Inferring accreted stellar mass fractions of central galaxies using random forest

Arshia Anjum and Sibabrata Biswal
Group 7
Machine Learning Course 2023

Abstract

The formation and evolution of the universe, and the galaxies in it is a much researched topic, with too less labour to look deep into it. Using Machine Learning to understand this evolution through simulations is the next best option to solve the mysteries within a satisfactory error bar. One of the main topics for the same questions the amount of mass formed inside the parent halo/galaxy, compared to the amount of mass accreted by the same through its satellite halos, and it gains significance as it provides insights in the assembly history of galaxies, the merger history as well as the interactions on which galaxy evolution depends. In this report, we have used random forest (RF) as a means of studying the same and predicting the ex-situ mass fractions (f_{acc}) of various galaxies using the TNG simulation.

1 Introduction

The origin and evolution of galaxies are two of the most active fields of astrophysical research. The Lambda Cold Dark Matter (or ΛCDM) hypothesis is the most recent manifestation of our knowledge of the origins of the Universe. It advances the big bang hypothesis by positing that most of the physical substances in the Universe is made up of a material known as dark matter. Galaxies arise in the ΛCDM structure creation paradigm by the cooling and condensation of gas at the centre of dark matter halos. According to the theory, galaxy formation occurs in two stages: an early rapid production of in-situ stars by gas cooling, followed by a later period of mass increase of ex-situ stars via accretion of smaller satellite galaxies. These satellite galaxies were earlier considered as the central galaxies of smaller halos. Satellite galaxies, or the subhalos as we call them after they fall into the larger halos, loose stellar mass through tidal stripping.

One of the reasons we can differentiate between the in-situ and ex-situ mass is that, the accreted stellar mass makes the outer regions of the parent halo, and are metal poor as compared to the in-situ mass. The next question which arises is, what is the importance of finding the ex-situ mass fraction (which will be referred as f_{acc} throughout the report). The ex-situ mass fraction of a galaxy is the fraction of its total mass that comes from accreted material, which includes gas and stars that were not originally formed within the galaxy itself, but were instead acquired through mergers or interactions with other galaxies. The ex-situ mass fraction derives its significanvce from the fact that it gives information on the he assembly history of galaxies, the merger history as well as the interactions on which galaxy evolution depends. Also, it comments of=n the galaxy properties, surroundings as well as the age of the galaxy.

As mentioned, we use RF to study the process of stellar assembly through the TNG simulations. IllustrisTNG is a suite of large volume, cosmological, gravo-magnetohydrodynamical simulations run with the moving-mesh code AREPO. The simulation solves coupled evolution of dark matter, cosmic gas, stars, supermassive black holes, starting with the highest redshift of 127 to 0, i.e. the present day.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 Data

2.1 The Chosen Data

As mentioned in the project proposal, the data used for the model is from the Illustris-TNG simulation. However, the data we were interested in was TNG-100, which was around 2TB in size, hence more rigorous to work with. Hence, we decided to write and test the code using the TNG50-4 simulation data, which is a low resolution simulation, but has the same data format as the high resolution TNG-100. Once the code is complete, we shall download the TNG-100 files, and run it using those files to get our model.

2.2 Data Extraction and Construction

Using an authentic API key, the data was extracted from the official website of the Illustris-TNG simulations. The way to go ahead with the procedure would be, to track the particles in the simulation at each redshift (point in time) and maintain a label for them. At each redshift, a friends-of-friends and subfind algorithm shall also be used to identify the halos and subhalos to which the particles belong. AS we keep track of them, in the present day data file, we would know the assembly history of the particle, hence be able to jusdge if it contributes to the in-situ mass or ex-situ mass of the galxy. Since this would have been a more cumbersome method, we used an already existing catalogue which does all this, and gives us the final labels of the particles. Therefore, the data we are working with are: the snapshots, the group catalogues, the offsets and the stellar assembly catalogues of the simulation.

To construct our data, we first work with the stellar assembly catalogue. After reading the file and sorting for redshift =0, we apply our first constraint, i.e. the mass of the central galaxies of the halos should be greater than $10^{10.16} M_{Sun}$. This constrain exists, because (i) The resolution limit for the TNG data is around $7.46 \times 10^8 M_{Sun}$ and (ii) the accuracy of morphology, rotation, and shape of the galaxies of interest deters below $10^9 M_{Sun}$. Hence, making us choose the galaxies whose mass is not less than $10^{10} M_{Sun}/h$. The second constraint had to be to check if the simulation gives faithful mock images for the chosen subhalos. In the context of TNG simulation, a faithful mock image refers to a computer-generated image or simulation that accurately represents a real-world phenomenon or system. There will be a need of SKIRT imaging data for the same. However, this was not possible to do here, as the data was low resolution. The third constraint was to check if the central galaxy of the halo is at least 0.5 magnitudes brighter than the satellite galaxies. However, this can be skipped, as this constraint does not have any manor impact on the model.

As we sort the stellar assembly data, we store the index of the subhalos (the subhalo ids). These ids are then used to get the galaxy features from the snapshot, group catalogue and offset files. The columns are then concatenated, and used for the model training.

