SHAPing the Gas: Understanding Gas Shapes in Dark Matter Haloes with Interpretable Machine Learning

Luis Fernando Machado Poletti Valle,[1]★ Camille Avestruz,[2,3] David J. Barnes,[4] Arya Farahi,[5]
Erwin T. Lau[6], Daisuke Nagai[1]
[1]*Department of Physics, Yale University, New Haven, CT 06520, U.S.A.*
[2]*Department of Physics, University of Michigan, Ann Arbor, MI, 48109, U.S.A.*
[3]*Leinweber Center for Theoretical Physics, University of Michigan, Ann Arbor, MI, 48109, U.S.A.*
[4]*Department of Physics, Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.*
[5]*The Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[6]*Department of Physics, University of Miami, Coral Gables, FL 33124, U.S.A.*

30 November 2020

ABSTRACT

The non-spherical shapes of dark matter and gas distributions introduce systematic uncertainties that affect observable-mass relations and selection functions of galaxy groups and clusters. However, the triaxial gas distributions depend on the non-linear physical processes of halo formation histories and baryonic physics, which are challenging to model accurately. In this study we explore a machine learning approach for modelling the dependence of gas shapes on dark matter and baryonic properties. With data from the Illustris-TNG hydrodynamical cosmological simulations, we develop a machine learning pipeline that applies XGBoost, an implementation of gradient boosted decision trees, to predict radial profiles of gas shapes from halo properties. We show that XGBoost models can accurately predict gas shape profiles in dark matter haloes. We also explore model interpretability with SHAP, a method that identifies the most predictive properties at different halo radii. We find that baryonic properties best predict gas shapes in halo cores, whereas dark matter shapes are the main predictors in the halo outskirts. This work demonstrates the power of interpretable machine learning in modelling observable properties of dark matter haloes in the era of multi-wavelength cosmological surveys.

Key words: cosmology: theory – dark matter – large-scale structure of Universe – galaxies: clusters: general – galaxies: groups: general – methods: numerical

1 INTRODUCTION

Upcoming large-scale surveys of galaxy clusters and groups in X-ray, microwave, optical wavelengths (such as eROSITA, CMB-S4, and the Rubin Observatory) will cover large cosmological volumes and provide accurate maps of the gas and dark matter distributions in galaxy clusters and groups, which in turn enable us to improve cosmological and astrophysical constraints (see Allen et al. 2011; Kravtsov & Borgani 2012; Pratt et al. 2019, for reviews). Baryonic physics dominate systematics in several branches of cluster-based cosmological constraints, such as using the Sunyaev-Zeldovich (SZ) effect measurements of pressure and density profiles (e.g., Amodeo et al. 2020) and cross-correlations of the thermal SZ and gravitational lensing signals (e.g., Hill & Spergel 2014; Van Waerbeke et al. 2014; Hojjati et al. 2017; Osato et al. 2018a, 2020). To maximize the scientific returns of these surveys, we must understand and model the non-linear physics of baryons in dark matter haloes of groups and clusters (see Wechsler & Tinker 2018, for review).

One of the common simplifying assumptions in modelling baryons in dark matter haloes is the spherical symmetry in the halo gas distribution. If not properly modeled, it will introduce systematic uncertainties that significantly affect observables and selection function of upcoming surveys. Specifically, the spherical assumption does not properly capture the scatter in observable-mass relations of galaxy clusters (Buote & Humphrey 2012; Chen et al. 2019; Ansarifard et al. 2020), introducing systematic uncertainties in model predictions such as those of multiwavelength cross-correlation of clusters and groups, which are impacted by orientation bias (e.g., Osato et al. 2018b).

Cosmological simulations have shown that dark matter shapes in haloes can be described as triaxial (e.g. Jing & Suto 2002; Allgood et al. 2006), which are dependent on halo formation history (e.g., Despali et al. 2014; Lau et al. 2021) and baryonic physics (Kazantzidis et al. 2004; Debattista et al. 2008; Bryan et al. 2013; Suto et al. 2017; Chua et al. 2019). Gas in massive haloes also follow a triaxial distribution that are dependent on the gravitational potential of the dark matter and the baryonic effects (e.g., Lau et al. 2011; Biffi et al. 2013). The non-linear relationship between gas shapes and a combination of multiple halo and baryonic properties makes it particularly suitable for an interpretable machine learning approach,

★ E-mail: luisfernando.machado@aya.yale.edu which is essential for computationally efficient models that also inform physical intuition about the properties in the simulated haloes.

This paper is structured as follows. §2 describes the Illustris-TNG cosmological hydrodynamical simulations used in this study. §3 motivates and describes our choice in machine learning approaches with the XGBoost model (Chen & Guestrin 2016) and the SHAP (SHapley Additive exPlanations) method (Shapley 1953; Lundberg & Lee 2017). §4 presents XGBoost model predictions of the radial profiles of gas shape in dark matter haloes based on different input subsets of halo and baryonic properties, the interpretation of our trained models with SHAP that identifies the most predictive halo properties, and a comparison of model performance with conditional abundance matching, an empirical non-machine learning based approach. Finally, §5 summarizes our main findings.

2 SIMULATION

2.1 Hydrodynamical Simulations

The goal of this paper is to determine how well different halo properties can predict the gas shape profile of galaxy groups and clusters, where predictions are made using a machine learning algorithm trained on simulated data (see §3 for details on the algorithm). We use data from the TNG300 simulation data from the IllustrisTNG project (Naiman et al. 2018; Pillepich et al. 2018; Nelson et al. 2018; Marinacci et al. 2018; Springel et al. 2018). For a comparison with lower mass haloes, we also use data from the TNG100 simulation. TNG100 and TNG300 are cosmological simulations created with the Arepo code (Springel 2010) in a periodic box with side lengths 75 Mpc (TNG100) and 205 Mpc (TNG300), and target masses $m_{\rm baryon} = 1.4 \times 10^6 M_\odot$ (TNG100) and $m_{\rm baryon} = 1.1 \times 10^7 M_\odot$ (TNG300). TNG100 and TNG300 were run with a flat cosmology consistent with Planck Collaboration et al. (2016), with $h = 0.6774$, $\Omega_\Lambda = 0.6911$, $\Omega_m = 0.3089$, $\Omega_b = 0.0486$, $\sigma_8 = 0.8159$, and $n_s = 0.9667$. IllustrisTNG has been shown to yield realistic clusters (e.g., Barnes et al. 2018; Vogelsberger et al. 2018). For more details regarding the Illustris-TNG simulations, we refer the reader to Pillepich et al. (2018).

2.2 Halo Catalog Contents

For the model training, we use the halo catalog information provided in the publicly available TNG300 datasets. Figure 1 shows the range of halo masses covered in the population, which includes a total of 2548 haloes. For the TNG100 population used in §4.3.1, we include haloes within the mass range $10^{12} M_\odot < M_{500c} < 10^{13} M_\odot$. We include common data on the haloes' redshifts, radii (R_{200c}, R_{500c}, $R_{\rm vir}$)[1] and masses (M_{200c} and M_{500c})[2]. We also include the density profiles for gas ($\rho_{\rm gas}(r)$), stars ($\rho_{\rm stars}(r)$), and dark matter ($\rho_{\rm dm}(r)$) within each halo. In addition to mass properties, we compute the localized triaxial shape parameters at different radial annuli, corresponding to the minor-to-major (S) and middle-to-major (Q) axis ratios, as described in §2.2.1. Furthermore, we include halo formation proxies such as the halo concentration $c_{\rm vir}$, the peak-centroid offset $X_{\rm off}$, and the dark matter surface pressure $P_{\rm dm}$ (measured within the spherical shell between $[0.8, 1.0] \times R_{\rm vir}$), which are described in §2.2.2. Finally, we include information about the halo accretion history via the mass accretion rate parameter Γ, as described in §2.2.3.

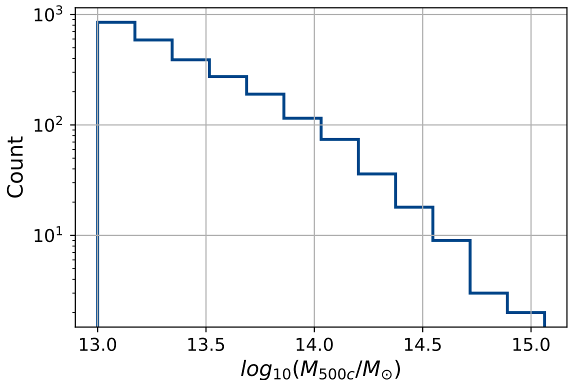

Figure 1. Distribution of M_{500c} for the halo population used in this study in the TNG300 dataset. We focus on haloes with $M_{500c} \geq 10^{13} M_\odot$, and note that the methodology in this paper can also be used to analyze wide halo mass ranges.

