mass distribution, changing the substructure on small scales, and hence the clustering measurement (van Daalen et al. 2014; Hellwing et al. 2016). To test whether this is causing the lower normalization at small scales, we extract a catalogue of features from the full hydro simulation (`L100Ref`) and use these as inputs to the `L050AGN + Zoom` model. We emphasize that these 'halo' features contain the contribution from both baryons and dark matter, but are otherwise identical to the features from a DMO simulation. The predicted clustering for this hybrid model application is shown in the second panel of Fig. 8; the normalization matches that of the `L100Ref` simulation, confirming that it is indeed baryonic effects causing the lower normalization on small scales. We stress that this is not strictly a fair use of the machine learning model, as it was trained on haloes from a DMO simulation, and as such the predictions should be taken with some caution. However, we argue this is a relatively 'clean' test of the impact of baryons on the halo, and the knock on effect on the clustering.

Other effects may also contribute to the deficit, such as differences in the parameters of the halo finder between DMO and hydro simulations, and for different resolution simulations. However, it seems clear that baryonic effects on haloes are a key contributor. A similar effect at small scales has been seen in semi-analytic models applied to DMO simulations (Contreras et al. 2015; Farrow et al. 2015). The machine learning model presented here allows us to cleanly test this effect on identical haloes.

We also compared the model predictions for the projected correlation function against those using a single subhalo feature (subhalo mass) to predict the stellar mass, applied to the P-Millennium volume. The normalization is underestimated in this simple model compared to the GAMA measurements, and this is particularly pronounced in the highest stellar mass bin. Full details are provided in Appendix B.

### 5.2 The GSMF

The left-hand panel of Fig. 9 shows the `L050AGN` model run on the `L100Ref` DMO simulation. We compare to the GSMF from the hydrodynamic `L100Ref` simulation, and it is clear that the high-mass end of the GSMF is not reproduced. While there are parameter differences between the models, it is not expected that the AGNdT9 model would fail to produce any $10^{12}$ M$_\odot$ galaxies in a (100 Mpc)$^3$ volume. In fact, the predictions broadly follow the model used for training, `L050AGN`, though underestimate the abundance of galaxies at the high-mass end ($> 10^{11}$ M$_\odot$). This additional underestimate is likely the result of a lack of training data at the high-mass end, due to the low number of high-mass galaxies in the `L050AGN` volume.

However, if we use the `L050AGN + Zoom` model we get much better agreement with the `L100Ref` simulation at the high-mass end. This demonstrates the effect of including the C-EAGLE zoom regions; the model is able to learn the baryonic properties of galaxies in the cluster regions, which are not present in `L050AGN`. Predictions at lower stellar masses are also consistent with both `L100Ref` and `L050AGN` down to $\sim 10^8$ M$_\odot$, the approximate resolution limit of the original simulations (Schaye et al. 2015), and where our predictions are approximately complete (see Appendix A).

We now turn our attention to the much larger P-Millennium DMO simulation. The right-hand panel of Fig. 9 shows predictions for the `L050AGN` and `L050AGN + Zoom` models on this volume, and while the former still completely misses the high-mass end, the model including zooms is able to predict stellar masses out to $\sim 10^{12}$ M$_\odot$. This extends the dynamic range of the GSMF beyond that accessible to the `L100Ref` hydrodynamic simulation, and

