Predicting galaxy properties from dark matter-only simulations using machine learning

Eddie Chua

Fall 2020

1 Problem Statement

The standard cold dark matter (ΛCDM) model has been successful in describing several properties of the universe, such as the Cosmic Microwave Background, and the formation and evolution of large-scale structure in the distribution of galaxies throughout the universe (e.g. Spergel *et al.*, 2003; Springel *et al.*, 2005). Within this paradigm, primordial Gaussian fluctuations lead to the growth of cold dark matter haloes, which form successively larger structures by accreting diffuse matter and merging with other haloes. Although the accreted haloes are subject to various disruptive processes (tidal stripping, tidal shocking and ram-pressure stripping etc.), many of these haloes are not completely destroyed, surviving as gravitationally bound subhaloes in their host haloes. The properties and evolution of haloes and subhaloes were initially studied using N-body simulations, which evolves the positions and velocities of a large number of massive (dark matter) particles over time. By being able to resolve these substructures, early N-body cosmological simulations have been important in understanding the under different environments (e.g. Springel *et al.*, 2001; Gao *et al.*, 2004; Macciò *et al.*, 2006; Diemand *et al.*, 2007; Springel *et al.*, 2008; Angulo *et al.*, 2009).

Under the two-stage hierarchical formation scenario proposed by White and Rees (1978), cooling of gas and subsequent star formation in the potential wells provided by these dark matter (DM) haloes give rise to luminous galaxies that trace the underlying DM distribution. Three main processes that are thought to be involved at this stage are cooling, star formation and feedback. In addition, many other processes such as dynamical friction, ram pressure, adiabatic contraction, black hole formation and growth etc. are also thought to contribute to galaxy formation and evolution. Because N-body simulations do not track the evolution of baryons, different techniques have been developed to map the properties of these simulated DM haloes to observables such as the galaxy luminosity function. These methods include semi-analytical modelling (SAM, e.g. Somerville and Primack, 1999; Benson *et al.*, 2003; Bower *et al.*, 2006) and abundance matching (e.g. Conroy and Wechsler, 2009; Guo *et al.*, 2010). SAM methods involve hybrid simulations which combine an N-body treatment of the dark matter and a semi-analytic treatment of the baryons. Here, instead of integrating the equations of motions of the baryons, all baryon physics are modelled using assumptions that are physically and observationally motivated. However, hybrid methods primarily based on N-body simulations are unable to provide a complete picture of galaxy formation, since baryons and DM are coupled and their co-evolution can have a significant impact on the structure of both DM haloes and subhaloes.

For this reason, hydrodynamic simulations, where the incorporation of galaxy formation processes allows the complex modelling of galaxies to occur in a full cosmological context, have become popular in recent years. In these simulations, the equations of motion of both the dark matter and the baryons are integrated alongside each other, allowing certain processes such as accretion and mergers to emerge naturally. Recent efforts have accelerated the development of cosmological hydrodynamics simulations, e.g. Illustris (Vogelsberger *et al.*, 2014a), EAGLE (Crain *et al.*, 2015), Illustris-TNG (Pillepich *et al.*, 2018) which model the formation of galaxies over a large range of masses. Although these work have been found to be capable of producing realistic galaxies, such simulations often span wide orders of magnitudes in mass and are computationally expensive, placing limits on both the mass resolution of objects, as well as the size of the simulation box. For example, the Illustris project simulated the evolution of 1820^3 dark matter particles and 1820^3 fluid (baryonic) elements from redshift $z = 127$ to the current epoch ($z = 0$) and required 19 million CPU hours for completion.

In this project, we propose an intermediate approach between the N-body based analytical approaches and the full hydrodynamic methods based on machine-learning. The goal will be to predict the results of hydrodynamic cosmological simulations using only N-body haloes. Ultimately, the results of such a trained model can subsequently be applied to more recent N-body simulations, which are performed in much larger box sizes ($\gtrsim 1$ Gpc^3) compared to hydrodynamic simulations (~ 100 Mpc^3), allowing for much better galaxy statistics for comparison with large-scale structure observations.

2 Data Source

The analysis presented here is based on the Illustris project (Vogelsberger *et al.*, 2014b,a; Genel *et al.*, 2014; Sijacki *et al.*, 2015), a series of cosmological simulations encompassing a volume 106.5 Mpc[1] a side and evolved in a ΛCDM cosmology consistent with WMAP-9 results. The simulations of the Illustris suite were carried out using the AREPO code (Springel, 2010), where the hydrodynamical equations are solved on a moving Voronoi mesh using a finite volume method.

The suite includes three realizations at different resolutions including gravity, hydrodynamics and key physical processes for galaxy formation. In this project, I will rely on the highest resolution run – Illustris-1, which follows 1820^3 DM particles and 1820^3 gas cells, with a mass resolution of $6.26 \times 10^6 M_\odot$ and $1.26 \times 10^6 M_\odot$ (for DM and baryons, respectively). For comparison to the full-physics run, an analogue DM-only (N-body) counterpart – Illustris-1-Dark is available, with the same initial conditions and corresponding resolution. I will focus only on the final snapshot (redshift $z = 0$[2]).

The data for Illustris is publicly available online (`https://www.illustris-project.org/data/`). This includes all 136 simulation snapshots (with all particle positions and velocities), as well as post-processed group catalogues of individual galaxies and haloes. Haloes, subhaloes, and their basic properties are obtained with the FOF and SUBFIND algorithms (Davis *et al.*, 1985; Springel *et al.*, 2001; Dolag *et al.*, 2009), at each of the 136 stored snapshots. A friends-of-friends (FOF) group finder (with linking length 0.2), which is an unsupervised clustering algorithm, is used to identify haloes. The SUBFIND algorithm then locates gravitationally bound objects (subhaloes) which are characterized hierarchically. The results of the FOF and SUBFIND algorithms are summarized in the publicly available group catalogues, which contain halo and galaxy properties such as the halo mass, stellar mass, star formation rate and photometrics. Concentrating on well-resolved haloes, Illustris contains ≈ 14000 of mass $10^{11} - 10^{14.5} M_\odot$. Some visualizations of the simulation box are shown in Figure 1a and 1b.

3 Methodology

3.1 Input features and pre-processing

The following table lists the features used for training, drawn from halo properties in the N-body simulation Illustris-Dark:

[1] 1 pc is $\approx 3.086 \times 10^{13}$ km.

[2] Note that larger redshift z corresponds to a younger universe.