3 Halo and Galaxy features used

We briefly describe the halo and galaxy features present in the data.

1. SubhaloBHMass: the estimated mass of a black hole that resides within a subhalo.
2. SubhaloGasMetalFractions: fraction of metals present in the gas component of a subhalo. (For carbon, nitrogen, oxygen, neon, magnesium, silicon, sulfur, calcium, iron, and nickel)
3. SubhaloGasMetalFractionsHalfRad: The fraction of metals present in the gas component of a subhalo within half of the subhalo's maximum circular velocity radius. (For carbon, nitrogen, oxygen, neon, magnesium, silicon, sulfur, calcium, iron, and nickel)
4. SubhaloGasMetalFractionsSfr: The fraction of metals present in the gas component of a subhalo that is actively forming stars. (For carbon, nitrogen, oxygen, neon, magnesium, silicon, sulfur, calcium, iron, and nickel)
5. SubhaloGasMetallicity: The metallicity of the gas component of a subhalo, defined as the fraction of the gas mass that is composed of heavy elements.
6. SubhaloGasMetallicityHalfRad: The metallicity of the gas component of a subhalo within half of the subhalo's maximum circular velocity radius.
7. SubhaloLen: The number of particles used to represent a subhalo in the simulation.

8. SubhaloMass: The total mass of a subhalo, including all components such as gas, stars, and dark matter.
9. SubhaloMassInHalfRad: The total mass of a subhalo within half of the subhalo's maximum circular velocity radius.
10. SubhaloMassInRad: The total mass of a subhalo within the subhalo's maximum circular velocity radius.
11. SubhaloSFRinHalfRad: The star formation rate within half of the subhalo's maximum circular velocity radius.
12. SubhaloSFRinRad: The star formation rate within the subhalo's maximum circular velocity radius.
13. SubhaloSpin: The angular momentum of a subhalo, which can affect its morphology and evolution.
14. SubhaloStarMetalFractions: The fraction of metals present in the star component of a subhalo. (For carbon, nitrogen, oxygen, neon, magnesium, silicon, sulfur, calcium, iron, and nickel)
15. SubhaloStarMetalFractionsHalfRad: The fraction of metals present in the star component of a subhalo within half of the subhalo's maximum circular velocity radius. (For carbon, nitrogen, oxygen, neon, magnesium, silicon, sulfur, calcium, iron, and nickel)
16. SubhaloStarMetallicity: The metallicity of the star component of a subhalo, defined as the fraction of the star mass that is composed of heavy elements.
17. SubhaloStarMetallicityHalfRad: The metallicity of the star component of a subhalo within half of the subhalo's maximum circular velocity radius.
18. SubhaloStellarPhotometrics: The properties of the stellar population in a subhalo, including luminosities and colors. (For U, B, V, K, g, r, i, z)
19. SubhaloVelDisp: The velocity dispersion of the stars in a subhalo.
20. SubhaloWindMass: The mass of gas that has been ejected from a subhalo due to feedback from star formation or black hole activity.
21. SubhaloHalfmassRad: The radius within which half of the total mass of a subhalo is contained.
22. SubhaloSFR: The star formation rate of a subhalo, measured in units of solar masses per year. This is the rate at which gas is being converted into stars within the subhalo.

The features related to satellite galaxies are not used, as the satellite galaxies mass limit is below the resolution of the simulation, and hence will not be a good testing criterion. Also, most of the galaxies studied do not have satellite galaxies above the range of $10^8 M_{Sun}$.

4 Machine Learning Methodology

4.1 Decision Trees

Decision trees, a type of machine learning algorithm, is used for classification and regression tasks. They work by recursively splitting the data-set into subsets based on the most informative feature, thus creating a tree-like structure. Decision trees can handle both categorical and numerical data, however, since we need numerical answers, we focus on the regression type trees. They are easy to interpret and visualize, and can handle noisy data.

4.2 Random Forest

Random forest, that utilizes decision trees for classification and regression, is an ensemble learning method. It works by constructing multiple decision trees using random subsets of the training data and features, and combining their predictions through averaging or voting. This helps in reducing over-fitting, increasing accuracy, and provides measures of feature importance. Random forest has several hyper-parameters such as the number of trees, the size of the subsets, and the depth of the trees. These can be tuned using cross-validation to find the optimal combination for the specific problem.

4.3 Description of the Model

We used RandomForestRegressor from the scikit-learn machine learning library to model the relationship between our input variables and the target variable. The random forest algorithm is an ensemble learning method that fits multiple decision tree models on randomly selected subsets of the data, and aggregates the predictions of each individual tree to improve overall predictive accuracy.

We utilized several hyperparameters to fine-tune the performance of the random forest regressor. One key hyperparameter is the number of decision trees in the forest, which we set to 100. Another important hyperparameter is the "bootstrap" setting, which determines whether each tree is fit on a bootstrapped sample of the data (with replacement) or the entire dataset. In our analysis, we set the bootstrap parameter to "True", which enables bootstrapping.