The halo data points used in our analysis are summarized in Table 1. Note that we use all information from the snapshot at $z = 0$, except for the mass accretion history, which naturally requires information from earlier snapshots.

2.2.1 *Shape Profiles*

We consider gas and dark matter distributions as triaxial ellipsoids with axis lengths $a \geq b \geq c$, and express the shape of such distributions in terms of the axis ratios $S \equiv c/a$ (minor-to-major) and $Q \equiv b/a$ (middle-to-major). Note that we always have $S \leq Q$, and that larger values of S and Q imply more spherical distributions. Note also that we can determine shape parameters for different particle types, which gives rise to gas ($S_{\rm gas}$, $Q_{\rm gas}$) and dark matter ($S_{\rm dm}$, $Q_{\rm dm}$) shape profiles.

The axis ratios are derived from the mass tensor:

$$\mathcal{M}_{ij} \equiv \frac{1}{\sum_p^N m_p} \sum_p^N m_p x_i x_j, \tag{1}$$

which is computed by starting with DM particles or gas cells with mass m_p within a given radial shell $[r, r + dr]$, where x_i is the position of the DM particle or gas cell in the direction $i = 1, 2, 3$ relative to the halo centre. The major, intermediate, and minor axes (a, b, c) of the triaxial ellipsoid are then assigned as the square roots of the sorted eigenvalues of the mass tensor. In the iteration steps, every particle or gas cell with ellipsoidal (or elliptical) radius $r_{\rm ep} = a\sqrt{(x'/a)^2 + (y'/b)^2 + (z'/c)^2} \leq R_{500c}$ (where x', y', z' are the coordinates of the particle in the frame of eigenvectors) is included in the computation of the mass tensor. The iteration repeats until the axis ratios converge. To ensure that the shape measurements are not affected by subhaloes, we remove particles in subhaloes when we compute the mass tensor.

Note that there is no unique way of estimating halo shape. There are other methods in the literature for computing the mass tensor. Some uses all particles interior to the radius instead of within a shell (e.g., Bryan et al. 2013). Some methods use a $r_{\rm ep}^{-2}$ weighting scheme (e.g., Dubinski & Carlberg 1991; Allgood et al. 2006; Schneider et al. 2012) to reduce effects of substructures; others use the spherical radius $r_{\rm sph} = a\sqrt{x'^2 + y'^2 + z'^2}$ instead when selecting the particles,

[1] R_{200c} (resp. R_{500c}) corresponds to the radius of a sphere with density equal to 200 (resp. 500) times the critical density of the Universe at the halo's redshift. $R_{\rm vir}$ refers to the definition from Bryan & Norman (1998).

[2] M_{200c} (resp. M_{500c}) is the halo mass contained within R_{200c} (resp. R_{500c}).

Group	List of Parameters
Global properties	M_{200c}; M_{500c}; R_{200c}; R_{500c}
Density Profiles	$\rho_{\rm gas}(r)$; $\rho_{\rm stars}(r)$; $\rho_{\rm dm}(r)$
Gas and Dark Matter Shape Profiles	$S_{\rm gas}(r)$; $Q_{\rm gas}(r)$; $S_{\rm dm}(r)$; $Q_{\rm dm}(r)$
Formation History Proxies	$c_{\rm vir}$; $X_{\rm off}$; $P_{\rm dm}$
Accretion History Parameters	$a\left(\frac{M}{M(z=0)}\right)$, $\frac{M}{M(z=0)} \in \{0.5, 0.7\}$; $\Gamma_{200c}(a)$, $a \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$

Table 1. Halo information used in the machine learning model. Shape profiles are obtained through the iterative inertia tensor procedure described in §2.2.1. Formation history proxies are described in §2.2.2. Mass accretion rate parameters are determined from tracing data, according to §2.2.3. Note that these profiles are interpolated into 25 radial bins, to maintain significant granularity while not overwhelming the learning process.

often without iterations (e.g., Cole & Lacey 1996; Bailin & Steinmetz 2005). Other methods calculate the shape of density contours by selecting particles based on their local density values (e.g., Jing & Suto 2002). Compared to these other shape measurement methods, our method has been shown to produce accurate halo shape as function of radius (Zemp et al. 2011).

2.2.2 *Halo Formation Proxies*

The halo formation history significantly impacts the gas contents within haloes, and can also alter the gas distribution shapes encoded by $S_{\rm gas}$. As a result, we include commonly used halo formation proxies as inputs in the machine learning model, to assess whether the model can find correlation between halo formation history and the final observed gas shapes in haloes. We include the following halo formation proxies and morphological metrics:

- $X_{\rm off}$ - Distance offset between the peak total mass density and the centre of mass for the total mass distribution. Large values of $X_{\rm off}$ indicate high merger activities and a less relaxed halo environment.
- $c_{\rm vir}$ - Defined by $c_{\rm vir} \equiv R_{\rm vir}/R_s$, where R_s is the scale radius found by fitting the DM density profile to the Navarro, Frenk and White (NFW) profile (Navarro et al. 1996). High halo concentration correlates with earlier formation epochs, which in turn indicate that a halo has had more time to relax.
- $P_{\rm dm}$ - Dark matter surface pressure, computed by including all particles within the spherical shell between $[0.8, 1.0]R_{\rm vir}$, as defined in Eq. 4 from Shaw et al. (2006).

2.2.3 *Mass Accretion Rate*

To compute the mass accretion rate, we use the merger trees computed with Sublink (Rodriguez-Gomez et al. 2015). The Sublink algorithm identifies the progenitor sequences across snapshots by identifying haloes that share the most bound particles between consecutive snapshots. This results in mass accretion history for every halo in the population. From the mass accretion history $M(z)$ we derive Γ via the following:

$$\Gamma(t) = \frac{\log_{10}(M(t)/(M(t - t_{\rm dyn}))}{\log_{10}(a(t))/(a(t - t_{\rm dyn}))}, \tag{2}$$

with $M = M_{200c}$. The dynamical time $t_{\rm dyn}$ is given by:

$$t_{\rm dyn}(z) = 2^{3/2} \cdot t_{\rm H}(z) \cdot \left(\frac{\rho_\Delta(z)}{\rho_c(z)}\right)^{-1/2} = 2^{3/2} \cdot t_{\rm H}(z) \cdot \Delta^{-1/2}, \tag{3}$$

which for $\Delta = 200c$ gives $t_{\rm dyn}(z) \approx t_{\rm H}(z)/5$.

We note that this analysis could be performed to obtain values for Γ at any selected radius. We choose to focus on Γ_{200c} since R_{200c} contains the outer regions of haloes, and therefore variations in M_{200c} reflect more accurately the broader merger activity of the halo. We also include the expansion factor at varied halo mass fractions relative to the present-day mass, $a\,(M/M(z=0))$, $M/M(z=0) \in \{0.5, 0.7\}$. Note that both Γ_{200c} and $a\,(M/M(z=0))$ encode instantaneous rates of mass accretion, but cover different timescales and provide a more complete picture of the mass evolution in the sampled haloes.

3 INTERPRETABLE MACHINE LEARNING MODEL

Our goal is twofold: making (a) prediction and (b) interpretation. We employ a machine learning model, XGBoost, to predict gas shape based on different halo properties. We then determine how the model made its predictions, by employing the SHAP method to quantify the relative importance of each feature in the final model prediction. We describe the motivation and pipeline for using XGBoost and SHAP in the following subsections.

3.1 Regression with Machine Learning

In this work, the primary goal is to develop a method to predict a continuous value (gas shape) based on the value of one or several predictor values (halo properties), which is a straightforward *regression analysis*. Among many machine learning models to perform such regression analyses, we chose XGBoost, a scalable model that has been vetted in multiple applications within the machine learning community.

XGBoost is an implementation of gradient boosted decision trees, a set of machine learning techniques used for regression and classification (Chen & Guestrin 2016). We note that XGBoost performs significantly faster and more accurately than similar gradient-boosted tree solutions. Additionally, it presents theoretically-backed design choices in its underlying optimization algorithms (see Natekin & Knoll 2013).

The crux of this paper is to accurately predict the gas shape profile for a galaxy group and cluster based on a given set of halo parameters. We train the XGBoost regressor model using properties from a sample population of galaxy groups and clusters. This population is divided into two random samples, training and validation, to allow for a proper evaluation of the resulting trained models. We also select different groups of halo properties to serve as input training features. The output target values are comprised of the gas shape (e.g., the ratio of the semi-major axis to the semi-minor axis, $S_{\rm gas}$) measured at a given halo-centric radius for each galaxy group or cluster in the training set. Note, we train a separate XGBoost model for each radius at which we want a shape prediction. We also note that the target data consists of values ranging between 0 and 1, and thus we must be careful in verifying that final model predictions satisfy this constraint. The trained models enable both a prediction and a ranking of relative feature importance. To assess the model performance, we evaluate each model on another subset of our galaxy groups and clusters (the *validation set*), predicting the gas shape from properties of galaxy groups or clusters that were not included in the training process. By applying SHAP, we can also determine which features have the most impact in the final shape prediction at any given radius.