Training Feature	Cat.	Log.	Description
halo mass M_{200}	✓	✓	total mass contained within the virial radius R_{200}, where R_{200} is the radius where the average density is 200 times that of the critical density of the universe.
halo velocity v	✓	✓	magnitude of halo velocity.
halo spin λ	✓	✓	magnitude of the halo spin vector, which denotes the halo angular momentum.
halo offset	✓	✓	ratio of distance between center of mass and location of the potential minimum, to the virial radius. A smaller offset denotes a more relaxed (less disturbed) halo.
halo concentration parameter c_{-2}			characterizes the density profile of the halo. Larger c_{-2} implies that the halo density is more centrally concentrated.
halo formation redshifts $z_{1/2}, z_{3/4}$			denotes the redshift when a halo has accreted 1/2 and 3/4 of its current mass. Smaller halo formation redshifts implies that the halo formed later.
halo shape parameters q_{15}, q_{200}, s_{15}, s_{200}			measures the halo shapes using an ellipsoidal approximation, where $q \equiv b/a$ and $c \equiv c/a$ and $a > b > c$ are the axis lengths of the density ellipsoids. $q = s = 1$ denotes a completely spherical halo. The two subscripts refer to the two radii where the halo shapes have been extracted and calculated.
current number of subhaloes $N_{\rm sub}$	✓		
velocity anisotropy β			denotes whether the orbits of the particles are tangentially biased ($\beta < 0$), radially biased ($\beta > 0$), or isotropic ($\beta = 0$).
measure of the surrounding "density" $D_{n,f=0.1}$, for $n = 1, 2, 3, 4, 5$	✓		$D_{n,f}$ is defined as the 3D distance to the nth nearest neighbour with a viral mass that is at least f times that of the halo under consideration, divided by the virial radius of the nth nearest neighbour.

Table 1: Input features used in this project. A check-mark in the second column denotes that the feature can be obtained directly from the SUBFIND catalogue. A check-mark in the third column denotes that the logarithm of the feature is taken.

All these features are taken to be continuous variables, with the exception of $N_{\rm sub}$, which is discrete. The features with check-marks are those that are either i) present in the online SUBFIND catalogue, or ii) can be easily calculated directly from it. Note that the remaining features have to be calculated from the full snapshots using the 6D (positions and velocities), which I have already done in previous work. Here, I have left out the detailed steps for these calculations and in favor of a more general description.

In the training data, the logarithm of some of these features are used, to 1) reduce their dynamic range (e.g. the halo mass spans > 3-4 orders of magnitude), and 2) to symmetrize the distribution (closer to normal). This is not possible for features which take on zero or negative values.

The data has also been pre-processed to remove haloes that have parameters with un-physical values. This means ensuring that:

1. halo formation redshifts ($z_{1/2}$, $z_{3/4}$) is between 0 and 127,
2. the concentration parameter $c_{-2} > 1$
3. the shape parameters q, s are larger than zero (and smaller than one).

These un-physical values can result when the algorithms in the calculations fail or do not converge properly. Such errors occur because some haloes have disturbed structures which are far from typical (e.g. haloes in the process of merging), rendering the calculation of the features or parameters inaccurate. Following the same

reasoning, I also remove outliers lying more than 5 standard deviations from the mean for each parameter. Finally, the features are standardized to zero-mean and unit standard deviation.

Note that because due to the nature of gravitational collapse and hierarchical structure formation, the number of haloes is dependent on the halo mass, approximately $n(M) \propto \frac{1}{M}$. Thus, low-mass haloes out-number high-mass ones, which can render it more difficult to reproduce characteristics on the high-mass end.

3.2 Output variables

In this project, we will focus on evaluating how well the following six halo and galaxy properties in the hydrodynamic simulation can be predicted:

Output variables	**Log.**	**Description**
halo mass M_{200}	✓	same as before, but measured in Illustris-1
galaxy stellar mass $m_{\rm gal}$	✓	total stellar mass in the halo
black hole mass $m_{\rm BH}$	✓	mass of central supermassive black hole
star formation rate SFR	✓	rate (per year) at which stars are formed in the halo
halo shape parameter s_{15}		same as before, but only for parameter s and for the shape in the inner regions of the halo
circularity parameter $f_{\rm circ}$		denotes the galaxy morphology, calculated from the fraction of stars considered as 'disk' stars. Larger and smaller values of $f_{\rm circ}$ are associated with disk galaxies and elliptical galaxies respectively.

Table 2: Features to be predicted, extracted for haloes in the hydrodynamic simulation Illustris-1.

Note that I will refer to the stellar component of the haloes as *galaxies*. As performed previously for the input features, haloes with un-physical values and outliers are removed. Here, we mainly ensure that the star formation rate is finite and that $f_{\rm circ} >= 0$.

3.3 Matching between inputs and outputs

Note that because the two simulations Illustris and Illustris-Dark are technically independent simulations, further processing has to be done to so that the input and output features described in the previous subsections are matched on a halo-to-halo basis. Fortunately, because both simulations are started from identical initial conditions, such a matching procedure can be done using the unique particle IDs in the simulations, with the precise strategy described in Rodriguez-Gomez *et al.* (2017). Briefly, for any given halo in Illustris, the matching halo in Illustris-Dark is the halo that contains the largest fraction of these IDs. Similarly, the process can be repeated starting from a subhalo in Illustris-Dark to find a match in Illustris. The final matched halo catalogue consists of haloes with successful matches in both directions.

After the removal of un-physical values, outliers and the matching procedure, the dataset consists of 12728 matched haloes. 80% of the haloes are randomly selected to be in the training set, and the remaining 20% forms the test set for model evaluation.

3.4 Artificial neural network

The task in this project is a supervised learning problem to map halo properties of Illustris-Dark to galaxy and halo properties in the hydrodynamic simulation Illustris. To address this supervised learning problem, I explore the use of an artificial neural network (ANN) to perform the regression. Since the relation between the Illustris and Illustris-Dark haloes can be non-linear, a neural network offers a flexible architecture to capture the non-linearities. Although we concentrate only on the Illustris simulation, which represents only a single realization with a particular choice of physics parameters, a neural network also offers the possibility of taking into account different simulation parameters in future work, enabling the predictions to be conditioned on the choice of physics parameters in the hydrodynamic simulations.

For the ANN, I used a fully-connected network, implemented using the Keras framework in TENSORFLOW 2. The hyper-parameters for the ANN are shown in Table 3 below:

Hyperparameter	**Chosen value**	**Values tested**
Hidden layers	2	1,2,3
Units per layer	64	16,32,64
Activation function	ReLU	
Regularization	$\ell 2$ with $\lambda = 10^{-3}$	
Loss function	MSE (for M_{200}, $m_{\rm gal}$, m_{BH}, SFR) cross-entropy for (s_{15} and $f_{\rm circ}$)	MSE, MAE
Optimizer	Adam	
Batch size	256	
Training epochs	450	200-500

Table 3: Hyperparameters of the fully-connected artificial neural network applied in this present study.

For the first four output features ($m_{\rm gal}$, m_{BH}, SFR), a single ANN with multiple outputs is used. Thus the final layer is a linear layer consisting of four output nodes, with mean-squared error (MSE) as the loss function.

Since the final two output features (s_{15} and $f_{\rm circ}$) are fractions lying between zero and one, a separate neural network is trained. Here, the final layer has two output nodes with sigmoid activation. The general cross-entropy is used for the loss function (i.e. $L = -\sum_{x\in \text{ data}} p(x) \log \hat{p}(x)$).

