Galaxy-by-galaxy Emulation of Cosmo-hydrodynamical Simulations of Galaxy Formation

Data Science Institute
COLUMBIA UNIVERSITY

Sicheng Li, Junsheng Shi, Shangzhi Liu, Chen Jin, Wen Zhan
Shy Genel, Yongseok Jo

Data Science Capstone Project
with Flatiron Institute &
Columbia Astrophysics Laboratory

Goal of Project

The CAMELS(Cosmology and Astrophysics with MachinE Learning Simulations) project is designed to model and investigate the formation and evolution of galaxies and the physical processes involved in them. In our capstone project, we are working based on the CAMELS dataset. In particular, we are working on 19 key features from two simulations (Illustris-TNG and SIMBA). The goal of this project is to use the result from one simulation to predict the other, and vice versa.

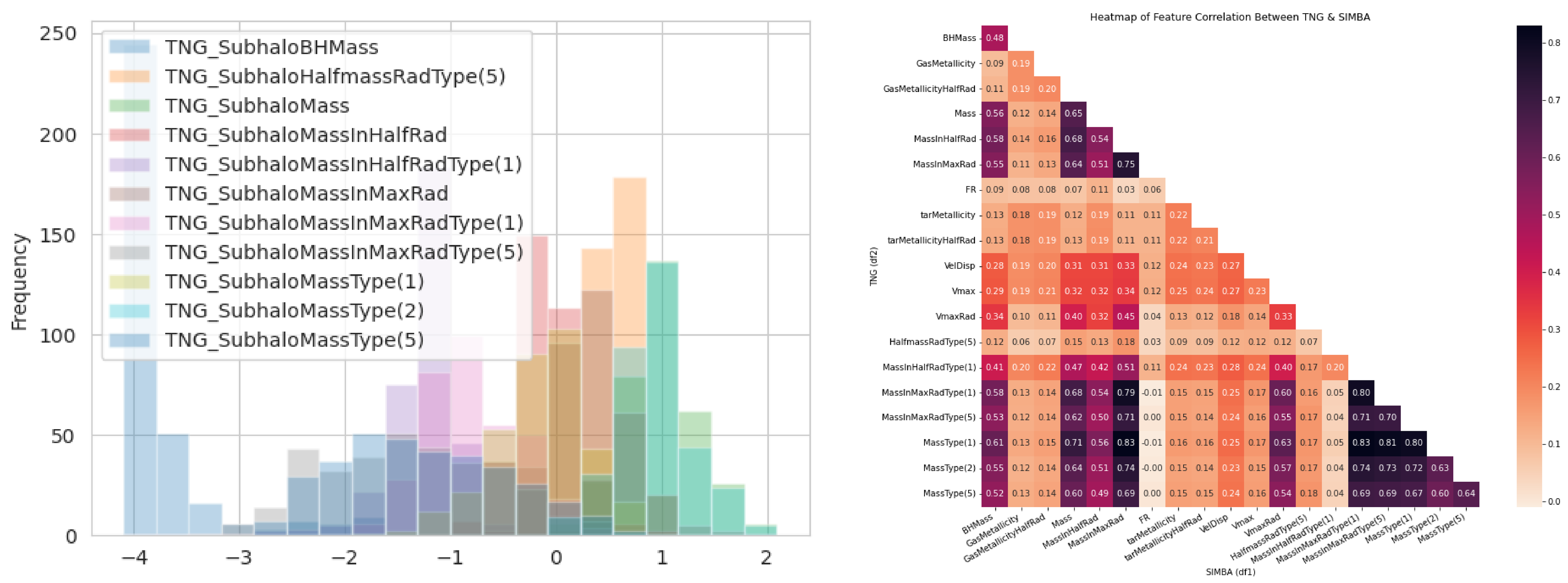

Figure 1. Adjusted Mass Distribution and TNG to SIMBA Heatmap.

Inner and Inter Simulation Prediction Model

Based on the given matching catalogs between TNG, SIMBA simulations and shadow versions of CV_0 TNG, we obtained our bijective matching data by matching 19 key features. Then we used these data to train our inner-simulation prediction model, which is to predict one feature from the other features in the same simulation, and our inter-simulation prediction model, which is to predict 19 features from TNG/SIMBA from the other simulation or shadow versions. We used Linear Regression, XGBoost Regression, and Multi Layer Perceptron Regression models, and focused on R2 score and RMSE to check if models work well.

