Mimicking the halo–galaxy connection using machine learning

Natalí S. M. de Santi, N. V. N. Rodrigues, A. D. Montero-Dorta, L. R. Abramo, B. Tucci and M. C. Artale

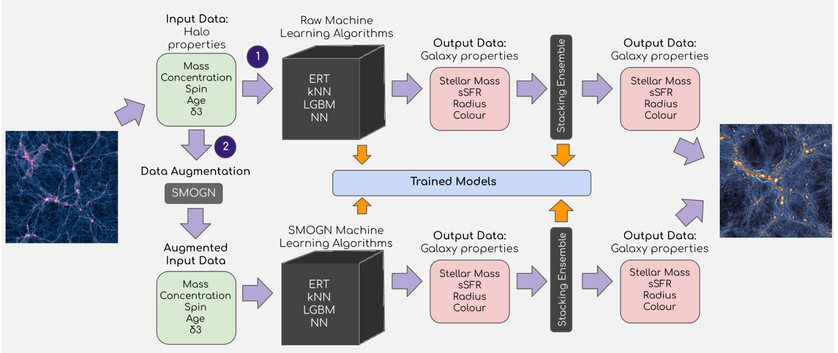
Institute of Physics, University of São Paulo, Brazil

Abstract

As far as we know, galaxies form inside dark matter halos and elucidating this connection is a key element in theories of galaxy formation and evolution. In this work, we propose a suite of machine learning (ML) tools to analyze these intricate relations in the IllustrisTNG300 magnetohydrodynamical simulation. We apply four individual algorithms: extremely randomized trees (ERT), K-nearest neighbors (kNN), light gradient boosting machine (LGBM), and neural networks (NN). Moreover, we combine the results of the different methods in a stacked model. In addition, we apply all these methods in an augmented dataset using the synthetic minority over-sampling technique for regression with Gaussian noise (SMOGN), in order to alleviate the problem of imbalanced data sets, and show that it improves the shape of the predicted distributions. Overall, all the ML algorithms produce consistent results in terms of predicting central galaxy properties from a set of input halo properties that include halo mass, concentration, spin, and halo overdensity. For stellar mass, the (predicted vs. real) Pearson correlation coefficient is 0.98, dropping down to 0.7-0.8 for specific star formation rate, colour, and size. We also demonstrate that our predictions are good enough to reproduce the power spectra of multiple galaxy populations, defined in terms of stellar mass, sSFR, colour, and size with high accuracy. Our analysis adds evidence to previous works indicating that certain galaxy properties cannot be reproduced using halo features alone.

Overall idea of the work

The idea of this work is to provide a ML suite able to get halo and predict central galaxy properties. We provide the trained models and the analysis of the results with the predictions:



The dataset - IllustrisTNG

We use the IllustrisTNG magnetohydrodynamical cosmological simulation [1], produced using the AREPO moving-mesh code. We analyze the largest box, IllustrisTNG300-1 (TNG300) that spans a side length of $205\ h^{-1}$Mpc and evolve 2500^3 dark matter (DM) particles of mass $4.0\times10^7\ h^{-1}\mathrm{M}_\odot$ and (initially) 2500^3 gas cells of mass $7.6\times10^6\ h^{-1}\mathrm{M}_\odot$. The input halo properties are:

- **Mass:** $M_{\rm vir}\ [h^{-1}\mathrm{M}_\odot]$, computed by adding up the mass of all gas cells and particles contained within a sphere of radius $R_{\rm vir}$;
- **Age:** $z_{1/2}$, a formation redshift, defined as the one at which half of the present day halo mass has been accreted into a single subhalo for the first time;
- **Spin:** $\lambda_{\rm halo} = \frac{|J|}{\sqrt{2}M_{\rm vir}V_{\rm vir}R_{\rm vir}}$;
- **Concentration:** $c_{\rm vir} = \frac{R_{\rm vir}}{R_{\rm s}}$;
- **Overdensity:** δ_3, defined as the number density of subhaloes within a sphere of radius $R = 3h^{-1}$Mpc.