3.2 Interpretability via SHAP Method

While machine learning models are powerful in making accurate prediction, explaining how the model reached a specific prediction is a difficult task. Our goal is not just to make the most accurate prediction, but also to understand the patterns picked up by our machine learning model in order to inform our physical intuition of the dependencies between gas shapes and halo properties.

We employ the SHAP (SHapley Additive exPlanations) method in order to explain the trained models. SHAP follows a game-theoretic approach to explain the model predictions by computing the relative contributions of each feature to the final predictions (Shapley 1953). Specifically, SHAP values represent the changes in the expected model prediction when conditioning on that feature (Lundberg & Lee 2017). SHAP's strong theoretical guarantees regarding the *accuracy* and *consistency* of the estimated feature contribution distinguishes this tool from other interpretability methods. We employ SHAP method in conjunction with the trained XGBoost model to interpret the trained models.

3.3 Data Pre-Processing and Rescaling

The study of galaxy cluster properties naturally brings about a challenge with the scales involved in the different observable data fields. In fact, the set of feature inputs presented in Table 1 contains values ranging from unit scales (e.g., $S_{\rm dm}$) to several orders of magnitude larger (e.g., M_{200c}). Machine learning models, in particular XGBoost, are highly sensitive to features at different scales, and tend to erroneously bias towards using the larger feature values as decision points, even if the underlying data does not correlate as strongly with the target field. To prevent this issue, and obtain meaningful impact from every feature in the dataset, we first perform a common preprocessing of the input data. This involves applying a logarithm to all fields in the input dataset, and then shifting each feature's distribution to one ranging from 0 to 1. As a result, every data field maintains their original sorting properties, while also allowing XGBoost to focus on the meaningful correlation between features.

Machine learning models also suffer from the risk of *overfitting*, whereby a model learns highly non-linear trends in the input parameter space, and fits its underlying model to predict the training target data very accurately, at the expense of accuracy when presented with new data. A common approach to prevent overfitting is the use of a training-test data split, in which a subsample from the entire available dataset is used for the training process, and the remaining data is saved for a validation step once the training is complete. This process allows us to verify that the model performs well with both the training and the test datasets, which indicates the model has not suffered from overfitting. In this study, we apply the commonly used 80-20 split, with 80% of the halo population is used for training, and the remaining 20% is used in the validation step.

3.4 Hyperparameter optimization

Machine learning models often have parameters that control the learning process, and such *hyperparameters* are not related to the actual feature data used during model training. These hyperpameters are fundamental in creating generalizable and accurate models, while preventing overfitting of the training data. As a first step in training such models, it is essential to perform hyperparameter optimization, the process through which we select optimal learning parameters before proceeding with the actual model training. Each hyperparameter controls a separate configuration in the learning process, with the most commonly noted as the main drivers of accuracy improvements being:

- Learning Rate: Controls the rate at which new trees are added to the model, in order to account for new information. Lower values lead to a slower addition of new trees, which slows down the learning process and prevents model overfitting.
- Number of Estimators: Number of boosted trees fitted via the training process. Larger values lead to more accurate models, at the risk of overfitting to the training data.

We perform a traditional grid search, via which every combination of the considered hyperparameters is used to train a different model on the same sample feature set. The models are compared via a mean squared error loss function (MSE), and we determine the optimal set of hyperparameters as the set that minimizes the MSE on the sample training set. If we consider a machine learning model trained on n haloes, each with target data (here, gas shape axis ratio obtained from the simulation) Y_i and predicted value $\hat{Y}_i$ (here, gas shape axis ratio predicted from the ML model), then the MSE value for this model is computed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \tag{4}$$

We also determine the relative improvement on the final MSE values gained from each hyperparameter. We performed this grid search at 3 different radii, to compare optimal hyperparameters in the inner, intermediate, and outer regions of the haloes. We note that the optimal hyperparameter values have a weak dependence on the radial bin selected, with the final improvements in MSE varying less than 5% across all sampled radial bins. Therefore, we select the most common set as the optimal set of hyperparameters throughout our analysis.

The set of considered hyperparameters, ranked by largest relative improvement in the final MSE, is described in Table 2. We note that the average MSE obtained with the optimal hyperparameters showed slight improvements (by ∼ 20%) over similar models trained with randomly selected hyperparameters.

3.5 Interpretability via Feature Ranking

We train XGBoost models using different subsets (described in Table 3) of features as the training data sets. We compare the model accuracy for each feature set as a function of radius, to determine which set of features contribute the most in determining the gas shape distributions in our halo population. In addition to the comparison across feature sets, we explore the relative importance of different features within a given feature set in order to determine which features contribute the most to the final models. We intend to quantify the predictive power of different halo properties on baryonic shape parameters. We then employ SHAP method to rank features based on

XGBoost Hyperparameter	Explored Values	Optimal Value
Learning Rate	$10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10$	10^{-1}
Number of Estimators	10, 50, 100, 300, 500, 700, 1000	300
Maximum Depth	3, 4, 5, 6, 7	5
Minimum Child Weight	3, 4, 5, 6	4
Gamma	0, 0.1, 0.2, 0.3	0
Subsample	0.5, 0.6, 0.7, 0.8, 0.9	0.8
Column Sample Rate By Tree	0.4, 0.5, 0.6, 0.7, 0.8	0.6
Alpha	$10^{-9}, 10^{-7}, 10^{-5}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 100$	10^{-2}

Table 2. List of hyperparameters searched when preparing the model, ranked by overall impact in the final model accuracy. The individual hyperparameters are described in §3.4. We note that *learning rate* and *number of estimators* provide over 90% of the combined improvement in model accuracy. Furthermore, the set of optimal hyperparameters only varied minimally across radial bins, and as a result we apply the same optimal set across all models, to maintain the consistency and limit the impact to the different feature sets under consideration.

their overall contribution to the predictions. Specifically, we employ `TreeSHAP` algorithm (Lundberg et al. 2020), which has been developed for estimating SHAP values using tree-based machine learning models such as XGBoost model. Features with large absolute average SHAP values contribute more to the prediction and hence more important.

The computed SHAP values for every trained model allow us to determine the most important features for every iteration of the XGBoost models, which in turn inform our physical intuition of the dependencies between gas shapes and several halo properties at varying radial ranges, from the core to the outer regions of groups and clusters.

We compared the relative feature importance determined from the XGBoost built-in feature ranking methods. We then followed previous comparisons between different implementations of `TreeSHAP` algorithm and converged on applying the "interventional" method of `TreeSHAP`, ensuring that it yields consistent results when compared to the best feature importance ranking methods from XGBoost (Chen et al. 2020).

4 RESULTS

4.1 Relationship between Shape Distribution and Halo Properties

Before applying XGBoost models to study the sampled haloes, we first analyze how the distribution of gas shapes varies with different halo properties. As an example, the panels of Figure 2 show the probability distribution function (PDF) of the gas shape axis ratio, $S \equiv c/a$, measured at three different radii. We compute the PDFs of gas shape parameters measured at inner ($0.13 < r/R_{200c} < 0.15$, *left*), intermediate ($0.45 < r/R_{200c} < 0.55$, *middle*), and outer ($2.0 < r/R_{200c} < 2.4$, *right*) radial bins, with individual distributions corresponding to groups/clusters selected by different halo properties. We select these sample radial bins based on the premise that different physical processes may impact the gas at different cluster-centric radii. We predict that baryonic effects such as cooling and energy feedback impact the gas in the core regions, and expect accretion to affect gas in the outskirts. The top panel shows gas shape axis ratio distributions of objects selected by halo concentration, $c_{\rm vir}$, and the bottom panel shows distributions selected by the localized dark matter shape axis ratio, $S_{\rm dm}$.

We first note that the PDF of $S_{\rm gas}$ values shift to smaller values towards larger radii, regardless of sample selection. This result generally agrees with previous studies (Samsing et al. 2012; Lau et al. 2012), and indicates that gas distributions are more spherical, on average, in the outskirts than they are in the inner regions.

Next, we see that the distribution of the gas shape in the core depends more heavily on halo concentration than it does on dark matter shapes. One possible explanation is that halo concentration is affected by baryonic processes (for instance, see Schaller et al. 2015), leading to a degeneracy between concentration and gas shape values.

In addition, the PDFs seem to depend more strongly on the dark matter shape parameter as we move towards outer radii. In the outer regions, gas particles trace more closely the dark matter potential, with more spherical gas distributions corresponding to more spherical dark matter distributions. The dark matter shapes do not heavily impact the PDFs of shape axis ratios in the inner radii.