The hyper-parameters have been chosen to give good accuracy, measured by 1) the MSE (or equivalently, R^2 value) across the output features on the validation set, and 2) the ability to reproduce the galaxy mass – halo mass, and the black hole mass – halo mass relations.

I found the ANN performance to be sensitive to the number of hidden layers and units but largely insensitive to the choice of the loss function (between MSE and MAE). For example, although using 1 hidden layer and 32 units results in a small loss (MSE), the trained ANN is unable to capture the galaxy mass – halo mass relation, especially at the high mass end.

We evaluate our results using a test set, as well as the ability to reproduce important statistical relations, such as the black hole - halo mass relation and the galaxy stellar mass function.

4 Evaluation and Final Results

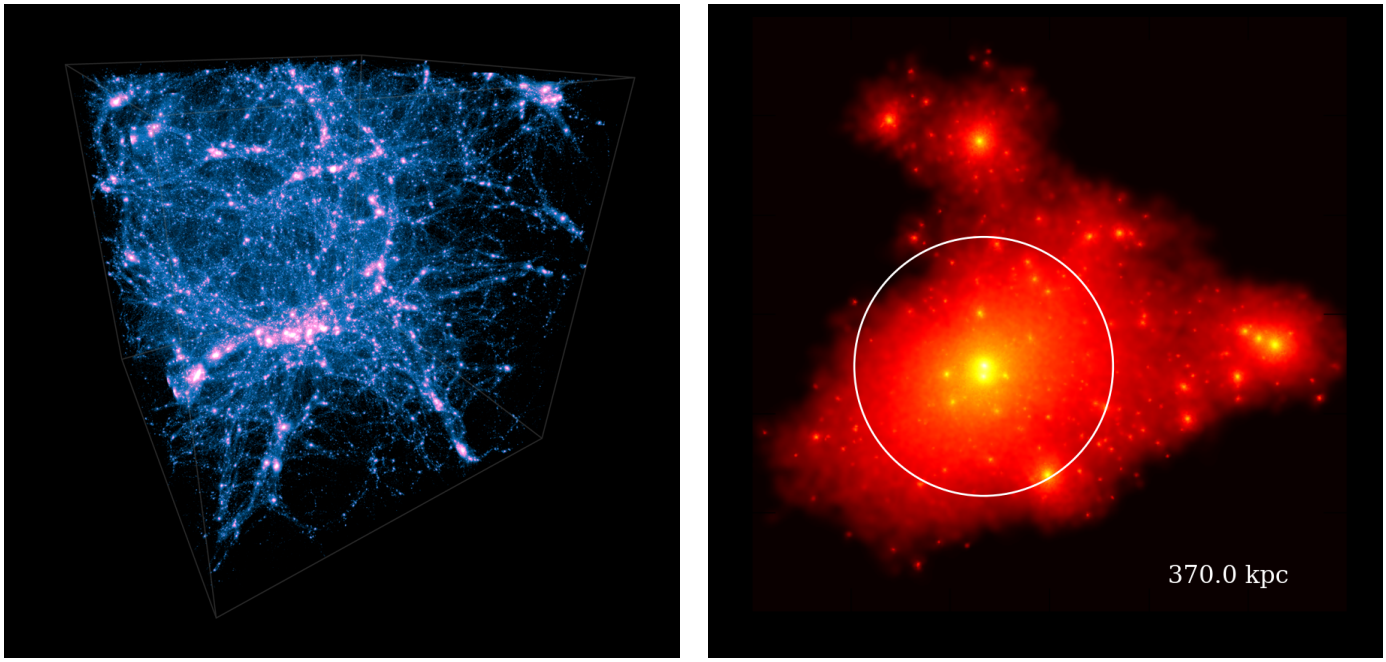
4.1 Visualization

I begin by showing visualizations of the simulations. Figure 1a shows a render of the dark matter density in the full Illustris-1 box at redshift $z = 0$ from an external perspective, obtained from the Illustris website[3]. This figure shows the large-scale distribution of dark matter in the box, where dark matter haloes appear as spots of high density (approximately 0-dimensional on large scales) and are connected by 1-dimensional filaments. Together, these constitute the cosmic web and reflects the time progression of gravitational collapse starting from the early universe: starting from 3-D distribution → 2-D objects (sheets) → 1-D objects (filaments) → 0-D objects (haloes).

This is project, we are not interested in the large-scale distribution of matter but the distribution of matter on the scale of a halo ($\lesssim$ 1 Mpc). Figure 1b shows the projected dark matter density, zoomed-in around a typical halo of size ($\approx 10^{13} M_\odot$). Here, I estimate the mass density with KDE, using the following cubic spline kernel:

$$W(x) = \begin{cases} \frac{1}{\pi h^3}\left[1 - \frac{3}{2}x^2\left(1 - \frac{x}{2}\right)\right], & \text{for } 0 \le x \le 1 \\ \frac{1}{\pi h^3}\frac{1}{4}\left(2 - x\right)^3, & \text{for } 1 < x \le 2 \\ 0, & \text{for } x > 2 \end{cases} \tag{1}$$

[3]Due to the large total number of dark matter particles $1820^3 \approx 6 \times 10^9$, substantial computational and memory resources are required to compute the densities and make such a visualization of the full box.



(a) Exterior view of the dark matter density distribution in the full Illustris-1 box at redshift zero, taken from the Illustris website. The 0-D objects (haloes) and 1-D objects (filaments) here constitute the cosmic web. Recall that each side of the box is 106.5 Mpc long.

(b) Projected dark matter density zoomed-in on a single halo, consisting of the central (most massive) subhalo and multiple smaller satellites. White circle shows the virial radius of this halo.

This cublic spline is frequently used for astrophysical simulations due to its finite support, thus a particle only contributes to a finite region determined by the smoothing length h. For simplicity in this visualization, I take $h = 1$, although h is typically determined for each particle as the distance to the Nth (e.g. $N = 32$) closest neighbor. The halo consists of the most massive subhalo, taken to be halo center, as well as multiple smaller satellite subhaloes, in the processes of being accreted and merging with the central subhalo. The white circle shows the virial radius R_{200} (see section 3.1 for the definition) of the halo, and gives an idea of the halo size.

4.2 Artificial neural network predictions

To show the performance of the ANN, Figure 2 shows the 2D histograms between the true vs predicted values for each of the six output variables on the test set. Solid blue lines show the fitted regression lines, with the R^2 value shown in the top left.

In the top row, we find the galaxy stellar mass, halo mass and black hole masses have R^2 values larger than 0.9, showing that with this set of input features, these three mass variables in Illustris can be predicted fairly well from using only the input (Illustris-Dark) features. In particular, the halo mass seems to be the easiest to predict, with R^2 very close to 1, and the histograms shows very little scatter from the diagonal. The stellar mass shows a scatter in the predicted values of approximately $\Delta \log(m_{\rm gal}) \pm 0.25$, fairly uniform across the stellar mass range. Interestingly, for the black hole mass, the scatter in the predicted values is smaller for more massive black holes.