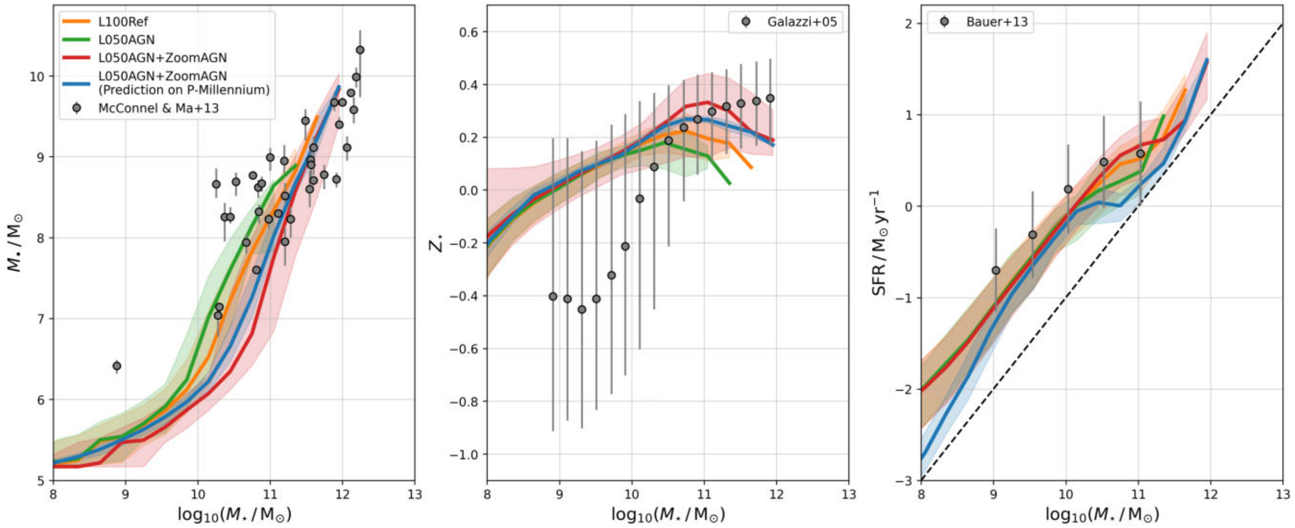

**Figure 10.** The black hole–stellar mass relation (left), stellar mass–metallicity relation (middle), and star-forming sequence (right). Observational constraints for each relation are shown, from McConnell & Ma (2013), Gallazzi et al. (2005), and Bauer et al. (2013), respectively. The relation in the `L100Ref` (orange), `L050AGN` (green), and `L050AGN + Zoom` (red) simulation sets is shown, as well as the predicted relation from the `L050AGN + Zoom` machine learning model applied to the P-Millennium simulation (blue). The median is given by the solid line in each case, and the 16th–84th percentile range is shown by the coloured shaded region. The dashed black line in the right-hand panel shows the cut used for passive galaxies, sSFR $< 10^{-11}$ yr$^{-1}$.

improves the statistics significantly. This is a significant achievement of the model – it is able to successfully extend the predictive range beyond that achievable with periodic hydrodynamic simulations. At lower stellar masses the predictions are consistent with both `L100Ref` and `L050AGN`. The predictions at the high-mass end are also in broad agreement with the observational constraints from Baldry et al. (2012).

The P-Millennium simulation is lower resolution than those used for training, which may impact the predicted properties of galaxies, particularly those close to the resolution limit. To test the impact of resolution, we applied the `L050AGN + Zoom` model to a lower resolution (100 Mpc)$^3$ DMO run, with eight times fewer particles. The predictions for the GSMF were identical, which confirms that differing resolution has no impact on the predicted properties; as long as the haloes are resolved, the halo features used for prediction are robust.

### 5.3 The black hole–stellar mass relation

We have demonstrated how the model is able to predict stellar masses with high accuracy, and produce a GSMF for the P-Millennium simulation volume. We now explore other key baryonic distribution functions. Fig. 10 shows the black hole–stellar mass relation, the stellar mass–metallicity relation, and the star-forming sequence. Each panel shows the relation in the `L100Ref`, `L050AGN`, and `L050AGN + Zoom` simulations, as well as the predicted relation for our `L050AGN + Zoom` model, with fiducial feature set, run on the P-Millennium simulation.

The black hole–stellar mass relation shows a rapid increase in the stellar mass above $M_\star \sim 10^{10}\,\mathrm{M}_\odot$, though the exact mass at which the relation turns upwards is dependent on the simulation. In `L050AGN + Zoom` the increase is at a higher mass compared to the two periodic simulations. This is not due to any parameter differences, since `L050AGN` has identical parameters, but may be due to the cluster environment delaying black hole accretion by starving the central regions of a galaxy of gas. Though van Son et al. (2019) note an excess of 'black hole monster galaxies' in cluster environments due to tidal stripping, this is a subdominant population compared to the main relation, so it does not increase the normalization of the black hole–stellar mass relation in these environments. The model predictions lie between the periodic and zoom relations, which is perhaps expected since both environments are providing training data from which the machine is making its predictions. Overall, the relation is predicted remarkably well, and the predictions extend the dynamic range to higher stellar and black hole masses than those achievable in `L100Ref` and `L050AGN`. At these higher masses the model is in good agreement with the observational results of McConnell & Ma (2013), though the scatter at fixed stellar mass is still underpredicted (as seen in Schaye et al. 2015).

### 5.4 The stellar mass–metallicity relation

Predictions from the model for the stellar mass–metallicity relation show similar behaviour. The model predictions lie between the relations from the periodic and zoom simulation sets at high stellar masses ($M_\star\,/\,\mathrm{M}_\odot > 10^{10}$), but closely follow the predictions below this, except at the very lowest stellar masses.

The scatter in both these relations is much tighter for the model predictions than in the original simulation sets. This is a reflection of the deterministic nature of the machine learning prediction, combined with the relatively limited feature set, which has been discussed in a number of previous works (e.g. Kamdar et al. 2016b; Moews et al. 2020). Historical halo features, such as the formation and assembly time, may help to increase the diversity of baryonic properties at fixed stellar mass. However, the predictions still lie within the uncertainties on observational constraints from Gallazzi et al. (2005) at all stellar masses.

### 5.5 The star-forming sequence

Finally, the right-hand panel of Fig. 10 shows the star-forming sequence, excluding passive galaxies (sSFR $< 10^{-11}$ yr$^{-1}$). As shown in Fig. 5 the model tends to underpredict SFRs of star-forming

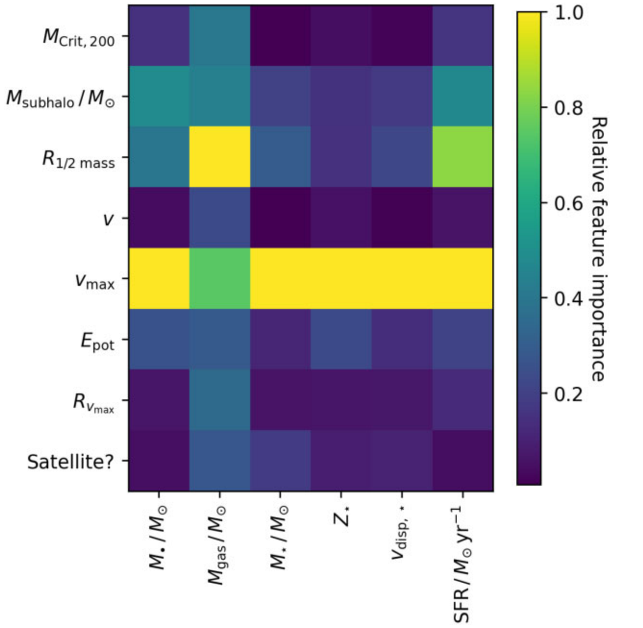

**Figure 11.** Matrix showing the relative importance (0 → 1, low to high importance) of each feature (*y*-axis) for each predictor quantity (*x*-axis), for the L050AGN + Zoom model. The importance is normalized by the maximum for each predictor.