While focusing on specific halo properties and select radial bins is informative, we note that it can be impractical to extend this analysis to all halo properties and all discretized radial bins. In addition, it is challenging to quantify this relative ranking of predictive power between halo properties at different radii. We use XGBoost to incorporate information from any number of halo properties in a single model to predict the gas shape axis ratio for any given radius. The next sections present the predictive performance of XGBoost models (see §4.2) and how we can leverage interpretable machine learning models to determine the most predictive halo properties at every radial bin (see §4.3).

4.2 Impact of feature subsets on model performance

In this subsection, we describe the performance of the XGBoost models trained with different subsets of halo properties. We selected 7 different sets of halo properties (*feature sets*), which are described in Table 3. We train one XGBoost model for each feature set and for each of the 25 radial bins. The localized gas shape parameter $S_{\rm gas}$ at a given radial bin is the target data for each model. We train a total of $7 \times 25 = 175$ regressor models.

For each trained model, we compute the MSE between the predicted values and the actual (target) values. We perform this computation for the reserved test sets, in order to prevent overfitting and to measure the final accuracy of our model for different radial bins and different feature set inputs.

Figure 4 illustrates the MSE of all models, normalized to the score obtained from the *Everything* model at each given radius. Note, the inclusion of more input features often, but not always, improves the final MSE results. This is not true towards small radii, where adding dark matter shape information reduces the accuracy of the resulting model. From Figure 2 we see that gas shapes weakly correlate with

Feature Set Name	Included Features
Everything	Gas density, DM density, DM Shapes, Accretion History
Everything - DM Shapes	Gas density, DM density, Accretion History
Everything - DM Shapes - Accretion History	Gas density, DM density
DM	DM density, DM Shapes, Accretion History
DM - DM Shapes	DM density, Accretion History
DM - Accretion History	DM density, DM Shapes
DM - DM Shapes - Accretion History	DM density

Table 3. Feature sets included in the XGBoost model training. These feature sets are mentioned in the labels for Fig. 4. Note that all models also include *Global Properties* and *Formation History Proxies*, as detailed in Table 1.

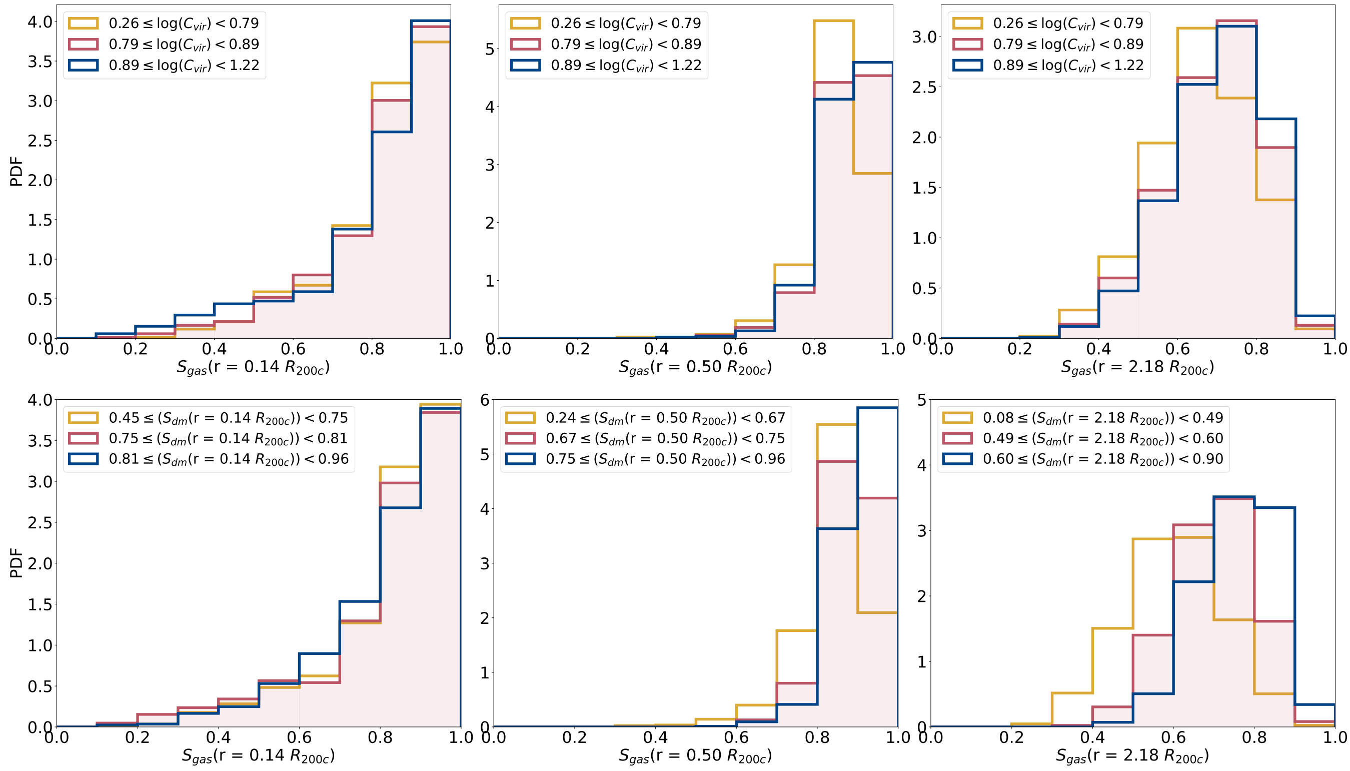

Figure 2. Distributions of gas shape parameters at three different radial bins: inner ($0.13 < r/R_{200c} < 0.15$, *left*), intermediate ($0.45 < r/R_{200c} < 0.55$, *middle*), and outer ($2.0 < r/R_{200c} < 2.4$, *right*). The halo populations are binned based on different halo parameters: halo concentration ($c_{\rm vir}$, *top*) and localized dark matter shape parameter ($S_{\rm dm}$, *bottom*). We note that gas shapes are generally more spherical (i.e. larger values of $S_{\rm gas}$) towards the inner radii. We also note that, at all radii, gas shapes vary weakly with halo concentration. On the other hand, gas shapes vary more strongly with dark matter shapes towards outer radial bins. This analysis indicates that the correlation between gas shapes and different halo parameters varies with radial bin, suggesting different relative impacts of physical effects.

dark matter shapes in the halo cores, and adding uncorrelated data to a machine learning model may lead to a reduction in model accuracy, as observed towards smaller radii in Figure 4 (for more details on the impact of additional data on machine learning models, see Nakkiran 2019; Advani & Saxe 2017).

We examine the performance of XGBoost in different radial regions of our sample haloes. In the inner regions ($r/R_{200c} \leq 0.2$), the predictive power of XGBoost largely improves with the inclusion of baryonic features, most notably gas density. In fact, the MSEs for the test dataset are mostly identical among the DM-only feature sets, and improve greatly for the feature sets that include gas information. This is to be expected, since gas distributions at the inner regions are largely affected by baryonic effects such as feedback, and cannot be accurately determined solely from information about the dark matter distribution. We discuss this effect further in §4.3.

In the intermediate regions ($0.2 \leq r/R_{200c} \leq 0.8$), the performance gap between different XGBoost models starts to shrink. The models including only dark matter training data present closer accuracy to the models including baryonic properties, and this trend continues towards larger radii. This shift suggests that the baryonic data is less necessary for predictions of gas shapes towards halo outskirts. In fact, at $r = 0.5R_{200c}$ the relative RMSEs of the dark matter-only models are only 10% away from the RMSE obtained with all of the available input data.