To compute the following galaxy properties:

- **Stellar mass:** $M_*\ [h^{-1}\mathrm{M}_\odot]$, defined as the total mass of all stellar particles bound to each subhalo;
- **Specific star formation rate:** sSFR $[\mathrm{yr}^{-1}h]$, defined as the star formation rate SFR $[\mathrm{yr}^{-1}\mathrm{M}_\odot]$ (computed as the sum of the star formation rate of all gas cells contained in each subhalo) per unit stellar mass: sSFR $=$ SFR$/M_*$;
- **Radius:** $R^{(*)}_{1/2}\ [h^{-1}\,\mathrm{kpc}]$, stellar (3D) half-mass radius, defined as the comoving radius containing half of the stellar mass of each subhalo;
- **Colour:** (g-i), derived from the magnitudes provided at the IllustrisTNG database.

Machine Learning tools

In this work we solve a regression problem to find the map $F(\mathbf{x} \to y)$ between the halo properties $\mathbf{x}$ and each galaxy feature y. To do this task we have used four well known different ML algorithms: **extremely randomized trees** (ERT) [2], **light gradient boosting machine (LGBM)** [4], **k-nearest neighbours (kNN)** [3] and **neural networks (NN)** [5]. The two first methods are tree-based, kNN is a clusterization algorithm and the last one is based in a complex arrange of neuron units.

Besides, we bring two new aspects:

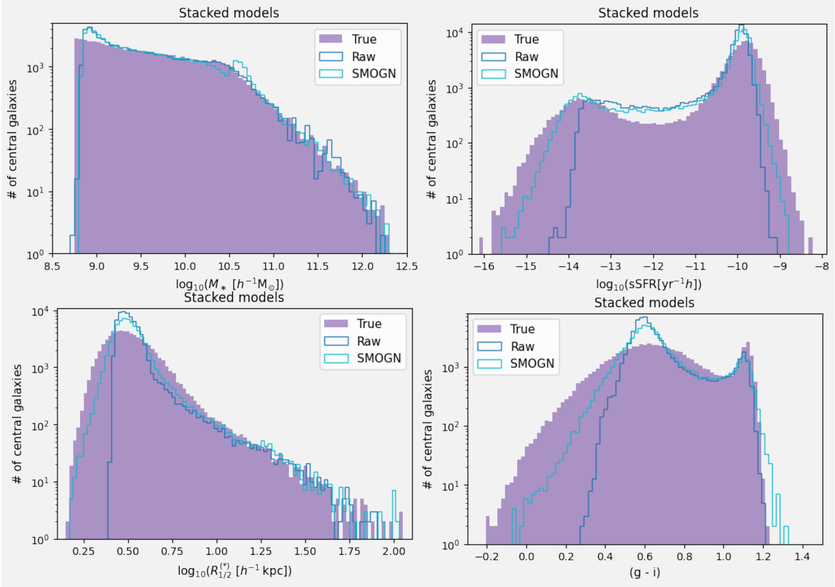
- **Stacking ensemble model:** used to combine the different predictions from the different ML algorithms in order to improve upon them, not choosing a single model. It forms linear combinations of different predictors [6] and summarizes their results;
- **Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN):** used to solve the problem of imbalanced data sets to perform prediction of underrepresented regions of the data. The algorithm initially bins the data for a given target variable and subsequently splits the resulting distribution in "rare" and "normal" bins. Rare bins are augmented, whereas normal bins are under-sampled [7].