In the bottom row, however, the predictions from the ANN are much less promising. For the star formation rate, the general trend appears to be captured. However, there is substantial scatter which results in a lower $R^2 = 0.7$. The model appears inadequate to predict the halo shape and circularity parameters of the Illustris haloes. For example, the predicted halo shapes have values which range only from 0.5 and 0.9, whereas the true values range between ≈ 0.3 and close to 1. Thus, the diversity in predicted shapes is much smaller compared to the true values, and the slope between the predicted and true values is smaller than one. A similar result applies for the circularity as well.

Note that parameters such as the stellar, halo and black hole mass are considered as more 'static' features, whereas the star formation rate, halo shape and circularities are associated with the dynamics of the halo and

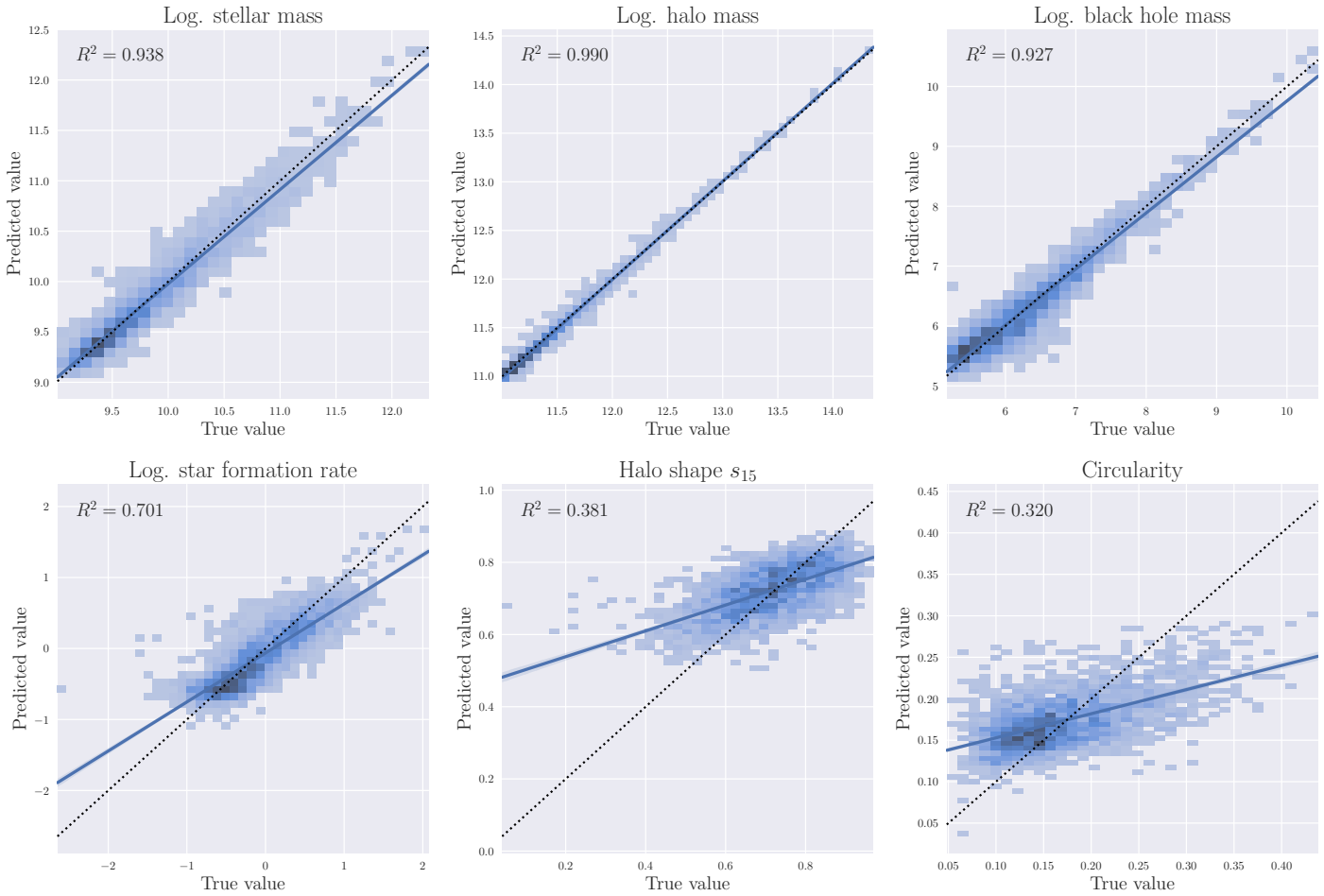


Figure 2: Results of the **artificial neural network model** described in section 3.4, showing the 2D histograms between the true vs predicted values for the six chosen output variables. Additionally, regression lines are shown as solid blue lines. The R^2 values are also shown in the top left corner. Whereas the three mass variables (on the top row) are well-predicted by the ANN, the model does not produce good predictions especially for the halo shape parameter s_{15} and the circularity.

its central galaxy. Thus, for the given N-body inputs, we conclude that the ANN model is able to predict the corresponding static features in the hydrodynamic run, but not the dynamical ones. This is unlikely to be a limitation of ANNs in general, but rather, that rather reflects that the predictors utilized here provide insufficient information to capture the full dynamics and other physics involved in hydrodynamics simulations.

4.3 Comparison with other models

In order to ensure that limitations do not lie solely in the ANN model, it is useful to compare against the predictions obtained using other regression models. In this section, I consider 1) a ridge regressor, and 2) a random forest model, using the implementations in the Python package SCIKIT-LEARN.

4.3.1 Ridge regression

It is interesting to see with the same input features, how well a linear model would work to predict the six outputs. Because some of the input features are correlated with one another, this can lead to high variance if linear regression without regularization is used. Here, I consider a Ridge regressor i.e. linear regression with $L2$-regularization. A separate model is trained for each of the six outputs, and 10-fold cross-validation is used to choose the best hyper-parameter λ for the regularization strength.