galaxies, and this is reflected in the star-forming sequence, where the normalization at $M_* / M_\odot = 10^{11}$ is lower than that in the original simulation sets, by up to −0.8 dex compared to L050AGN + Zoom. The scatter at fixed stellar mass is comparable to the simulation sets (±0.25 dex), though this may be partly due to truly quiescent galaxies in the simulation sets that have residual star formation when predicted in the model. In general, however, the star-forming sequence is broadly reproduced, is in good agreement both above and below the characteristic mass, and lies within the uncertainties on observational constraints from Bauer et al. (2013).

## 6 FEATURE EXPLORATION

Feature importance in ERT can be evaluated from the relative position of a given feature in the tree; the closer to the root node in the ensemble of trees, the higher the importance. In order to evaluate the feature importance for each predictor, we retrain the model on each predictor individually. Fig. 11 shows a matrix of each predictor against each feature, coloured by their relative importance. The order of relative importance is generally the same for all predictors. $V_{\rm max}$ is by far the most important feature; Kamdar et al. (2016b) attributed a similarly high importance for $V_{\rm max}$ in their machine learning model trained on Illustris. A number of other studies have highlighted the importance of $V_{\rm max}$ for predicting baryonic properties. Matthee et al. (2017) showed that, in EAGLE, $V_{\rm max}$ is a key predictor of the stellar mass, more so than the halo mass. Chaves-Montero et al. (2016) use a SHAM technique to test the recovery of the clustering of galaxies in EAGLE and find a similarly strong dependence on $V_{\rm max}$. The circumgalactic medium mass fraction, at fixed halo mass, has also been shown to correlate strongly with $V_{\rm max}$ (when parametrized as a ratio with the virial circular velocity, closely related to the halo binding energy), in both EAGLE and Illustris (Davies et al. 2019, 2020; Oppenheimer et al. 2020); the authors of these studies argue that a high $V_{\rm max}$ corresponds to an early collapse time for a halo, which leads to greater black hole growth, which in turn ejects more of the circumgalactic medium mass. This has a big impact on the latter baryonic properties of the galaxy, such as its star formation history and morphology. This explains the strong importance of $V_{\rm max}$ in our feature set for the majority of our baryonic predictors.

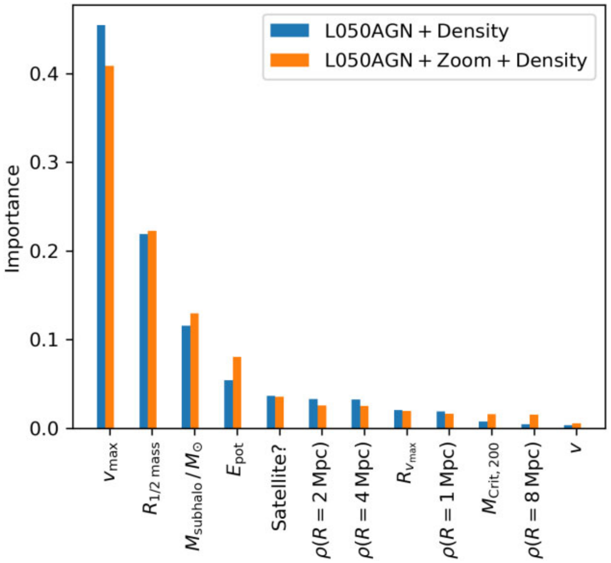

**Figure 12.** Relative feature importance as described by ERT, across all features simultaneously. L050AGN (blue) and L050AGN + Zoom (orange) machine learning models are shown, including additional features describing the local density on different scales ($\rho(R)$).

Interestingly, for the gas mass, the half-mass radius is instead the most important feature. $V_{\rm max}$ is still of high importance, but at a similar level to the subhalo mass and total halo mass ($M_{\rm crit,\,200}$). This suggests that the size of the underlying dark matter halo is closely related to its current gas mass. A similarly strong correlation between (H I) gas mass and size has been found observationally, though with the stellar component rather than dark matter (Catinella et al. 2012).

The peculiar velocity is the least important feature for all predictors, as expected. Interestingly, features that encode the local halo environment, such as $M_{\rm crit,\,200}$ and its status as a satellite or central, are also two of the least important features. This suggests that the properties of the subhalo itself mostly determine the baryonic properties, however this does not necessarily mean that 'nature' rather than 'nurture' is the dominant evolutionary process. Instead, other subhalo features may encode environmental information, e.g. satellites are clear outliers in the $M_{200} - M_{\rm subhalo}$ plane.

We also evaluated the effect of including local density features, $\rho(R)$. Fig. 12 shows the feature importance for all predictors, in the L050AGN and L050AGN + Zoom machine learning models. None of these local features dominates the feature importance, but the density on intermediate scales ($R$ = [2, 4] Mpc) has a higher importance than on the smallest and largest scales ($R$ = [1, 8] Mpc, respectively). The order of feature importance is otherwise mostly preserved.