This trend becomes even more clear towards the outer halo radii (beyond R_{200c}). The relative accuracy of the *DM Density + DM*

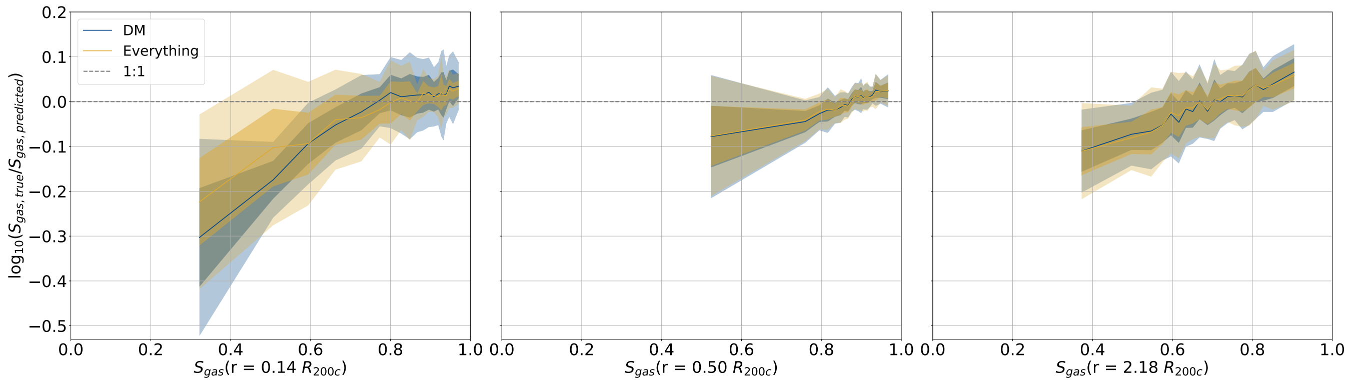

Figure 3. Predictive performance of the XGBoost models at different radial bins, namely inner ($0.13 < r/R_{200c} < 0.15$, *left*), intermediate ($0.45 < r/R_{200c} < 0.55$, *middle*), and outer ($2.0 < r/R_{200c} < 2.4$, *right*). The solid line indicates the median value of $\log_{10}(S_{\rm gas,\,true}/S_{\rm gas,\,predicted})$, for both the *DM* model and the *Everything* model (as described in Table 3), in bins of $S_{\rm gas}$. The horizontal gray line indicates the 1:1 line, which would happen for a perfectly accurate model. The shaded regions indicate the log-normal scatter (1σ and 2σ) around the median line. Both the *DM* and *Everything* models are more accurate towards outer radial bins. In addition, the models differ most in the inner radial bin, and become more similar towards outer radii.

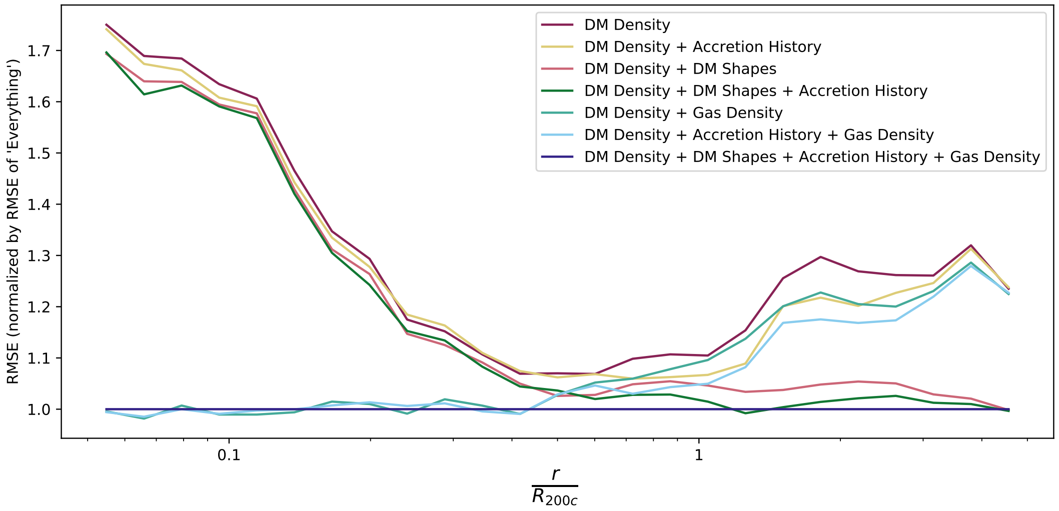

Figure 4. Model accuracy (measured by the root mean squared error, RMSE; lower RMSEs indicate more accurate models) vs. radius (in units of R_{200c}) for 7 different feature sets. We trained a XGBoost regressor for every one of the 25 radial bins available, for every one of the 7 feature subsets. All models predict $S_{\rm gas}$, the localized gas short-to-major axis ratio at the different radial bins. The feature sets are described in detail in §4.2. The RMSE values are normalized by the values obtained with all features (labeled *Everything* in Table 3), to indicate relative accuracy. The models with only DM input data perform best towards the cluster outskirts, since the gas distributions in the outskirts are largely determined by the underlying gravitational potential. On the other hand, accuracy worsens towards the inner regions, as the gas distributions are largely affected by baryonic effects (e.g., feedback and cooling), and cannot be accurately determined from dark matter information alone. Furthermore, the DM-only models present similar accuracy levels from the inner regions up to around R_{200c}, at which point the models that include dark matter shape perform considerably better. This improvement indicates that baryonic properties are the main predictors of gas shape towards the inner regions, while dark matter shape is an important predictor of gas shapes towards the cluster outskirts. Finally, accretion history parameters have a minor impact in model accuracy, as the dark matter shape profiles largely account for the predictive power of the DM-only models.

Shapes and *DM Density* + *DM Shapes* + *Accretion History* models continues to improve towards larger radii, suggesting that dark matter properties are sufficient in predicting gas shapes in the halo outskirts. However, models without dark matter shapes (e.g., *DM Density* + *Gas Density*) perform relatively worse in these regions. Combined, these results indicate that dark matter shapes are a powerful and necessary predictor of gas shapes beyond $0.5R_{200c}$.

Physically, this observation is consistent with outer regions of haloes being largely determined by their dark matter distributions, and are not as strongly affected by baryonic effects (Nagai et al. 2007; Chan et al. 2015; Schaller et al. 2015).

We summarize the trends of different XGBoost models at different radial bins in Figure 3. We compare the accuracy of *Everything* and *DM* models at three different radial bins: $r/R_{200c} \in [0.13, 0.15], [0.45, 0.55], [2.0, 2.4]$. Both models improve in accuracy and become more similar to one another towards larger radii. This agrees with the results from Figure 4, where we show the model accuracy as a function of radius, demonstrating that dark matter information is the most predictive halo data set outside the halo core $r/R_{200c} > 0.5$. Adding baryonic information does not significantly increase the model accuracy outside the core. On the other hand, in the halo core, the model generally under-predicts $S_{\rm gas}$ at true

$S_{\rm gas} \lesssim 0.7$, and continues to under-predict $S_{\rm gas}$ at smaller true $S_{\rm gas}$ values. Adding baryonic information improves the accuracy of the prediction around true $S_{\rm gas} \sim 0.3$, as the fractional difference between model and true $S_{\rm gas}$ values decreases from $\sim 50\%$ to $\sim 30\%$. At higher true $S_{\rm gas}$ values, however, adding baryonic information does not significantly improve the prediction accuracy.

4.3 Most Predictive Halo Properties

We now focus on the relative feature importance rankings obtained from XGBoost at every radial bin and for every feature subset. Table 4 presents a summary of the findings. The feature importance rankings indicate the most predictive halo properties at every radial bin, which in turn provides insights into physical processes that drive gas shapes at different scales.

For any given XGBoost model, SHAP determines the relative feature importances in predicting the target gas shape values. The feature importance rankings are obtained from the `TreeSHAP` algorithm described in §3.5. Figure 5 shows example `TreeSHAP` results for the DM feature set at two different radii. Note that the summary results in Table 4 are obtained from a collection of such figures. For instance, in the inner radial bin, $\rho_{\rm gas}$ is the most predictive halo property among the DM features, closely followed by global mass properties and dark matter shape parameters. On the other hand, in the outer radial bin, dark matter shapes and densities are the most predictive features.

The `TreeSHAP` algorithm allows for an even deeper analysis of the relative feature importances. In particular, note that in the cluster outskirts (around $2R_{200c}$), across all dark matter features, the dark matter short-to-major axis ratio $S_{\rm dm}$ is the most powerful predictor of $S_{\rm gas}$. Furthermore, the impact of the middle-to-major axis ratio $Q_{\rm dm}$ is considerably smaller than the impact of $S_{\rm dm}$ in the final $S_{\rm gas}$ values. This is correlated with the presence of accreting filaments in the outer regions of galaxy clusters, which are mainly defined by one preferential alignment direction. Further studies on the shape of filaments and the effect of substructure on the final gas and dark matter shapes would shed light on this effect.

Among all halo properties, localized gas densities $\rho_{\rm gas}$ are the most predictive for localized gas shapes. In fact, the gas density information is closely tied to the distribution of the gas contents in haloes, and therefore is naturally the main predictor of gas shape profiles across all radial ranges. In addition, dark matter shape is a significant predictor of gas shapes, particularly towards the halo outskirts. This correlation is to be expected, since the gas contents follow the underlying gravitational potential, which in turn is largely determined by the dark matter distribution. The main findings come from the feature rankings in the DM-only feature sets.

Towards the halo cores, the localized gas density $\rho_{\rm gas}$ is the main predictor of gas shapes, whereas towards the outer regions, the underlying localized dark matter shapes become the most relevant features. In addition, there is a significant contribution from the dark matter density profiles across all radial ranges. This discrepancy in feature importance rankings at different radial scales connects back to the greater impact of baryonic physics in the cores of galaxy clusters, which decreases towards the outskirts and gives place to the increasing effect of the dark matter gravitational potential further from the halo cores.