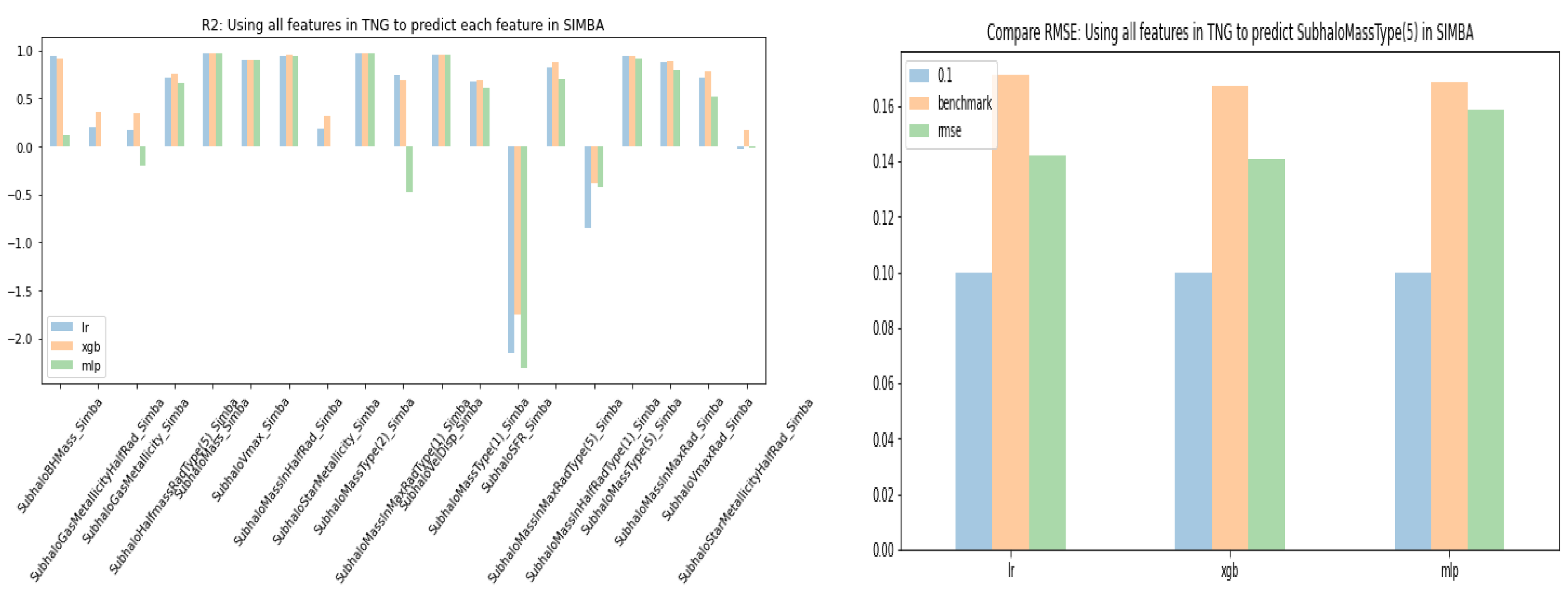

Figure 2. R square and RMSE in inter-simulation prediction model (TNG to SIMBA).

Particles Part

The data involving particles of each subhalo is organized in snapshots of simulation runs. The accessible data include 4 types of particles, which are Gas, Dark Matter, Stars & Wind and Black Hole, which are what we focused on. Initially, our task was to correctly align particles to their own subhalo by using the column SubhaloLenType as an offset indicator. Then, to verify the alignment result, we leveraged cartesian spatial coordinates to visualize the distribution of particles.

After confirming that particles are assigned to their corresponding group (Subhalo), we considered employing the particle's spatial information and relation by building a graph neural network, so that we can better understand and predict the simulations. Currently, we are able to achieve a good performance in the same subhalo with a simple network structure, and we are figuring out how to train a generalized network to all subhalos with different graph size and structure.