## 7 DISCUSSION AND CONCLUSIONS

We have demonstrated the effectiveness of machine learning methods in modelling the complex relationships between galaxies and their host haloes by training a machine learning model to directly learn this mapping. By combining hydro and DMO simulations we avoid baryonic effects on haloes that would bias predictions. And by using a training set consisting of both periodic and zoom simulations of galaxy clusters, we include rare environments that may not be present in typical periodic simulations, allowing the model to be applied to much larger volume DMO simulations, increasing the dynamic range, and allowing the evaluation of clustering statistics over much larger scales. Our conclusions are as follows:

(i) The model successfully predicts the stellar mass, stellar velocity dispersion, and black hole mass, and provides reasonable predictions for the star formation rate, stellar metallicity, and total gas mass. Even where the stellar metallicity shows some dispersion in the prediction, the overall distribution is recovered.

(ii) Star formation rates and gas masses are biased low due to the effect of quiescent, gas-poor galaxies, and some suggestions for improving this are put forward, including the use of historical halo features.

(iii) Adding features representing the local density leads to a negligible increase in the predictive accuracy for most properties, except the gas mass, which shows significant improvement, particularly in cluster environments.

(iv) We apply the trained model to the P-MILLENNIUM simulation and analyse the projected two-point correlation function. We are able to predict the clustering of galaxies out to much larger scales than in the periodic hydro simulations ($> 10\,h^{-1}$ Mpc), as well as analyse the clustering of rarer, high-mass galaxies, and find that EAGLE is in good agreement with observational constraints from GAMA on large scales. On smaller scales we conclude that baryonic effects on haloes affect the clustering statistics.

(v) The predicted GSMF is in excellent agreement with that given by the periodic hydro simulations at low and intermediate masses, and extends the relation to higher masses.

(vi) The black hole–stellar mass and stellar mass–metallicity relations are well reproduced, though with less scatter, as seen in other machine learning models.

(vii) The normalization of the star-forming sequence is slightly underpredicted at the characteristic mass, which reflects both the lower normalization in the training data, but also the lower predicted stellar masses on the test set. However, the general form is in good agreement.

(viii) $V_{\rm max}$ is the most important feature in all simulation sets. Measures of the local environment, such as the satellite flag, host halo mass, and local density, do not show high importance in any of the models.

We stress that our model is not intended as a replacement of traditional galaxy formation models: it is in fact wholly reliant on such models to train from. It does, however, provide a means of expanding the predictions from such models to much larger periodic volumes. These larger volumes are useful for a number of science questions. Galaxy clustering is a particularly important application we have demonstrated here, allowing us to test the clustering statistics of high-resolution hydrodynamic simulations in the high-mass, large-separation regime. As demonstrated by Jo & Kim (2019), additional features, such as the halo merger history and its assembly and formation time, are expected to have a significant positive impact on the prediction accuracy. While we have found that features describing the local environment are not highly important, additional parameters describing, for example, the tidal shear (e.g. Lucie-Smith, Peiris & Pontzen 2019) may also encode more useful information for the machine to learn from. It may also be possible to make predictions at multiple redshifts simultaneously by providing the machine with the scale factor, as demonstrated in Moster et al. (2021).

The C-EAGLE sample provides a wealth of training data on rich cluster environments, however those environments on the opposite end of the overdensity distribution, extreme underdensities, or *cosmic voids* are less well sampled in our training set. Void regions do not have as obvious an effect on their constituent galaxies properties as rich cluster environments, where galaxy mergers are far more common and extreme processes such as ram-pressure stripping occur, however noticeable effects are still seen in voids in the fiducial periodic EAGLE volumes (Paillas et al. 2017; Xu et al. 2020). Larger, more significantly underdense regions are, as for overdense regions, not well sampled in the periodic volumes, however such voids are an important constituent of the Universe, making up ~60 per cent of the cosmic volume (Pan et al. 2012). In future work we will use resimulations of a range of overdensities down to low redshift to better populate this region of overdensity space, reducing generalization errors for galaxies in these environments.

We have focused on six key baryonic properties, but other baryonic properties are simple to add, including the emission properties of galaxies if combined with post-processing pipelines. This will allow for the construction of extremely large light-cones (as demonstrated in Hearin et al. 2020, using their empirical modelling plus simulation-calibrated approach), necessary for making predictions for wide-field surveys from the upcoming Roman and Euclid space-based observatories (Potter, Stadel & Teyssier 2017). To this end, in future work we will explore predictions during the epoch of reionization, where we will leverage the FLARES simulations (Lovell et al. 2021). A unique aspect of FLARES is that it consists of resimulations of a range of overdensities, providing training data in extreme overdense and underdense environments, which may aid predictions of galaxy properties across all environments.

## ACKNOWLEDGEMENTS

The authors wish to thank the anonymous referee for a detailed report that improved this manuscript. We also wish to thank Daniel Farrow for providing the GAMA clustering data, John Helly for help reading the P-Millennium data, and Aswin Vijayan, Rob Crain, Joop Schaye, and Scott Kay for helpful comments and discussions. We also thank the EAGLE team for their efforts in developing the EAGLE simulation code.

We acknowledge the following open-source software packages used in the analysis: scipy (Virtanen et al. 2020), Astropy (Robitaille et al. 2013), matplotlib (Hunter 2007), seaborn (Waskom 2021), and halotools (Hearin et al. 2017, v0.7).