The metrics used in the analysis are:

- **Mean Squared Error:** $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{\rm pred} - y_i^{\rm true}\right)^2$
- **Pearson Correlation Coefficient:** $\mathrm{PCC} = \frac{\mathrm{cov}\left(y^{\rm pred}, y^{\rm true}\right)}{\sigma_{y^{\rm pred}}\sigma_{y^{\rm true}}}$

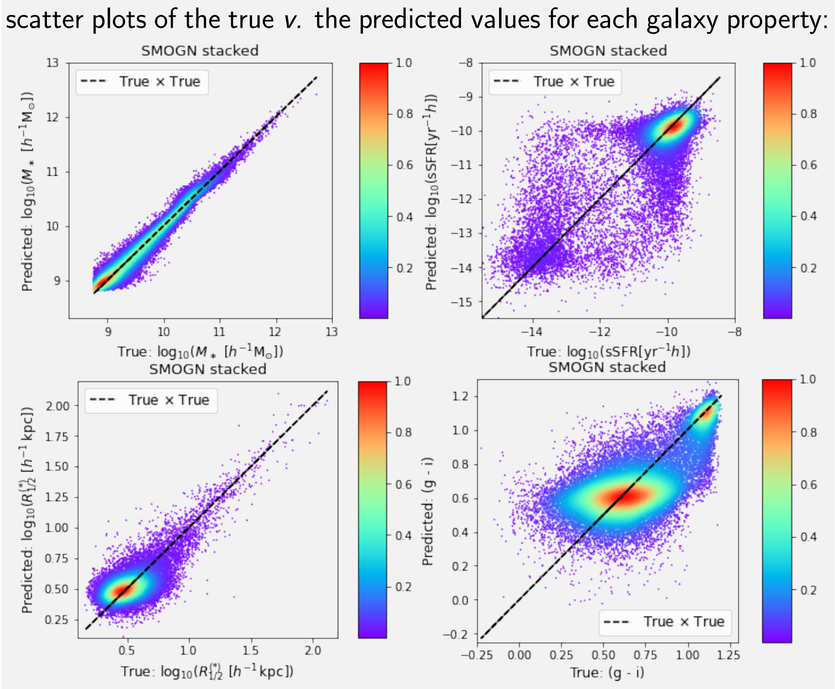
Distributions of the galaxy properties

This work primarily tries to reproduce the frequencies (distributions) of each galaxy property:



Predicted versus True relation

The predictions on a one-to-one basis are quantitatively given by the scatter plots of the true *v.* the predicted values for each galaxy property:

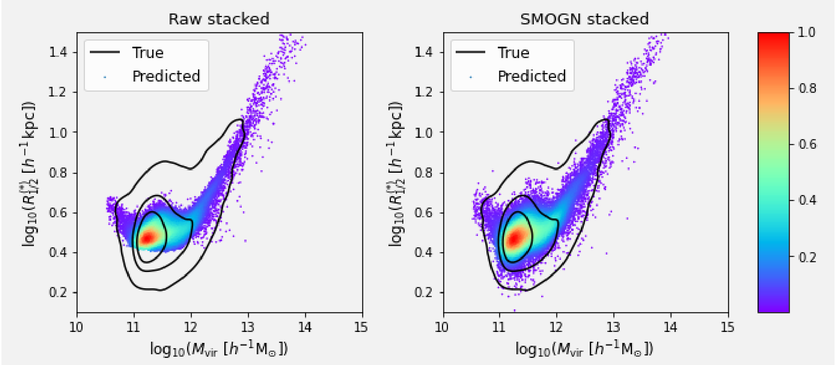


Numerically, it is given by the scores:

Property	Raw		SMOGN	
	MSE	PCC	MSE	PCC
Stellar mass	0.017	0.98	0.018	0.98
sSFR	0.691	0.79	0.747	0.79
Radius	0.012	0.75	0.014	0.71
Colour	0.032	0.71	0.036	0.67

Halo/Galaxy mass relation

The radius *v.* halo mass relations demonstrate that the machine produces fairly realistic relations between the properties:

