Inferring accreted stellar mass fractions of central galaxies using random forest

Arshia Anjum and Sibabrata Biswal

Supervisor- Dr. Subhankar Mishra

Machine Learning Project

Introduction

- Lambda CDM model
- In-situ and ex-situ mass fractions
- Galaxies arise in the Lambda CDM 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.
- The accreted stellar mass makes the outer regions of the parent halo, and are metal poor as compared to the in-situ mass.
- The ex-situ mass fraction derives its significance from the fact that it gives information on the assembly history of galaxies, the merger history as well as the interactions on which galaxy evolution depends. Also, it comments on the galaxy properties, surroundings as well as the age of the galaxy.

Data

The Chosen Data

- As mentioned in the project proposal, the data used for the model is from the Illustris-TNG simulation.
- TNG50-4 simulation data, which is a low resolution simulation, but has the same data format as the high resolution TNG-100.

Data Extraction and Construction

- 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 judge if it contributes to the in-situ mass or ex-situ mass of the galxy.

Data

Data Extraction and Construction

- The snapshots, the group catalogues, the offsets and the stellar assembly catalogues of the simulation.
- 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 constraint 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.
- 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 major impact on the model.

Data

Data Extraction and Construction

- 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.

Halo and Galaxy features used

- SubhaloBHMass
- SubhaloGasMetalFractions: (For carbon, nitrogen, oxygen, neon, magnesium, silicon, sulfur, calcium, iron, and nickel)
- SubhaloGasMetalFractionsHalfRad: (For carbon, nitrogen, oxygen, neon, magnesium, silicon, sulfur, calcium, iron, and nickel)
- SubhaloGasMetalFractionsSfr: (For carbon, nitrogen, oxygen, neon, magnesium, silicon, sulfur, calcium, iron, and nickel)
- SubhaloGasMetallicity
- SubhaloGasMetallicityHalfRad
- SubhaloLen
- SubhaloMass
- SubhaloMassInHalfRad
- SubhaloMassInRad
- SubhaloSFRinHalfRad
- SubhaloSFRinRad
- SubhaloSpin
- SubhaloStarMetalFractions: (For carbon, nitrogen, oxygen, neon, magnesium, silicon, sulfur, calcium, iron, and nickel)
- SubhaloStarMetalFractionsHalfRad: (For carbon, nitrogen, oxygen, neon, magnesium, silicon, sulfur, calcium, iron, and nickel)
- SubhaloStarMetallicity
- SubhaloStarMetallicityHalfRad
- SubhaloStellarPhotometrics: (For U, B, V, K, g, r, i, z)
- SubhaloVelDisp
- SubhaloWindMass
- SubhaloHalfmassRad
- SubhaloSFR

Machine Learning Methodology

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.

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.

Macine Learning Methodology

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.
- 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 200. Another important hyperparameter is the “bootstrap" setting which we set to "True", that 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.

Provisional Results

TNG50-4

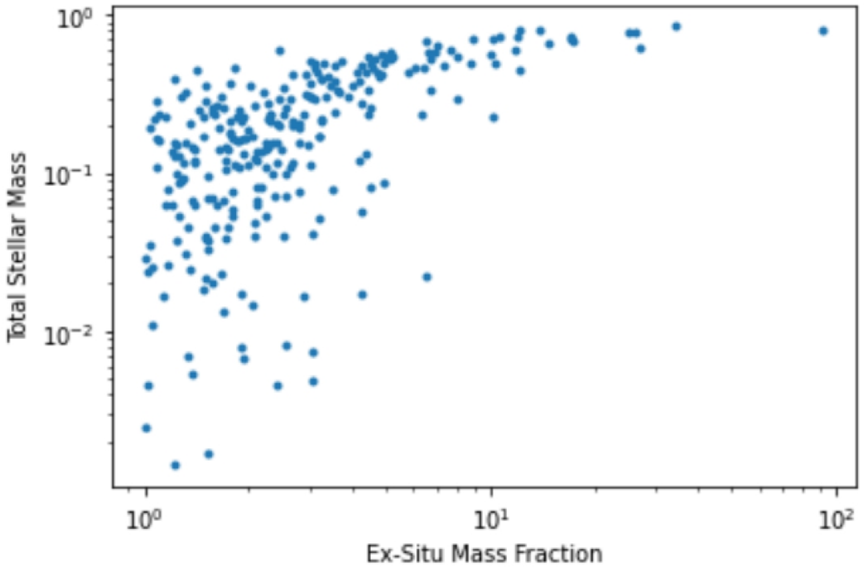
- The subhalos that can be used in the dataset are just 265 out of 22869 total subhalos. This is due to the low resolution data used for the training.

- Checking the luminosity of the central galaxies compared to the satellite galaxies as a constraint, was not helpful, as it should have been.

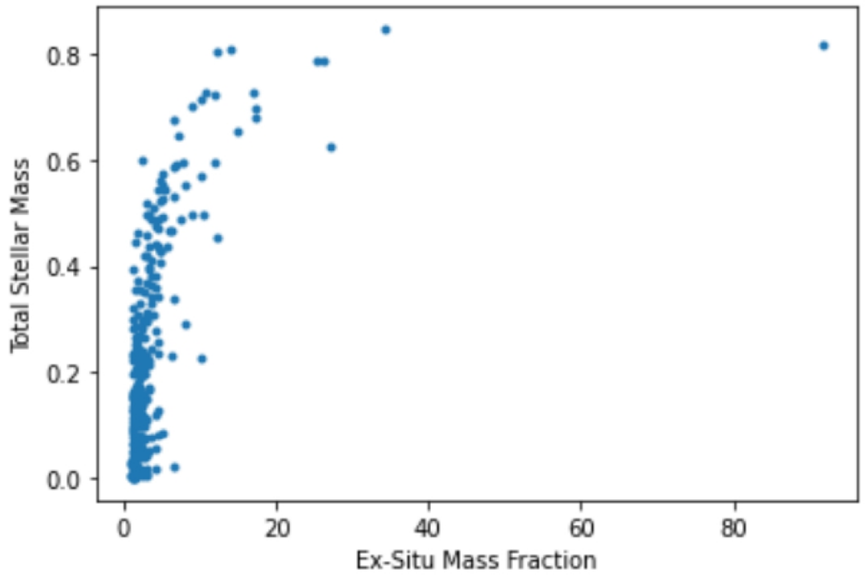
- The major features that we got from the model were: SubhaloMass, SubhaloStellarPhotometrics(V), SubhaloMassInHalfRad, SubhaloStarMetallicityHalfRad, SubhaloStellarPhotometrics(I)

- The accuracy of the model without Feature Engineering:
  R2 Score: 0.763904
  OOB Score: 0.668257

- R2 score in reference papaer for TNG100 data: 92.3%



The relationship between certain features and exsitu mass fraction has high standard deviation, which can be fixed with either a log log plot, or maximum normalisation.



Future Work

- As regards our future plans in the project, we are interested in carrying out the same procedure in TNG300 and TNG50 high resolution data
- We would focus on the specific properties in both the data, one has better statistical properties, while the other has better structural properties.
- 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.

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]].