This work used the DiRAC@Durham facility managed by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). The equipment was funded by BEIS capital funding via STFC capital grants ST/K00042X/1, ST/P002293/1, ST/R002371/1, and ST/S002502/1, Durham University, and STFC operations grant ST/R000832/1. DiRAC is part of the National e-Infrastructure. PAT acknowledges support from the STFC (grant number ST/P000525/1). CCL acknowledges support from the Royal Society under grant RGF/EA/181016. CMB acknowledges support from STFC grant ST/T000244/1. MS is supported by the Netherlands Organisation for Scientific Research (NWO) through VENI grant 639.041.749. YMB gratefully acknowledges funding from the NWO through VENI grant number 639.041.751.

## DATA AVAILABILITY

The public EAGLE database can be used to access the subhalo properties for the periodic hydrodynamic simulations in this paper

(McAlpine et al. 2016). Other data underlying this article will be shared on reasonable request to the corresponding author. The code used to train and analyse the models, and produce all plots, is made available at github.com/christopherlovell/ML-cosmo.

## REFERENCES

Agarwal S., Davé R., Bassett B. A., 2018, MNRAS, 478, 3410
Artale M. C. et al., 2017, MNRAS, 470, 1771
Bahé Y. M. et al., 2016, MNRAS, 456, 1115
Bahé Y. M. et al., 2017, MNRAS, 470, 4186
Baldry I. K. et al., 2012, MNRAS, 421, 621
Ball N. M., Brunner R. J., 2010, Int. J. Mod. Phys. D, 19, 1049
Barnes D. J., Kay S. T., Henson M. A., McCarthy I. G., Schaye J., Jenkins A., 2017a, MNRAS, 465, 213
Barnes D. J. et al., 2017b, MNRAS, 471, 1088
Bauer A. E. et al., 2013, MNRAS, 434, 209
Baugh C. M., 2006, Rep. Prog. Phys., 69, 3101
Baugh C. M. et al., 2019, MNRAS, 483, 4922
Behroozi P. S., Conroy C., Wechsler R. H., 2010, ApJ, 717, 379
Benson A. J., 2010, Phys. Rep., 495, 33
Catinella B. et al., 2012, A&A, 544, A65
Chabrier G., 2003, PASP, 115, 763
Chaves-Montero J., Angulo R. E., Schaye J., Schaller M., Crain R. A., Furlong M., Theuns T., 2016, MNRAS, 460, 3100
Contreras S., Baugh C. M., Norberg P., Padilla N., 2015, MNRAS, 452, 1861
Crain R. A. et al., 2009, MNRAS, 399, 1773
Crain R. A. et al., 2015, MNRAS, 450, 1937
Crain R. A. et al., 2017, MNRAS, 464, 4204
Davé R., Thompson R. J., Hopkins P. F., 2016, MNRAS, 462, 3265
Davé R., Anglés-Alcázar D., Narayanan D., Li Q., Rafieferantsoa M. H., Appleby S., 2019, MNRAS, 486, 2827
Davies J. J., Crain R. A., McCarthy I. G., Oppenheimer B. D., Schaye J., Schaller M., McAlpine S., 2019, MNRAS, 485, 3783
Davies J. J., Crain R. A., Oppenheimer B. D., Schaye J., 2020, MNRAS, 491, 4462
Davis M., Peebles P. J. E., 1983, ApJ, 267, 465
Dolag K., Borgani S., Murante G., Springel V., 2009, MNRAS, 399, 497
Farrow D. J. et al., 2015, MNRAS, 454, 2120
Feng Y., Di-Matteo T., Croft R. A., Bird S., Battaglia N., Wilkins S., 2016, MNRAS, 455, 2778
Fluke C. J., Jacobs C., 2020, WIREs Data Mining and Knowledge Discovery, 10, e1349
Furlong M. et al., 2015, MNRAS, 450, 4486
Furlong M. et al., 2017, MNRAS, 465, 722
Gallazzi A., Charlot S., Brinchmann J., White S. D. M., Tremonti C. A., 2005, MNRAS, 362, 41
Geurts P., Ernst D., Wehenkel L., 2006, Mach Learn, 63, 3
Gonzalez-Perez V., Lacey C. G., Baugh C. M., Lagos C. D. P., Helly J., Campbell D. J. R., Mitchell P. D., 2014, MNRAS, 439, 264
Hearin A. P. et al., 2017, AJ, 154, 190
Hearin A., Korytov D., Kovacs E., Benson A., Aung H., Bradshaw C., Campbell D., LSST Dark Energy Science Collaboration, 2020, MNRAS, 495, 5040
Hellwing W. A., Schaller M., Frenk C. S., Theuns T., Schaye J., Bower R. G., Crain R. A., 2016, MNRAS, 461, L11
Henriques B. M. B., White S. D. M., Thomas P. A., Angulo R., Guo Q., Lemson G., Springel V., Overzier R., 2015, MNRAS, 451, 2663