The resulting predictions are shown in the 2D histograms in Figure 3. In the top row, although the R^2

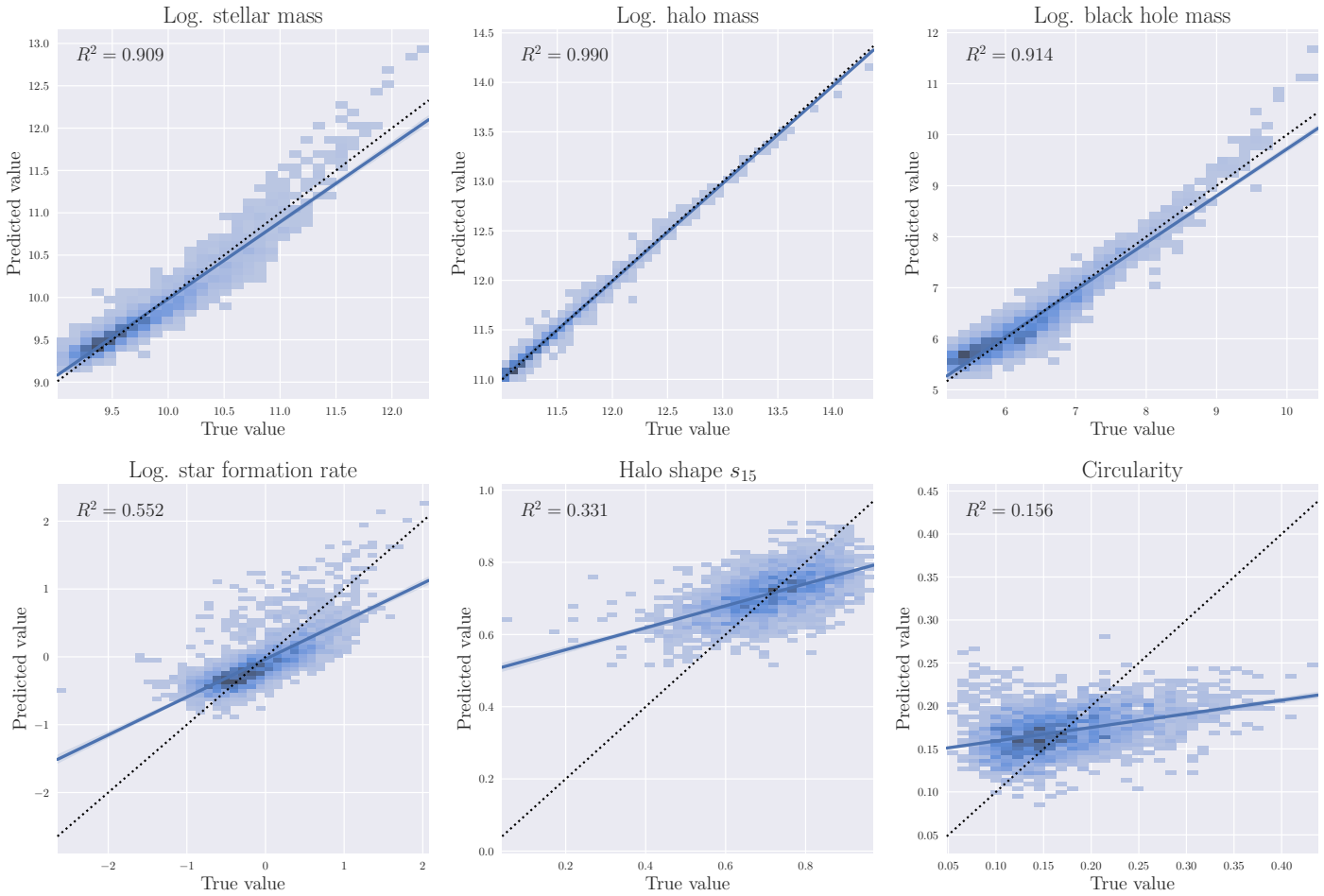


Figure 3: Results of using **ridge regression** for prediction, similar to Figure 2. Although the R^2 values of for the masses (first row) are high, we can see that a linear model fails to capture the high mass end well.

values for the stellar, halo and black hole mass are high ($R^2 > 0.9$), we can see from the large deviations that the model fails at the high-mass end. Since lower mass haloes are more numerous than high mass halos, the linear model achieves a low MSE by fitting primarily to the low-mass haloes. It is not hard to see that the residuals between the predicted and true values are mass dependant, signifying high bias. Unsurprisingly, the predictions for the lower row are also worse compared to the ANN model

We note that better results can probably be obtained by feature engineering (e.g. including higher powers of the input features), but that is outside the scope of this project.

4.3.2 Random forest

In this section, we consider a random forest model to perform the regression, which is an ensemble learning method. By combining bagging and randomized feature selection, a random forest aggregates the results of multiple weaker learners and has a reduced variance compared to a single decision tree. A separate random forest with 200 trees is trained for each of the six output variables.

Figure 4 shows the resulting 2D histograms of predicted vs true values using random forest models. Even with no hyper-parameter tuning, the results obtained here are very similar to that of the ANN model, both in terms of the histograms and the R^2 values. This means that the stellar, halo and black hole masses are predicted well, while the halo shape and circularity parameters are not. These results validate our hypothesis that the ANN architecture is unlikely to be the limiting factor hampering the performance on the halo shape and circularity.

One advantage of a random forest model is its interpretability compared to an ANN. In SCIKIT-LEARN, the fraction of trees that a feature contributes to is combined with the associated decrease in impurity to calculate