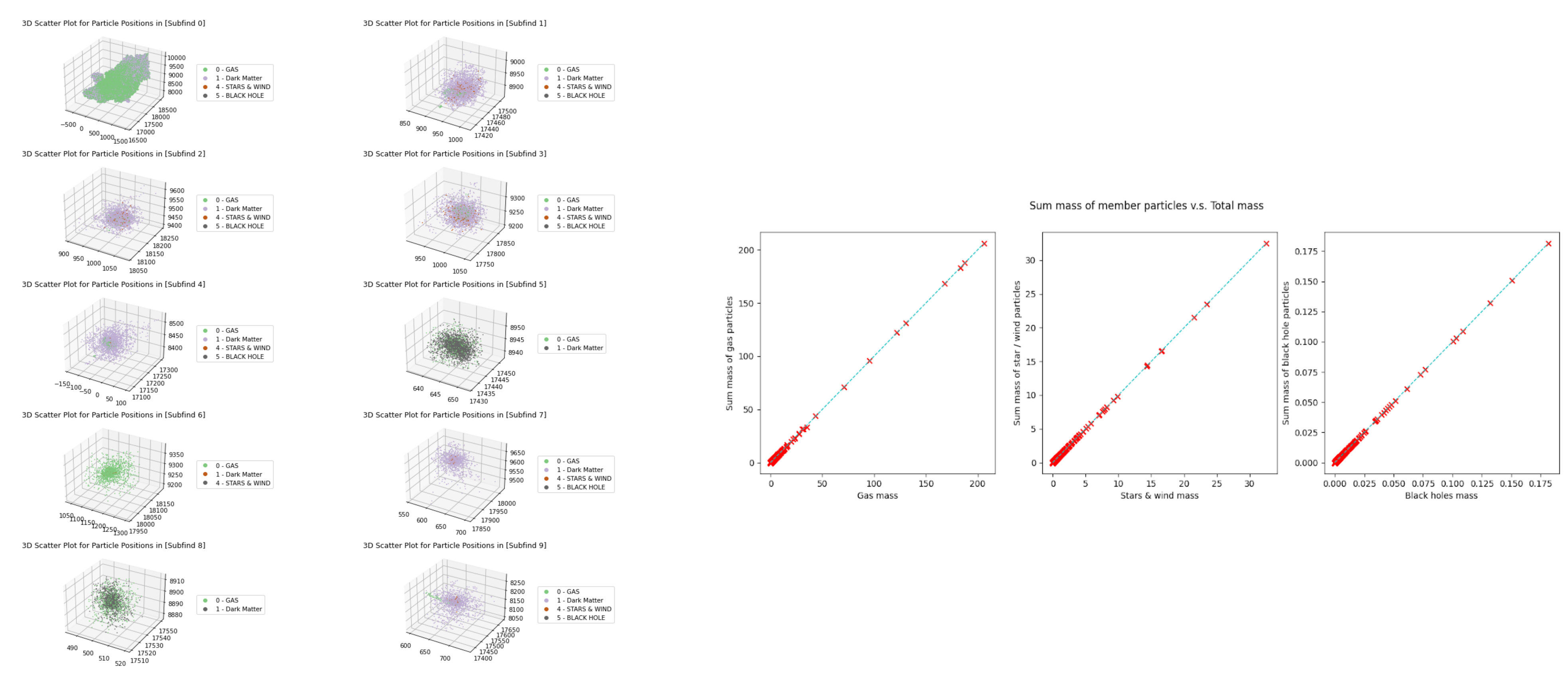

Figure 3. : 3D Scatter Plot of Particle and Sum Mass of Member Particles v.s. Total Mass.

Conclusion and Recommendations

We proposed a data-driven model to predict the results of Cosmo-hydrodynamical Simulations of Galaxy Formation. From 3D scatter plots, particles belonging to the same subhalo cluster in elliptical shape, which meet our expectations. Also the second scatter plot can showcase that particles are assigned correctly because the value of mass are matched.

Acknowledgments

We would like to extend our sincere thanks to the support from Professor Shy Genel and Yongseok Jo.

References

Dylan Nelson, Volker Springel, Annalisa Pillepich, Vicente Rodriguez Gomez, Paul Torrey, Shy Genel, Mark Vogelsberger, Ruediger Pakmor, Federico Marinacci, Rainer Weinberger, et al. The illustristng simulations: public data release. Computational Astrophysics and Cosmology, 6(1):1–29, 2019.