Henriques B. M. B., Yates R. M., Fu J., Guo Q., Kauffmann G., Srisawat C., Thomas P. A., White S. D. M., 2020, MNRAS, 491, 5795
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Icaza-Lizaola M., Bower R. G., Norberg P., Cole S., Schaller M., Egan S., 2021, MNRAS, 507, 4584
Jo Y., Kim J.-h., 2019, MNRAS, 489, 3565
Kamdar H. M., Turk M. J., Brunner R. J., 2016a, MNRAS, 455, 642
Kamdar H. M., Turk M. J., Brunner R. J., 2016b, MNRAS, 457, 1162
Katz N., White S. D. M., 1993, ApJ, 412, 455
Lagos C. d. P. et al., 2015, MNRAS, 452, 3815
Legrand L. et al., 2019, MNRAS, 486, 5468
Lovell C. C., Vijayan A. P., Thomas P. A., Wilkins S. M., Barnes D. J., Irodotou D., Roper W., 2021, MNRAS, 500, 2127
Lucie-Smith L., Peiris H. V., Pontzen A., 2019, MNRAS, 490, 331
McAlpine S. et al., 2016, Astron. Comput., 15, 72
McAlpine S., Bower R. G., Harrison C. M., Crain R. A., Schaller M., Schaye J., Theuns T., 2017, MNRAS, 468, 3395
McCarthy I. G., Schaye J., Bird S., Le Brun A. M. C., 2017, MNRAS, 465, 2936
McConnell N. J., Ma C.-P., 2013, ApJ, 764, 184
Matthee J., Schaye J., Crain R. A., Schaller M., Bower R., Theuns T., 2017, MNRAS, 465, 2381
Mitchell P. D., Schaye J., 2021, preprint (arXiv:2103.10966)
Moews B., Davé R., Mitra S., Hassan S., Cui W., 2021, MNRAS, 504, 4024
Moster B. P., Somerville R. S., Maulbetsch C., van den Bosch F. C., Macciò A. V., Naab T., Oser L., 2010, ApJ, 710, 903
Moster B. P., Naab T., White S. D. M., 2013, MNRAS, 428, 3121
Moster B. P., Naab T., Lindström M., O'Leary J. A., 2021, MNRAS, 507, 2115
Neistein E., Khochfar S., Dalla Vecchia C., Schaye J., 2012, MNRAS, 421, 3579
Oppenheimer B. D. et al., 2020, MNRAS, 491, 2939
Paillas E., Lagos C. D. P., Padilla N., Tissera P., Helly J., Schaller M., 2017, MNRAS, 470, 4434
Pan D. C., Vogeley M. S., Hoyle F., Choi Y.-Y., Park C., 2012, MNRAS, 421, 926
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Peebles P. J. E., 1980, The Large-Scale Structure of the Universe. Princeton Univ. Press, Princeton, NJ
Pillepich A. et al., 2018, MNRAS, 473, 4077
Planck Collaboration I, 2014, A&A, 571, A1
Potter D., Stadel J., Teyssier R., 2017, Comput. Astrophys. Cosmol., 4, 2
Robitaille T. P. et al., 2013, A&A, 558, A33
Safonova S., Norberg P., Cole S., 2021, MNRAS, 505, 325
Schaller M. et al., 2015, MNRAS, 451, 1247
Schaye J. et al., 2015, MNRAS, 446, 521
Scott D. W., 1979, Biometrika, 66, 605
Simha V., Weinberg D. H., Davé R., Fardal M., Katz N., Oppenheimer B. D., 2012, MNRAS, 423, 3458
Sinigaglia F., Kitaura F.-S., Balaguera-Antolínez A., Nagamine K., Ata M., Shimizu I., Sánchez-Benavente M., 2021, ApJ, 921, 66
Somerville R. S., Davé R., 2015, ARA&A, 53, 51
Stiskalek R., Desmond H., Holvey T., Jones M. G., 2021, MNRAS, 506, 3205
Stone M., 1974, J. R. Stat. Soc. B, 36, 111
Sullivan D., Iliev I. T., Dixon K. L., 2018, MNRAS, 473, 38
Tormen G., Bouchet F. R., White S. D. M., 1997, MNRAS, 286, 865
Trayford J. W. et al., 2015, MNRAS, 452, 2879
Trayford J. W. et al., 2017, MNRAS, 470, 771
van Daalen M. P., Schaye J., McCarthy I. G., Booth C. M., Dalla Vecchia C., 2014, MNRAS, 440, 2997
van Son L. A. C. et al., 2019, MNRAS, 485, 396
Velliscig M. et al., 2015, MNRAS, 454, 3328
Vijayan A. P., Lovell C. C., Wilkins S. M., Thomas P. A., Barnes D. J., Irodotou D., Kuusisto J., Roper W. J., 2021, MNRAS, 501, 3289
Virtanen P. et al., 2020, Nat. Methods, 17, 261
Vogelsberger M. et al., 2014, MNRAS, 444, 1518
Waskom M. L., 2021, J. Open Source Softw., 6, 3021
Wechsler R. H., Tinker J. L., 2018, ARA&A, 56, 435
Weinberg D. H., Colombi S., Davé R., Katz N., 2008, ApJ, 678, 6
Xu W. et al., 2020, MNRAS, 498, 1839
Xu X., Ho S., Trac H., Schneider J., Poczos B., Ntampaka M., 2013, ApJ, 772, 147