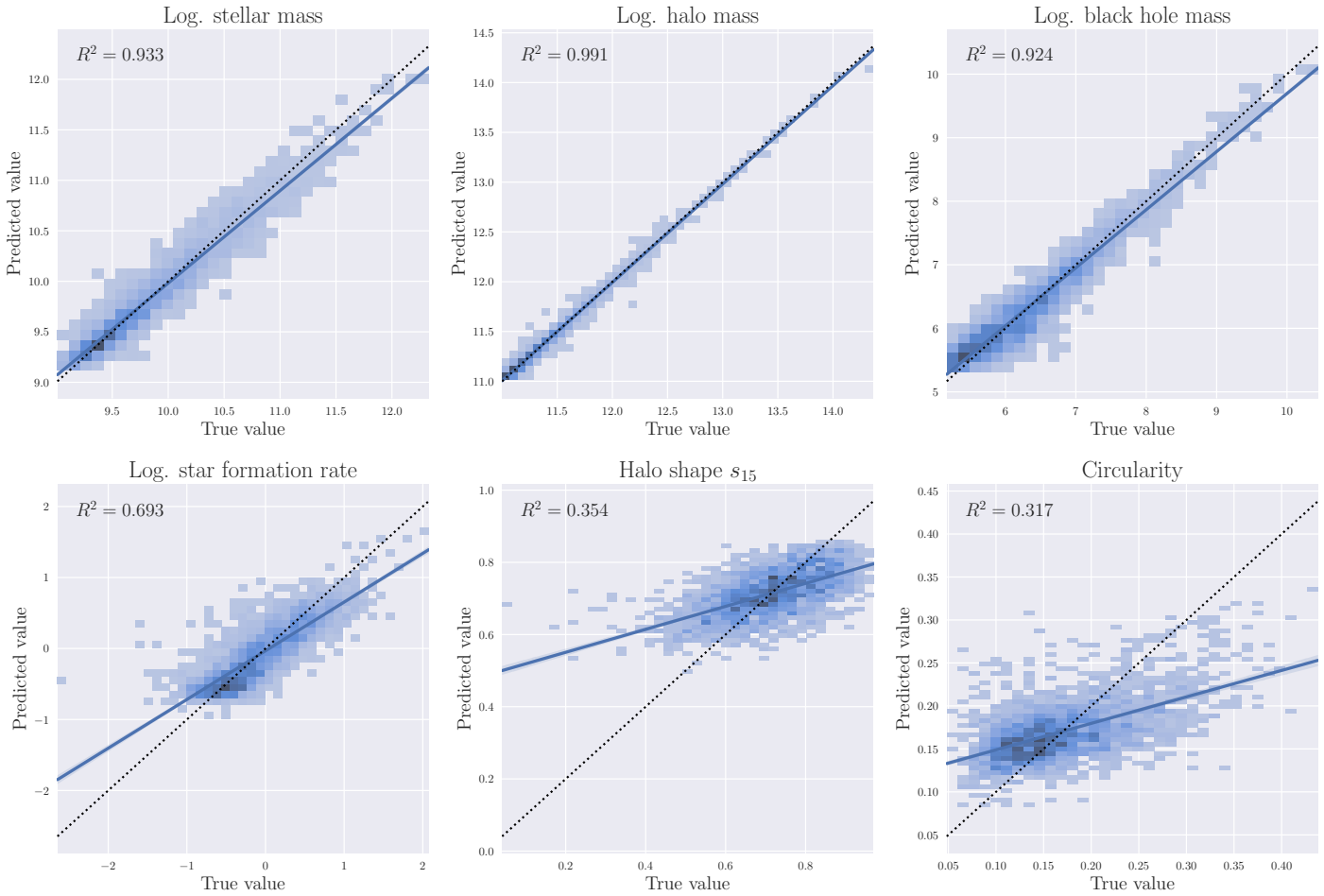


Figure 4: Results of using **random forest models** for prediction, similar to Figure 2. Note that a different random forest is trained for each output.

the relative importance of the features. Although the relative importance is evaluated on the training data and thus does not necessarily correspond to predictive features on the test data, we use it here to estimate the importance of the features. Figure 5 shows the relative importance of the top five features in predicting the six dependant variables. Except for the halo shape parameter s_{15}, the halo mass is by far the most important predictor for the other variables, and can be used almost exclusively to predict the stellar, halo and black hole masses. For the star formation rate, the halo concentration and the formation redshift are the next two most important features. Although the top predictor for the halo shape s_{15} is the corresponding parameter in the N-body counterpart, we recall from the previous sections that none of the machine learning models give good predictions on the test set.

4.4 Derived halo relations

Thus far, we have compared the predicted and true values of the output variables on a halo to halo basis. In galaxy formation, the ratio of galaxy mass to halo mass is extremely important because it reflects the galaxy formation efficiency, i.e. the fraction of primordial gas that the halo has managed to turn into stars. Through observational measurements, it is well-known that the galaxy formation efficiency is mass dependant due to the contrasting physical processes that dominate low-mass vs high-mass haloes. As such, this quantity is one of the main ones used to calibrate parameters and validate hydrodynamic simulations. Instead of predicting the galaxy mass ratio directly (i.e. including it as one of the output variables), I will use the ratio of the predicted masses as an additional diagnostic.

Another important relation is the black hole – halo mass relation, first identified through observations, which found a tight linear relation between the mass of the supermassive black hole and its host halo. This meant

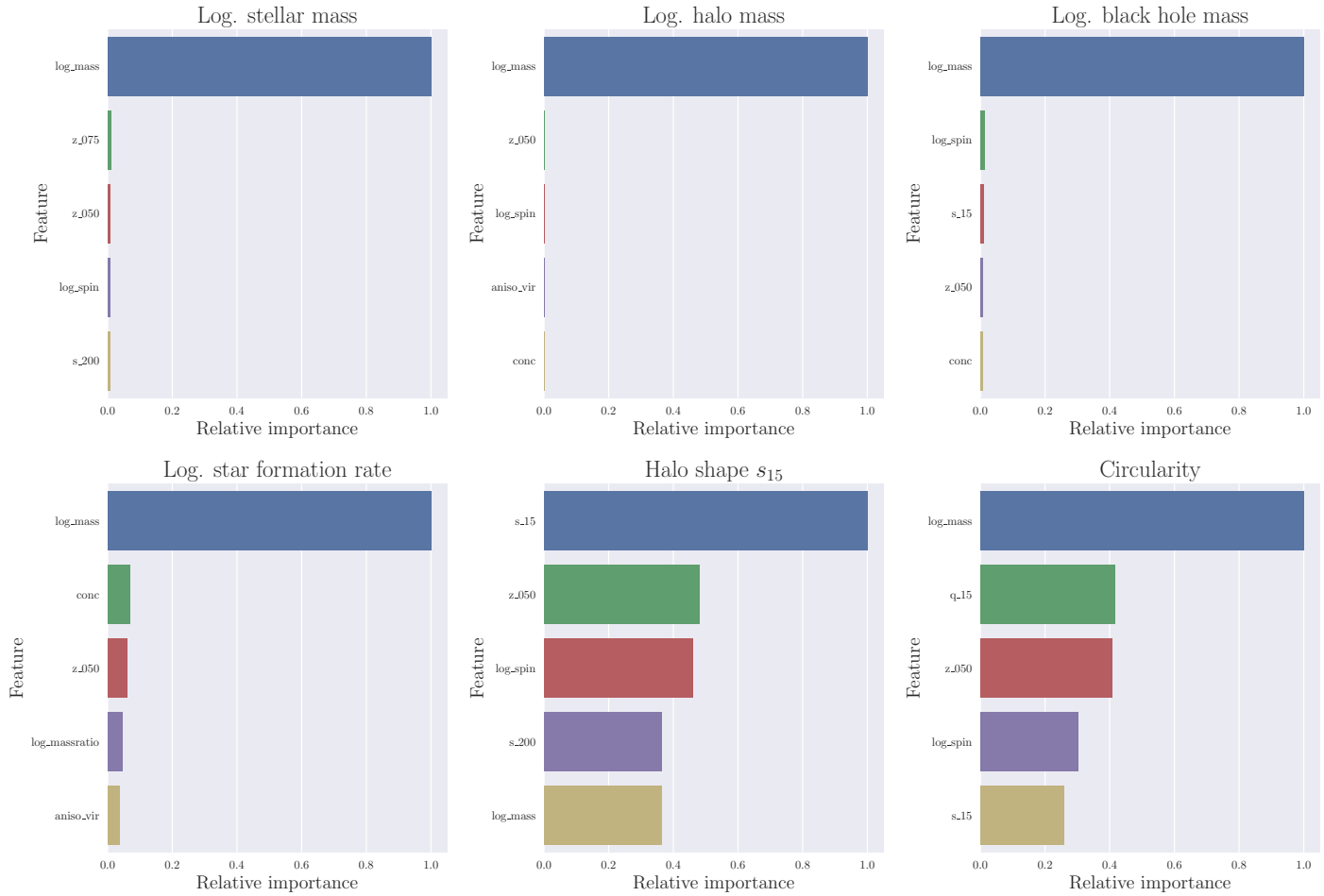


Figure 5: The relative importance of the top five predictors in predicting the different outputs, obtained from the random forest models.

that the growth of the supermassive black hole is closely linked to the growth of the halo, despite the fact that the supermassive black hole is found at the center and thus interacts only with a small region of the halo. As such, models of black hole physics used in simulations are informed and constrained by the black hole – halo mass relation.