4.3.1 Predictive properties in low mass haloes

In this study we mainly focus on modelling gas shapes in massive haloes from the TNG300 simulations. However, the same machine learning pipeline can be applied to analyze halo properties in any simulation, and the results can shed light on differences in physical models between simulations. For one instance of this type of comparison, we apply the XGBoost + SHAP pipeline to gas shapes in TNG100 haloes. We focus on low mass galaxy groups ($10^{12} - 10^{13} M_\odot$) in the TNG100 simulation, and expect different SHAP feature rankings corresponding to different relative impact of baryonic effects vs. gravitational effects in the gas shapes in these lower mass objects. We apply the triaxial model to the TNG100 sampled haloes, and note that the machine learning pipeline can help verify how accurate this assumption is in this scenario.

Figure 6 shows the SHAP feature importance rankings from a XGBoost model trained on TNG100 haloes. Unlike the rankings from TNG300, the inner radial bin ranking from TNG100 finds halo concentration to be the most predictive dark matter feature for gas shapes at halo cores. Since TNG100 contains more lower mass haloes than TNG300, this difference suggests mass dependence in the relative importance of halo properties in determining gas distributions. Since gas distributions are affected by halo formation history and baryonic processes, determining the most predictive features can illuminate the relative impact of different physical processes on the halo formation and evolution. Hence, this pipeline can inform our physical understanding of different implementations of baryonic effects across different simulations.

4.3.2 Limitations of SHAP

We acknowledge a limitation regarding the use of SHAP in this study, which may be expanded upon in future works. We note that SHAP cannot always be used to determine *causal relationships* (Ma & Tourani 2020). When applying this exercise to future simulations, we need to carefully distinguish predictive power from physical effect. For instance, while we find that dark matter shapes are more predictive of gas shapes in the halo outskirts, we cannot infer which physical processes lead to this correlation from SHAP alone. The combination of XGBoost and SHAP helps in predicting values of gas shapes based on halo properties, and may lead to future scientific questions, but cannot answer these questions without a deeper analysis of the underlying physics.

4.4 Comparisons with Conditional Abundance Matching

Finally, we compare the XGBoost performance with an analytical method of modelling correlated halo properties. We use Conditional Abundance Matching (CAM, Masaki et al. 2013; Hearin & Watson 2013) to generate another set of predicted gas shapes based on a few halo properties. Specifically, we utilize the `conditional_abunmatch` method from halotools (Hearin et al. 2017). In general, CAM provides an ansatz for the dependence of a given halo observable on a primary and a secondary halo properties. In this study, we use the halo mass M_{500c} and the global dark matter shape $S_{\rm dm}$ as the respective primary and secondary halo properties. By design, CAM matches the cumulative probability density distribution functions (CDF) of $S_{\rm gas}$ and of $S_{\rm dm}$ in a given M_{500c} bin,

$$F(S_{\rm gas}|M_{500c}) = F(S_{\rm dm}|M_{500c}), \tag{5}$$

where $F(X|Y) = \int_{-\infty}^{X} P(x|Y)dx$ is the CDF. CAM matches the rank-ordered lists of $S_{\rm gas}$ and $S_{\rm dm}$. Note that this method assumes a

Radial Range (R_{200c})	*Everything* model	*DM* model	*DM minus DM Shape* model
0.05 – 0.2	Halo Mass, $\rho_{\rm gas}$	Halo Mass, $S_{\rm dm}$	Halo Mass, Γ
0.2 – 0.8	$\rho_{\rm gas}$, DM Shape	DM Shape, $\rho_{\rm dm}$	$\rho_{\rm dm}$, Γ
0.8 – 2	$X_{\rm off}$, DM Shape	$X_{\rm off}$, DM Shape	$X_{\rm off}$, Γ
2 – 3	DM Shape, $P_{\rm dm}$	DM Shape, $P_{\rm dm}$	$P_{\rm dm}$, Γ

Table 4. Most important features for the XGBoost models at different radial ranges. The feature importances are measured for two feature set models, *Everything*, *DM* and *DM minus DM Shapes*. The *Everything* model gains most of its predictive power from the gas density profiles, which explains the better accuracy towards inner radii, where astrophysical effects are more pronounced. The *DM* model relies on halo mass information to predict gas shape in the inner radii, but gives more weight to DM shape and DM density for the outer regions. In the intermediate regions, there is also contribution from the mass accretion rate information, though it plays a secondary role in the predictions.

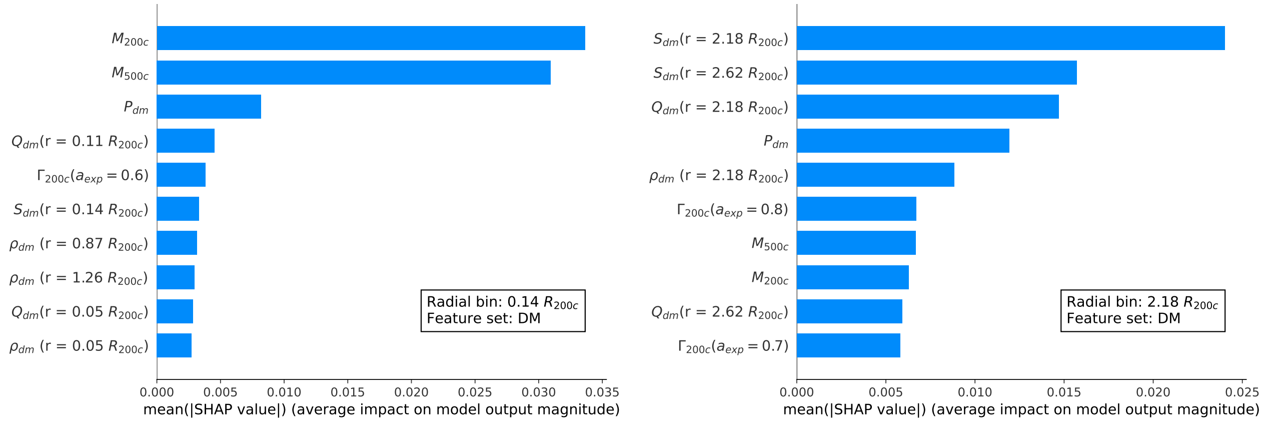

Figure 5. Example feature importance rankings determined by SHAP for two radial bins, *left:* inner ($r = 0.14R_{200c}$), and *right:* outer ($r = 2.18R_{200c}$) in. TNG300 data. In both models, we use the dark matter (DM) only feature set to train the XGBoost models and predict gas shape $S_{\rm gas}$ at the corresponding radial bins. Each figure shows the 10 most predictive features from the selected feature set, ranked by the SHAP algorithm described in §4.3. In particular, global halo masses and localized DM shapes are the strongest DM-only predictor of gas shapes in inner bins, while $S_{\rm dm}$ is the strongest predictor of $S_{\rm gas}$ in the outer regions of haloes.

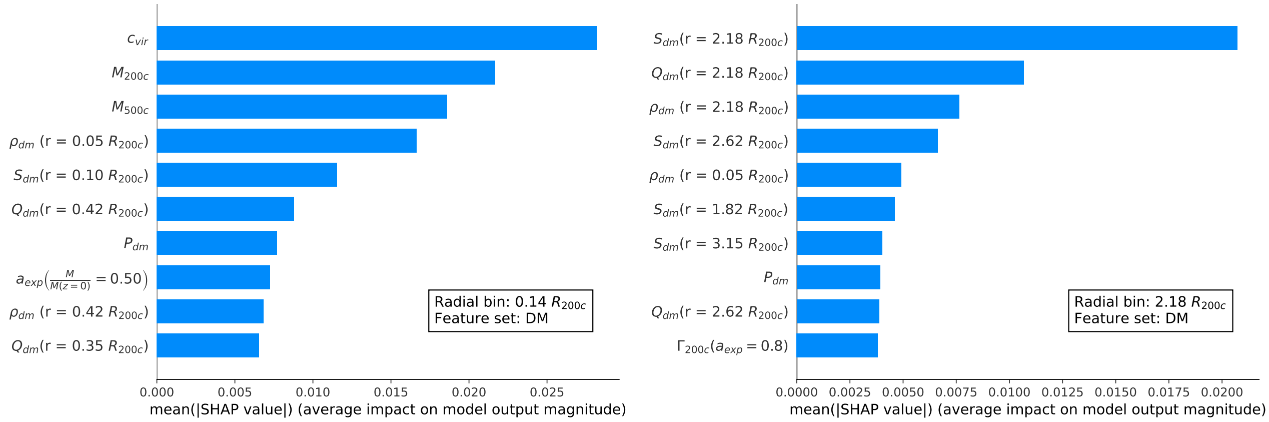

Figure 6. Similar to Figure 5, but using TNG100 haloes. Compared to Figure 5, the inner radius model finds halo concentration $c_{\rm vir}$ to be the most predictive halo property, suggesting different impacts of baryonic physics in the evolution of core gas shapes in the lower mass haloes.

perfect correlation between the target $S_{\rm gas}$ and the secondary parameter $S_{\rm dm}$ at any given M_{500c}, which is generally not true. To account for this assumption, we introduce stochasticity between $S_{\rm gas}$ and $S_{\rm dm}$ by re-shuffling the rank-ordered list of $S_{\rm gas}$ values. The reshuffling happens based on the Spearman correlation coefficient between $S_{\rm gas}$ and $S_{\rm dm}$ computed from the simulation. This procedure results in a relation between $S_{\rm gas}$ and $S_{\rm dm}$ that closely resembles what we measure in the simulation. Thus, the CAM model is by nature *empirical*, in that it entirely depends on the input simulation (or input observation) on which it is calibrated.