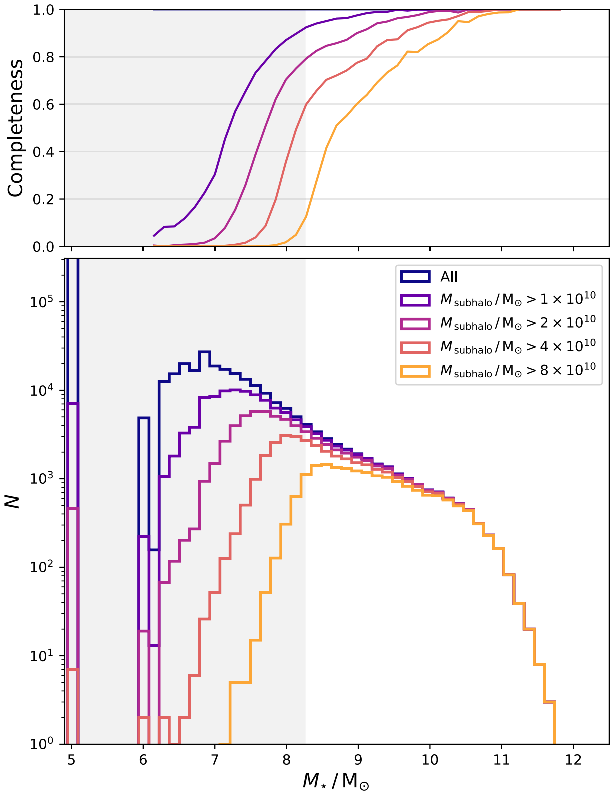

**Figure A1.** Histogram of stellar masses in the L100Ref simulation for different DMO subhalo mass limits from the matched subhalo list. The grey shaded area designates the stellar mass resolution limit ($M_\star$ / M$_\odot$ > 1.8 × $10^8$, or 100 star particles at the initial baryon mass).

## APPENDIX A: STELLAR MASS COMPLETENESS

Before model training we pre-select haloes based on their dark matter properties only to ensure the same selection can be applied to any DMO simulation the model is applied to. This is intended to avoid a situation where a model is applied to haloes with properties that were not present in the training set. Since the selection is done on DMO properties only, we here check whether galaxies below the resolution limit in the hydro simulation are included, and the incompleteness of galaxies above the resolution limit. Fig. A1 shows a histogram of stellar mass in the L100Ref simulation for different DMO subhalo mass cuts. For even the strictest subhalo mass limit there are large numbers of subhaloes with stellar masses below the resolution limit; this suggests their baryonic properties are highly unresolved. However, the important quantity is the completeness at fixed stellar mass. For a subhalo mass limit of $M_{\rm subhalo}$ / M$_\odot$ > $10^{10}$ the completeness is greater than 95 per cent above the stellar mass resolution limit ($M_\star$ / M$_\odot$ > 1.8 × $10^8$, approximately equal to 100 star particles at the initial baryon mass, i.e. ignoring stellar evolution mass-loss), and 100 per cent complete above 5 × $10^9$ M$_\odot$. We use a subhalo mass limit of $M_{\rm subhalo}$ / M$_\odot$ > $10^{10}$ throughout the rest of the text.

## APPENDIX B: ISOTONIC FITS TO A SINGLE FEATURE

In order to provide a qualitative assessment of the ERT model we choose to fit a simple model to the relationship between each predictor and a *single* feature. We use subhalo mass and $V_{\rm max}$ as our chosen features as these are commonly used in SHAM approaches. We fit each relation with an Isotonic regression model, which ensures monotonicity. We do this for the training set, and evaluate the performance on the test set. Each relation and the corresponding fits are shown in Fig. B1. The percentage of galaxies where the predicted value is within 0.2 dex of the true value is quoted in each panel. In each case this percentage is lower than that achieved with the ERT model.

We also show the Pearson correlation coefficient for the ERT model as well as the Isotonic regression model for each feature in Fig. B2. The ERT model outperforms the Isotonic regression model for all predictors, though the performance is comparable using $V_{\rm max}$ for the stellar mass, stellar velocity dispersion, and stellar metallicity. This is expected from the strong correlation between the predictor and $V_{\rm max}$ in each of these cases, shown in Fig. B1. Fig. 11 also shows that these three predictors are particularly dependent on $V_{\rm max}$, whereas other predictors have greater contributions from other features. It is also interesting to see that subhalo mass is the more accurate predictor for gas mass, black hole mass, and star formation rate, compared to $V_{\rm max}$, which highlights that using one or the other feature in an SHAM approach may not lead to optimized predictions for all galaxy features – the ML approach, on the other hand, simply incorporates all features, and chooses the best for each predictor.

In Fig. B3, we show the impact of using the Isotonic regression model (using subhalo mass as the feature) on the projected correlation function and the GSMF. The GSMF is mostly reproduced, as expected due to the strong correlation between feature and predictor. However, the projected correlation function (for 11 < $M_\star$ / M$_\odot$ < 11.5) shows a deficit in the normalization compared to the ERT model, particularly on small scales. One explanation is that high-mass satellite galaxies, which are not common in the training set, may be more common in the larger P-Millennium volume. The ERT model then handles these objects better than the Isotonic model, utilizing other features that are more important in these environments (e.g. the satellite flag).