In this section, we validate how well the predictions from the three models examined in the previous sections (ANN, random forest and ridge regression) are able to reproduce these two relations. Here, I am not interested in a halo to halo comparison, but rather if these relations can be reproduced on average using the masses predicted using the models.

The upper row of Figure 6 plots the galaxy mass ratio as a function of halo mass for the three models. Results using the predicted masses are shown in blue, with points showing individual haloes. The blue curve shows the median relation, obtained by segmenting the x-axis into 11 halo mass bins in logarithmic space. The blue shaded region shows the $25th$ to $75th$ percentile of the distribution in each mass bin. I compare these results to the median curve using the true masses of the test set. For the ANN (left-most panel) and the random forest (middle panel) models, the predicted masses capture the overall shape of the galaxy mass ratio, which peaks at approximately $10^{12}\mathrm{M}_{\odot}$. While the agreement is very good at lower masses ($< 10^{12}\mathrm{M}_{\odot}$), there is a larger discrepancy in the slope of this relation for more massive haloes in both models. Despite the high R^2 values in the Ridge regression model for the masses (as discussed previously for Figure 4.3.1), the linear model (right-most panel) fails completely at capturing the correct shape.

The lower row shows the black hole – halo mass relation using the true and predicted masses. Once again, the ANN and random forest models is able to reproduce this relationship. Since the curve is close to linear, the linear model performs well here, but still exhibits deviations at the high mass end.

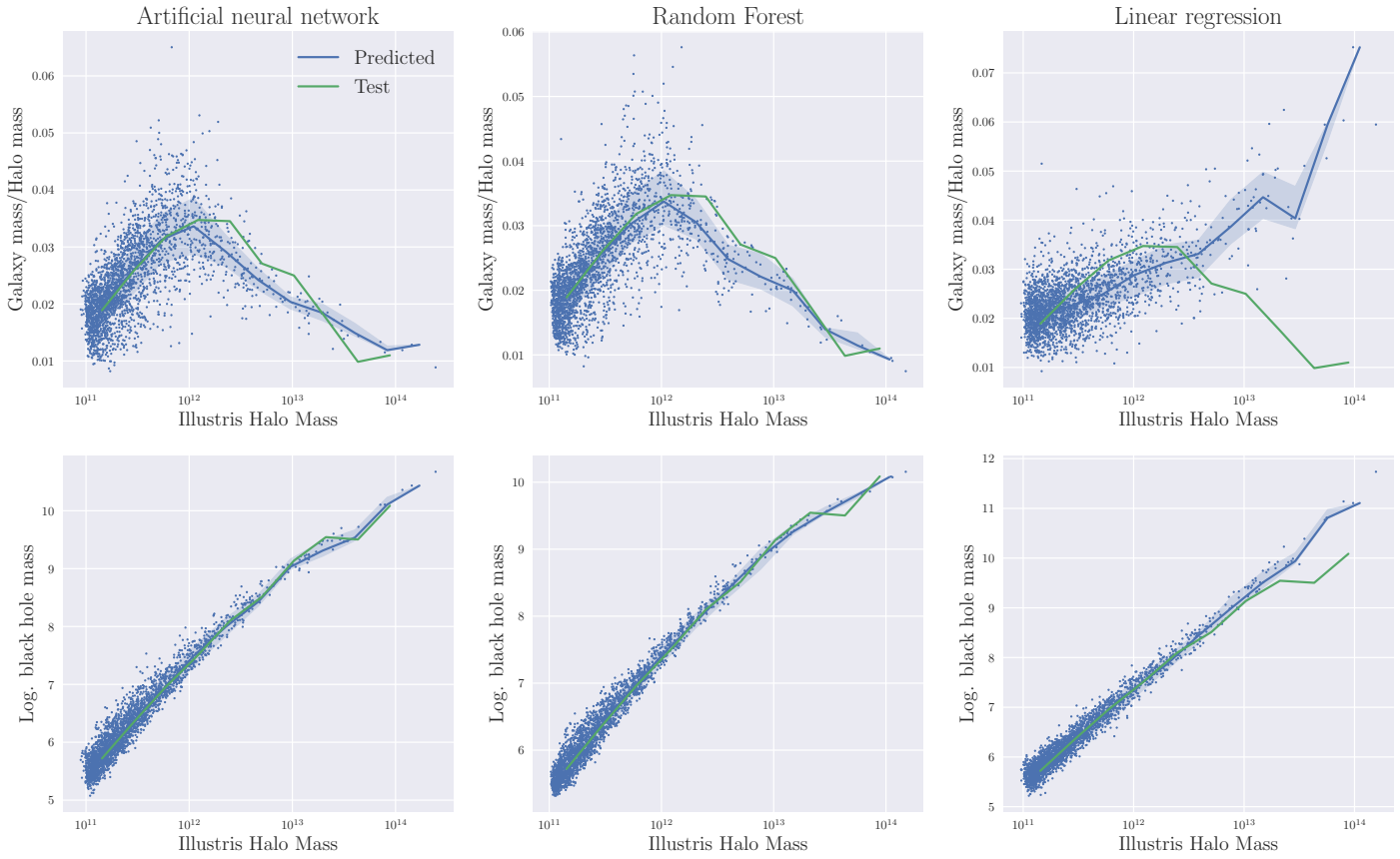


Figure 6

5 Discussion and conclusions

In this project, I set out to investigate how well an artificial neural network can predict galaxy and halo properties in the Illustris simulation, using only halo features in the N-body simulation Illustris-Dark. By matching the haloes on a one-to-one basis between the two simulations, this was a multiple-input, multiple output regression task. In total, there were 22 input features and 6 output variables, and the dataset consisted of 12728 haloes.

Evaluating the performance on a test set, I found that the model predicts the stellar, halo and black hole mass well. This includes capturing the shape of the galaxy mass fraction as a function of halo mass. For these variables, the results we obtained agree with those in other papers. For example, Kamdar *et al.* (2016) trained Extremely Randomized Trees (ERT) models using halo properties from (the hydrodynamic simulation) Illustris. Using a smaller and different set of parameters, the authors found that their ERT model was successful in modelling the results of galaxy formation. Note that this differs from the work in this project which uses properties from halo counterparts in the N-body simulation Illustris-Dark. Similar work was carried out by Jo and Kim (2019), who trained an ERT model using the newer simulation IllustrisTNG. They then applied the model to the much larger N-body simulation MultiDark-Planck to produce a galaxy catalogue. However, results from many numerical simulations (including Illustris and IllustrisTNG) have shown that galaxy formation physics have a back-reaction on the dark matter haloes, which causes their properties to change as well. In this project, this was evident from the halo shape parameter s_{15}, which was very different between halo counterparts in Illustris and Illustris-Dark. This implies that the ERT models trained solely on the hydrodynamics simulations might have limitations when applied to N-body simulations. My project eliminates such discrepancies by using the N-body features as inputs. Furthermore, these papers were restricted to the 'static' features (stellar, halo and black hole masses) and did not examine the effects on other properties such as the shape and the morphology (parametrized by the circularity).