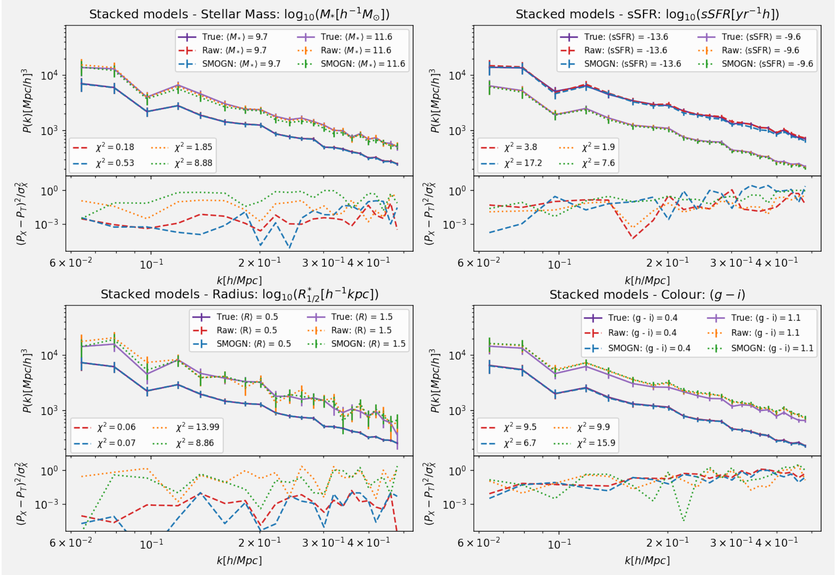


Power spectrum

By splitting those galaxies in two populations according to these properties, and then computing the clustering of each population, we can check whether our predictions are able to separate the galaxies correctly, according to the types of haloes that they inhabit. The error bars in the power spectrum were estimated using the FKP variance [8] and we have measured the residuals

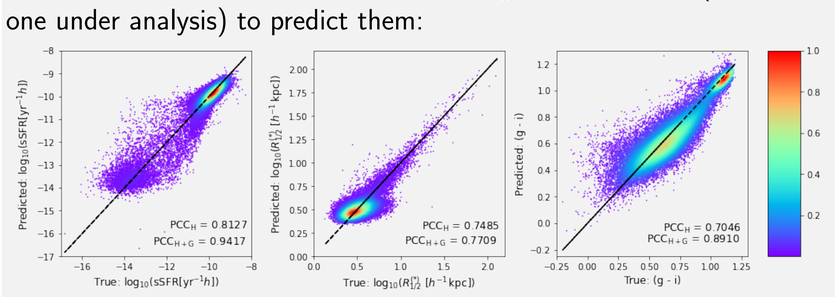
$$\chi^2(k) = \frac{\left[P^{Pred}_{\alpha,i}(k) - P^{True}_{\alpha,i}(k)\right]^2}{\left[\sigma^{Pred}_{P_{\alpha,i}}\right]^2}, \qquad (1)$$

for each bin of k and showed their sum over all their bins.



Using not only halo properties

In searching to improve the predictions for galaxy sSFR, radius, and colour, we have tried to include the proper galaxy properties (except the one under analysis) to predict them:



Conclusions

- The overall distributions are improved using the SMOGN augmented dataset, predicting galaxies in their tails;
- The scores for the raw models are still better than the SMOGN results;
- ML models applied in the raw dataset tend to concentrate on the bulk of the distributions, which hinders the prediction of scatter. SMOGN increases the desired scatter presented in the simulation;
- All the values of χ^2 summed up in k are lower than the number of degrees of freedom (22 bins of k), showing the good results in the clustering properties;
- The inclusion of other galaxy properties led to improvements in the metrics and a decreasing in the scatter in the one-to-one galaxy properties.

Acknowledgments

NSMS acknowledges financial support from FAPESP, grant 19/13108-0.

References

[1] Pillepich A., et al., 2018, MNRAS, 473, 4077.
[2] Geurts, P. 2006, Machine Learning, 63, 3.
[3] Ivezic Z., et al., 2014, Princeton University Press.
[4] Ke G., Meng Q., Finley T., et al., 2017, 30, Advances in Neural Information Processing Systems.
[5] Chollet F., 2017, Deep Learning with Python. Manning Publications Company.
[6] Breiman L., 1996, Machine Learning, 24.
[7] Branco P., , et al., 2017, Proceedings of Machine Learning Research Vol. 74.
[8] Feldman, H. A., , et al., 1994, ApJ, 426, 23. doi:10.1086/174036.