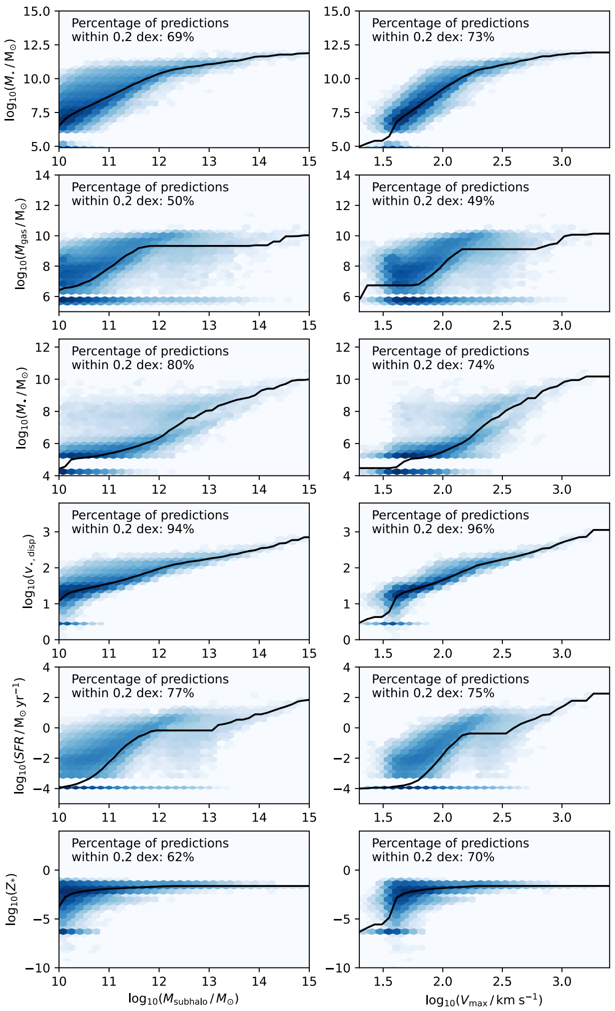

**Figure B1.** Relations between features commonly used in SHAM approaches (Subhalo mass and $V_\mathrm{max}$; $x$-axis) and each predictor ($y$-axis). Each panel shows a 2D histogram of the distribution (blue) alongside a fitted monotonic linear relation (black line). The percentage of galaxies where the predicted value is within 0.2 dex of the true value is quoted in each panel.

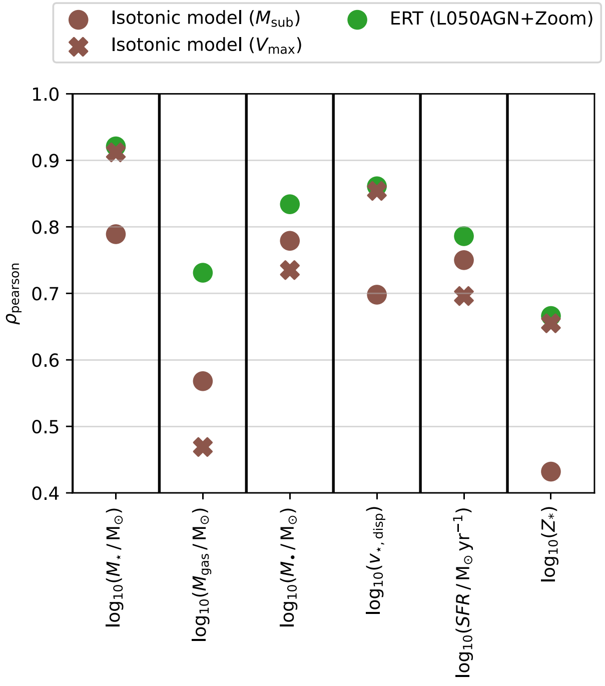

**Figure B2.** Pearson correlation coefficient for the ERT model (L050AGN + ZoomAGN) as well Isotonic regression models trained using subhalo mass and $V_\mathrm{max}$. Each predictor is shown on the $x$-axis.

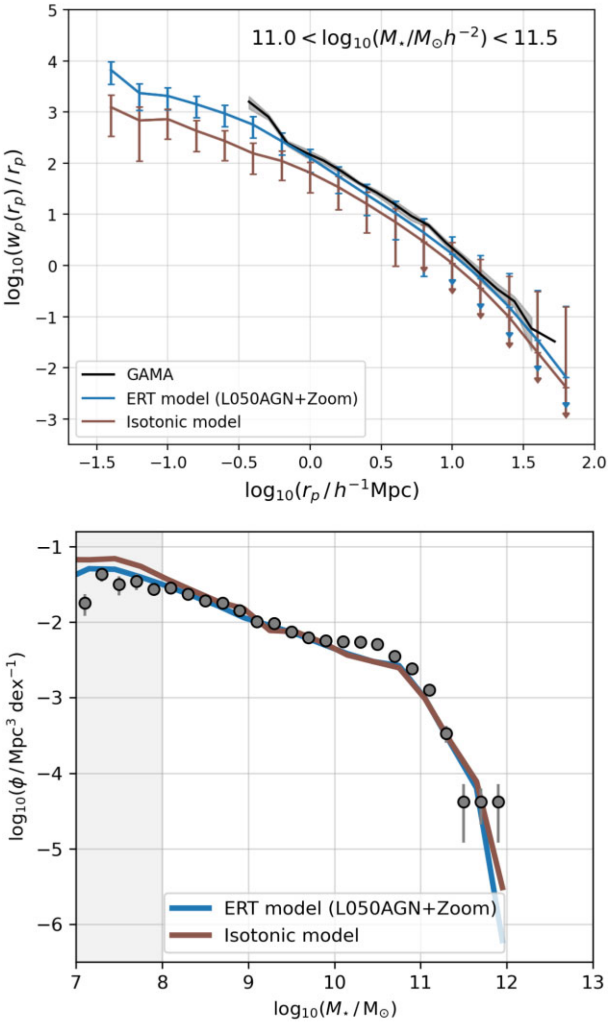

**Figure B3.** Predictions for the projected correlation function (top panel) and GSMF (bottom panel) using the Isotonic regression model (using subhalo mass; brown lines), compared with the ERT model (blue) with all features.

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