For these two features, the model was unable to achieve good performance, which we attribute to their more

dynamical nature. This suggests that the input features used in ours models are insufficient to capture the results of complicated gas and other physics. For example, the halo and galaxy merger history can play a large role in determining the morphology of the merger remnant and thus also the halo shape. Better results could perhaps be obtained if the entire halo formation history was taken into account e.g. by using a recurrent neural network. Larger scale information about the density distribution could also be incorporated via convolutional neural networks (CNN), which have also been used to map the large-scale distribution of matter from Illustris-Dark to Illustris (Kasmanoff *et al.*, 2020).

In conclusion, machine learning and other deep learning methods has become more preponderant in the recent years. In this project, we have demonstrated the application of a simple artificial neural network and showed that it produced results similar to, or even superior to that from a random forest model. Although the question is open as to whether a set of features based solely on N-body simulations can serve as sufficient predictors, further research along this front will be useful in developing faster and less numerically intensive simulations.

References

D. N. Spergel, L. Verde, H. V. Peiris, E. Komatsu, M. R. Nolta, C. L. Bennett, M. Halpern, G. Hinshaw, N. Jarosik, A. Kogut, M. Limon, S. S. Meyer, L. Page, G. S. Tucker, J. L. Weiland, E. Wollack, and E. L. Wright, ApJS **148**, 175 (2003), arXiv:astro-ph/0302209 .

V. Springel, S. D. M. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J. A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg, and F. Pearce, **435**, 629 (2005), astro-ph/0504097 .

V. Springel, S. D. M. White, G. Tormen, and G. Kauffmann, MNRAS **328**, 726 (2001), arXiv:astro-ph/0012055 .

L. Gao, S. D. M. White, A. Jenkins, F. Stoehr, and V. Springel, MNRAS **355**, 819 (2004), arXiv:astro-ph/0404589 .

A. V. Macciò, B. Moore, J. Stadel, and J. Diemand, MNRAS **366**, 1529 (2006), arXiv:astro-ph/0506125 .

J. Diemand, M. Kuhlen, and P. Madau, **667**, 859 (2007), astro-ph/0703337 .

V. Springel, J. Wang, M. Vogelsberger, A. Ludlow, A. Jenkins, A. Helmi, J. F. Navarro, C. S. Frenk, and S. D. M. White, MNRAS **391**, 1685 (2008), arXiv:0809.0898 .

R. E. Angulo, C. G. Lacey, C. M. Baugh, and C. S. Frenk, MNRAS **399**, 983 (2009), arXiv:0810.2177 .

S. D. M. White and M. J. Rees, MNRAS **183**, 341 (1978).

R. S. Somerville and J. R. Primack, MNRAS **310**, 1087 (1999), arXiv:astro-ph/9802268 .

A. J. Benson, R. G. Bower, C. S. Frenk, C. G. Lacey, C. M. Baugh, and S. Cole, ApJ **599**, 38 (2003), arXiv:astro-ph/0302450 .

R. G. Bower, A. J. Benson, R. Malbon, J. C. Helly, C. S. Frenk, C. M. Baugh, S. Cole, and C. G. Lacey, **370**, 645 (2006), arXiv:astro-ph/0511338 .

C. Conroy and R. H. Wechsler, ApJ **696**, 620 (2009), arXiv:0805.3346 .

Q. Guo, S. White, C. Li, and M. Boylan-Kolchin, MNRAS **404**, 1111 (2010), arXiv:0909.4305 [astro-ph.CO] .

M. Vogelsberger, S. Genel, V. Springel, P. Torrey, D. Sijacki, D. Xu, G. Snyder, D. Nelson, and L. Hernquist, MNRAS **444**, 1518 (2014a), arXiv:1405.2921 .

R. A. Crain, J. Schaye, R. G. Bower, M. Furlong, M. Schaller, T. Theuns, C. Dalla Vecchia, C. S. Frenk, I. G. McCarthy, J. C. Helly, A. Jenkins, Y. M. Rosas-Guevara, S. D. M. White, and J. W. Trayford, MNRAS **450**, 1937 (2015), arXiv:1501.01311 .

A. Pillepich, V. Springel, D. Nelson, S. Genel, J. Naiman, R. Pakmor, L. Hernquist, P. Torrey, M. Vogelsberger, R. Weinberger, and F. Marinacci, MNRAS **473**, 4077 (2018), arXiv:1703.02970 .

M. Vogelsberger, S. Genel, V. Springel, P. Torrey, D. Sijacki, D. Xu, G. Snyder, S. Bird, D. Nelson, and L. Hernquist, **509**, 177 (2014b), arXiv:1405.1418 .

S. Genel, M. Vogelsberger, V. Springel, D. Sijacki, D. Nelson, G. Snyder, V. Rodriguez-Gomez, P. Torrey, and L. Hernquist, MNRAS **445**, 175 (2014), arXiv:1405.3749 .

D. Sijacki, M. Vogelsberger, S. Genel, V. Springel, P. Torrey, G. F. Snyder, D. Nelson, and L. Hernquist, MNRAS **452**, 575 (2015), arXiv:1408.6842 .

V. Springel, MNRAS **401**, 791 (2010), arXiv:0901.4107 [astro-ph.CO] .

M. Davis, G. Efstathiou, C. S. Frenk, and S. D. M. White, **292**, 371 (1985).

K. Dolag, S. Borgani, G. Murante, and V. Springel, MNRAS **399**, 497 (2009), arXiv:0808.3401 .

V. Rodriguez-Gomez, L. V. Sales, S. Genel, A. Pillepich, J. Zjupa, D. Nelson, B. Griffen, P. Torrey, G. F. Snyder, M. Vogelsberger, V. Springel, C.-P. Ma, and L. Hernquist, MNRAS **467**, 3083 (2017), arXiv:1609.09498 .

H. M. Kamdar, M. J. Turk, and R. J. Brunner, MNRAS **457**, 1162 (2016), arXiv:1510.07659 [astro-ph.GA] .

Y. Jo and J.-h. Kim, MNRAS **489**, 3565 (2019), arXiv:1908.09844 [astro-ph.GA] .

N. Kasmanoff, F. Villaescusa-Navarro, J. Tinker, and S. Ho, arXiv e-prints , arXiv:2012.00186 (2020), arXiv:2012.00186 [astro-ph.CO] .