We also utilized the "out-of-bag" (OOB) score as a metric to evaluate the performance of our model. The OOB score measures the predictive accuracy of the model on data points that were not included in the training set for each individual tree. This provides an estimate of how well the model is likely to generalize to new, unseen data.

5 Provisional Results

As seen in the TNG50-4 data, the following are the provisional results:

1. The subhalos that can be used in the dataset are just 265 out of 22869 total subhalos. This is due to the low resolution data used for the training.
2. Checking the luminosity of the central galaxies compared to the satellite galaxies as a constraint, was not helpful, as it should have been.
3. The major features that we got from the model were: SubhaloMass, SubhaloMassInHalfRad, SubhaloBHMass, SubhaloStarMetalFractions, SubhaloStarMetallicity

Some of the relations observed in the model are expressed as a plot in Fig 1.

6 Future Plans

As regards our future plans in the project, we are interested in carrying out the same procedure in TNG300 and TNG50 high resolution data, as mentioned during proposal. We would focus on the specific properties in both the data, one has better statistical properties, while the other has better structural properties. Also, we will apply mass limits in the data set to split it into 2, which we could not apply in this due to low data. We will also be applying the SKIRT Imaging data constraints, and additionally focussing on the observable features alone. Lastly, our highly ambitious goal of trying a similar procedure for black hole systems also remains.

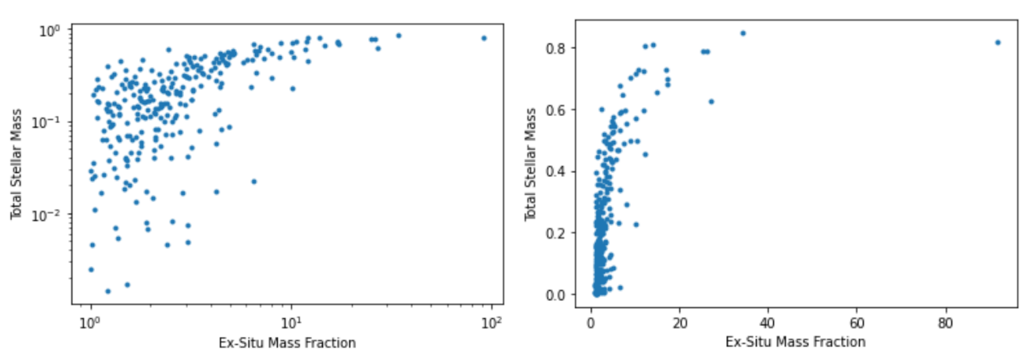

Figure 1: The Ex-Situ Stellar mass fraction's relationship with the Total Stellar mass of the subhalo (i) Log-Log Function (ii) Without Log

References

[1] R. Shi, W. Wang, Z. Li, J. Han, J. Shi, V. Rodriguez-Gomez, Y. Peng and Q. Li, Mon. Not. Roy. Astron. Soc. **515** (2022) no.3, 3938-3955 doi:10.1093/mnras/stac1541 [arXiv:2112.07203 [astro-ph.GA]].

[2] V. Rodriguez-Gomez, A. Pillepich, L. V. Sales, S. Genel, M. Vogelsberger, Q. Zhu, S. Wellons, D. Nelson, P. Torrey and V. Springel, *et al.* Mon. Not. Roy. Astron. Soc. **458** (2016) no.3, 2371-2390 doi:10.1093/mnras/stw456 [arXiv:1511.08804 [astro-ph.GA]].

[3] A. Pillepich, M. Vogelsberger, A. Deason, V. Rodriguez-Gomez, S. Genel, D. Nelson, P. Torrey, L. V. Sales, F. Marinacci and V. Springel, *et al.* Mon. Not. Roy. Astron. Soc. **444** (2014) no.1, 237-249 doi:10.1093/mnras/stu1408 [arXiv:1406.1174 [astro-ph.GA]].

[4] D. Montenegro-Taborda, V. Rodriguez-Gomez, A. Pillepich, V. Avila-Reese, L. V. Sales, A. Rodríguez-Puebla and L. Hernquist, doi:10.1093/mnras/stad586 [arXiv:2302.10943 [astro-ph.GA]].

[5] S. Tacchella, B. Diemer, L. Hernquist, S. Genel, F. Marinacci, D. Nelson, A. Pillepich, V. Rodriguez-Gomez, L. V. Sales and V. Springel, *et al.* Mon. Not. Roy. Astron. Soc. **487** (2019) no.4, 5416-5440 doi:10.1093/mnras/stz1657 [arXiv:1904.12860 [astro-ph.GA]].

[6] S. Wellons, P. Torrey, C. P. Ma, V. Rodriguez-Gomez, M. Vogelsberger, M. Kriek, P. van Dokkum, E. Nelson, S. Genel and A. Pillepich, *et al.* Mon. Not. Roy. Astron. Soc. **449** (2015) no.1, 361-372 doi:10.1093/mnras/stv303 [arXiv:1411.0667 [astro-ph.GA]].