The CAM model describes the dependence of $S_{\rm gas}$ on two halo parameters. This is different from the XGBoost model, wherein we in general consider a wider set of halo parameters. To achieve a fair comparison between XGBoost and CAM, we use two different

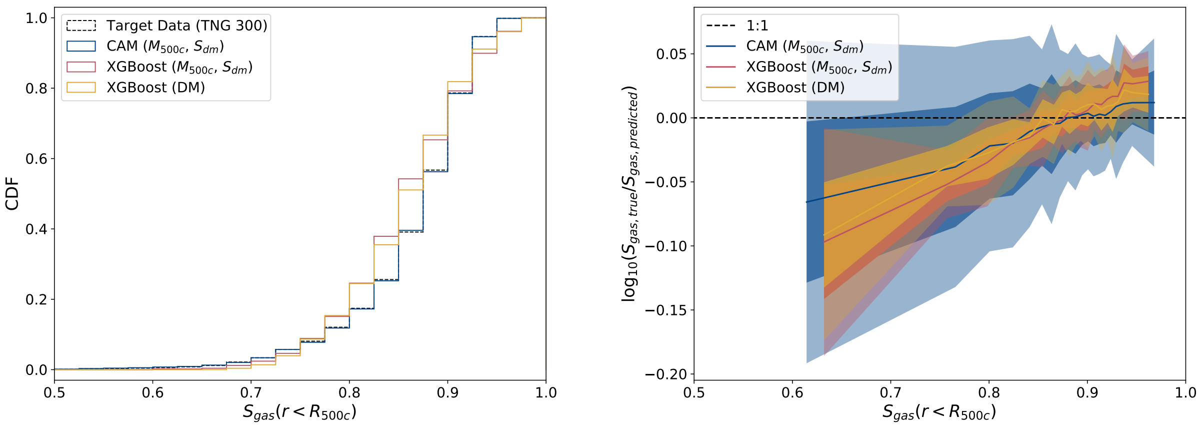

Figure 7. Trade-offs between XGBoost and Conditional Abundance Matching (CAM) models. The target gas shape parameter, $S_{\rm gas}(r < R_{500c})$, includes all gas particles within R_{500c} for each halo. We compare three models to the target data: (1) a CAM model obtained by using M_{500c} and global $S_{\rm dm}$ as primary and secondary parameters, respectively; (2) a XGBoost model trained with only M_{500c} and $S_{\rm dm}$ as inputs; and (3) a XGBoost model trained with the DM feature set (described in Table 3), which includes all dark matter features. *Left:* we compare the cumulative distributions (CDFs) obtained from each model against the target distribution of gas shape parameters. CAM is primarily designed to reproduce the target distribution more closely than XGBoost, whereas XGBoost is not as capable of reproducing the extrema of the target data range. *Right:* we compare the median and scatter of the predicted values at fixed bins of target gas shape $S_{\rm gas}$, similarly to Figure 3. The horizontal gray line indicates the ideal model. The solid lines indicate the median predicted values for each model, and the shaded regions indicate the log-normal scatter (1σ and 2σ) around the median line. XGBoost outperforms CAM in both the median values and the scatter ranges, even when only trained using the same exact parameter inputs, confirming that XGBoost manages to learn more non-linear correlations from training data. While CAM is designed to match given distributions of data, it is not able to predict new values of a given property on a halo-by-halo basis as accurately as a machine learning method such as XGBoost.

Model	KS Statistic	KS p-value	Spearman Correlation	RMSE
CAM (M_{500c}, $c_{\rm vir}$) (not plotted)	8×10^{-3}	0.95	0.3	0.09
CAM (M_{500c}, $S_{\rm dm}$)	6×10^{-3}	0.95	0.5	0.07
XGBoost (M_{500c}, $S_{\rm dm}$)	0.14	10^{-7}	0.5	0.05
XGBoost (DM)	0.12	5×10^{-6}	0.6	0.04

Table 5. Quantifying trade-offs between XGBoost and CAM from Figure 7. The two-sample Kolmogorov-Smirnov (KS) statistic and p-value between the target and model CDFs from *left* panel and Spearman correlation coefficient and RMSE of the true and predicted gas shape values from the *right* panel. The KS values illustrate that CAM best reproduces the CDF, as it is designed to do. The RMSE illustrates that XGBoost best captures the variance on a halo-by-halo basis.

feature sets: the *DM* feature set, as described in Table 3; and a set including only M_{500c} and $S_{\rm dm}$, for a true apples-to-apples comparison. Finally, we choose to target global gas shape parameters, which we compute by including all gas cells within R_{500c}. We choose this target data to compare the predictive power of CAM and XGBoost on a quantity that summarizes the gas distribution of the entire halo, instead of the shape measurement at specific radial bins.

We first quantify the differences in the distributions. We perform the two-sample Kolmogorov-Smirnov (KS) test for two scenarios: (1) true TNG300 measurements compared to XGBoost predictions, and (2) true TNG300 measurements compared to the CAM predictions, and display the results in Table 5. The two-sample KS statistic tests whether two samples have the same distribution by comparing the maximum difference D in their cumulative density distribution functions (CDF), which we show in the left panel of Figure 7. A smaller value of D and a larger p-value indicate that the two samples are more likely to be drawn from the same distribution. For CAM, the KS statistic is $D = 8 \times 10^{-3}$ (with $p = 0.95$). For XGBoost, $D = 0.14$ (with $p = 10^{-7}$) for the dataset with M_{500} and $S_{\rm dm}$ and $D = 0.12$ (with $p = 5 \times 10^{-6}$) for the full *DM* dataset. The KS statistic for CAM is smaller than those for XGBoost. This illustrates that the distribution from the CAM predictions better matches the distribution of the true TNG300 measurements, the target data of the models.

Second, we compare the accuracy of gas shape predictions between CAM and XGBoost. In the right panel of Figure 7 we show the ratio $S_{\rm gas,true}/S_{\rm gas,predicted}$ as a function of $S_{\rm gas,true}$, between CAM and XGBoost. This figure shows that XGBoost outperforms CAM in model accuracy. The scatter in XGBoost ($\sim 10\%$) is significantly smaller than that of CAM ($\sim 30\%$), indicating that XGBoost performs better than CAM in predicting $S_{\rm gas}$. At $S_{\rm gas,true} = 0.4$, the XGBoost values are $\sim 40\%$ more accurate than those from CAM.

These results suggest that there are important trade-offs between the two methodologies. CAM can reproduce the overall target distribution, but is unable to use information from more than two features, and it is also less accurate in its predictions. XGBoost provides more accurate models, but cannot fully reproduce the overall target CDF, and has trouble recreating the target outliers (for instance, see Figure 3 from Ntampaka et al. 2019). XGBoost also presents several advantages: it predicts target values more accurately for individual haloes; it can learn from as many halo properties as needed; and it manages to account for more non-linear correlations between different features and the target data. Furthermore, SHAP allows for interpretability of the resulting XGBoost models, and provides feature rankings that can motivate and inform the understanding of halo formation and baryonic effects.

5 CONCLUSIONS

In this work, we present interpretable machine learning models that predict gas shapes in dark matter haloes. We train predictor models with XGBoost, an implementation of gradient boosted trees (Chen & Guestrin 2016), with different subsets of input halo properties from the Illustris-TNG simulations. We use `TreeSHAP` implementation of SHAP (Shapley 1953; Lundberg & Lee 2017; Lundberg et al. 2020), a game-theory-based method to quantify the relative predictive power of different halo properties in the XGBoost models that predict gas shape at several halo radii. The main findings are:

- With dark matter information only, the XGBoost models predict gas shapes with a mean error of $\lesssim 20\%$ around the true $S_{\rm gas}$ values at radii $r/R_{200c} \geq 0.5$ in haloes with $M_{500c} \geq 10^{13} M_{\odot}$. The model accuracy of dark matter only predictions of gas shape improves towards larger radii. The predictive power reflects that accretion processes driven by the underlying dark matter distribution more heavily affect gas shapes in the outer regions of haloes, whereas baryonic effects fundamentally impact the gas shapes in the halo cores (See Figures 3 and 4).
- For TNG300 groups and clusters ($M_{500c} \geq 10^{13} M_{\odot}$), `TreeSHAP` identifies halo mass as the most important predictor of gas shapes at the core ($r/R_{200c} \leq 0.2$), and dark matter shape is the main predictor in the outer regions ($r/R_{200c} \geq 0.2$) of the haloes (See Figure 5).
- In galaxy size haloes ($10^{12} M_{\odot} \geq M_{500c} \geq 10^{13} M_{\odot}$) from TNG100, `TreeSHAP` identifies the halo concentration $c_{\rm vir}$ as the most important predictor of gas shapes in the inner ($r/R_{200c} \leq 0.2$) gas shapes while dark matter shape remains the most important predictor in the outer regions ($r/R_{200c} \geq 0.2$) (See Figure 6).
- We compare XGBoost with an alternative model, conditional abundance matching (see Figure 7). While CAM reproduces the cumulative distribution function of gas shapes (by construction), XGBoost generates more accurate predictions of gas shapes in the sampled haloes with the same input features to both models. We also note that CAM predictions improve across all metrics when using features identified from the XGBoost and SHAP pipeline (M_{500c} and $S_{\rm dm}$).

To our knowledge, the current work demonstrates one of the first applications of interpretable machine learning techniques (with XGBoost and SHAP) in connecting baryons with the underlying dark matter haloes. We specifically focus on understanding the physical modelling of the distribution of the hot X-ray emitting gas in galaxy groups and clusters. By applying machine learning techniques to the outputs of modern hydrodynamical cosmological simulations, we identified important physical properties that predict gas shapes at different radii in a dark matter halo. This work demonstrates that insights provided by interpretable machine learning approach can advance a physically-motivated and computationally efficient halo models for upcoming multi-wavelength cosmological surveys.

ACKNOWLEDGEMENTS

We thank Dhayaa Anbajagane, Han Aung, Andrew Hearin, Phil Mansfield, and Michelle Ntampaka for their constructive comments that improved the contents and presentation of the paper. CA was supported by the Leinweber Center for Theoretical Physics and the LSA Collegiate Fellowship at the University of Michigan. DB acknowledges support through NSF grant AST-1814053. AF is supported by a Michigan Institute for Data Science (MIDAS) Fellowship. EL acknowledges the University of Miami for support. We gratefully acknowledge use of the Harvard Odyssey and Yale Grace clusters as part of this work.

DATA AVAILABILITY

The TNG simulation data used in this paper is publicly available on `https://www.tng-project.org`. The scripts used to analyze the data and generate the plots are available upon request.

REFERENCES

Advani M. S., Saxe A. M., 2017, High-dimensional dynamics of generalization error in neural networks (`arXiv:1710.03667`)
Allen S. W., Evrard A. E., Mantz A. B., 2011, ARA&A, 49, 409
Allgood B., Flores R. A., Primack J. R., Kravtsov A. V., Wechsler R. H., Faltenbacher A., Bullock J. S., 2006, MNRAS, 367, 1781
Amodeo S., et al., 2020, arXiv e-prints, p. arXiv:2009.05558
Ansarifard S., et al., 2020, A&A, 634, A113
Bailin J., Steinmetz M., 2005, ApJ, 627, 647
Barnes D. J., et al., 2018, MNRAS, 481, 1809
Biffi V., Dolag K., Böhringer H., 2013, MNRAS, 428, 1395
Bryan G. L., Norman M. L., 1998, ApJ, 495, 80
Bryan S. E., Kay S. T., Duffy A. R., Schaye J., Dalla Vecchia C., Booth C. M., 2013, MNRAS, 429, 3316
Buote D. A., Humphrey P. J., 2012, MNRAS, 420, 1693
Chan T. K., Kereš D., Oñorbe J., Hopkins P. F., Muratov A. L., Faucher-Giguère C. A., Quataert E., 2015, MNRAS, 454, 2981
Chen T., Guestrin C., 2016, arXiv e-prints, p. arXiv:1603.02754
Chen H., Avestruz C., Kravtsov A. V., Lau E. T., Nagai D., 2019, MNRAS, 490, 2380
Chen H., Janizek J. D., Lundberg S., Lee S.-I., 2020, True to the Model or True to the Data? (`arXiv:2006.16234`)
Chua K. T. E., Pillepich A., Vogelsberger M., Hernquist L., 2019, MNRAS, 484, 476
Cole S., Lacey C., 1996, MNRAS, 281, 716
Debattista V. P., Moore B., Quinn T., Kazantzidis S., Maas R., Mayer L., Read J., Stadel J., 2008, ApJ, 681, 1076
Despali G., Giocoli C., Tormen G., 2014, MNRAS, 443, 3208
Dubinski J., Carlberg R. G., 1991, ApJ, 378, 496
Hearin A. P., Watson D. F., 2013, MNRAS, 435, 1313
Hearin A. P., et al., 2017, AJ, 154, 190
Hill J. C., Spergel D. N., 2014, J. Cosmology Astropart. Phys., 2014, 030
Hojjati A., et al., 2017, MNRAS, 471, 1565
Jing Y. P., Suto Y., 2002, ApJ, 574, 538
Kazantzidis S., Kravtsov A. V., Zentner A. R., Allgood B., Nagai D., Moore B., 2004, ApJ, 611, L73
Kravtsov A. V., Borgani S., 2012, ARA&A, 50, 353
Lau E. T., Nagai D., Kravtsov A. V., Zentner A. R., 2011, ApJ, 734, 93
Lau E. T., Nagai D., Kravtsov A. V., Vikhlinin A., Zentner A. R., 2012, ApJ, 755, 116
Lau E. T., Hearin A. P., Nagai D., Cappelluti N., 2021, MNRAS, 500, 1029
Lundberg S., Lee S.-I., 2017, A Unified Approach to Interpreting Model Predictions (`arXiv:1705.07874`)
Lundberg S. M., et al., 2020, Nature Machine Intelligence, 2, 56

Ma S., Tourani R., 2020, Predictive and Causal Implications of using Shapley Value for Model Interpretation (arXiv:2008.05052)

Marinacci F., et al., 2018, MNRAS, 480, 5113

Masaki S., Lin Y.-T., Yoshida N., 2013, MNRAS, 436, 2286

Nagai D., Kravtsov A. V., Vikhlinin A., 2007, ApJ, 668, 1

Naiman J. P., et al., 2018, MNRAS, 477, 1206

Nakkiran P., 2019, More Data Can Hurt for Linear Regression: Sample-wise Double Descent (arXiv:1912.07242)

Natekin A., Knoll A., 2013, Frontiers in Neurorobotics, 7, 21

Navarro J. F., Frenk C. S., White S. D. M., 1996, ApJ, 462, 563

Nelson D., et al., 2018, MNRAS, 475, 624

Ntampaka M., et al., 2019, ApJ, 876, 82

Osato K., Flender S., Nagai D., Shirasaki M., Yoshida N., 2018a, MNRAS, 475, 532

Osato K., Nishimichi T., Oguri M., Takada M., Okumura T., 2018b, MNRAS, 477, 2141

Osato K., Shirasaki M., Miyatake H., Nagai D., Yoshida N., Oguri M., Takahashi R., 2020, MNRAS, 492, 4780

Pillepich A., et al., 2018, MNRAS, 475, 648

Planck Collaboration et al., 2016, A&A, 594, A13

Pratt G. W., Arnaud M., Biviano A., Eckert D., Ettori S., Nagai D., Okabe N., Reiprich T. H., 2019, Space Sci. Rev., 215, 25

Rodriguez-Gomez V., et al., 2015, MNRAS, 449, 49

Samsing J., Skielboe A., Hansen S. H., 2012, ApJ, 748, 21

Schaller M., et al., 2015, MNRAS, 451, 1247

Schneider M. D., Frenk C. S., Cole S., 2012, J. Cosmology Astropart. Phys., 2012, 030

Shapley L. S., 1953, Annals of Mathematics Studies, 28, 307

Shaw L. D., Weller J., Ostriker J. P., Bode P., 2006, ApJ, 646, 815

Springel V., 2010, MNRAS, 401, 791

Springel V., et al., 2018, MNRAS, 475, 676

Suto D., Peirani S., Dubois Y., Kitayama T., Nishimichi T., Sasaki S., Suto Y., 2017, PASJ, 69, 14

Van Waerbeke L., Hinshaw G., Murray N., 2014, Phys. Rev. D, 89, 023508

Vogelsberger M., et al., 2018, MNRAS, 474, 2073

Wechsler R. H., Tinker J. L., 2018, ARA&A, 56, 435

Zemp M., Gnedin O. Y., Gnedin N. Y., Kravtsov A. V., 2011, ApJS, 197, 30

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