OPEN ACCESS

Predicting the UV Escape Fraction of the First Galaxies in the Renaissance Simulations with Machine Learning

Ben C. Sherwin[1,2], Snigdaa S. Sethuram[2], Corey Brummel-Smith[2], and John H. Wise[2]

Published November 2023 • © 2023. The Author(s). Published by the American Astronomical Society.

Research Notes of the AAS, Volume 7, Number 11

Citation Ben C. Sherwin *et al* 2023 *Res. Notes AAS* **7** 242

DOI 10.3847/2515-5172/ad0cc5

[1] University of Florida, Department of Physics, 2001 Museum Rd, Gainesville, FL 32611, USA

[2] Georgia Institute of Technology, School of Physics, Center for Relativistic Astrophysics, 837 State St NW, Atlanta, GA 30332, USA

Ben C. Sherwin https://orcid.org/0009-0006-6662-5056

Snigdaa S. Sethuram https://orcid.org/0000-0001-7908-3934

Corey Brummel-Smith https://orcid.org/0000-0001-6204-5181

John H. Wise https://orcid.org/0000-0003-1173-8847

1. Received October 2023
2. Revised October 2023
3. Accepted November 2023
4. Published November 2023

Galaxy formation; Reionization; High-redshift galaxies

What are corridors?

Abstract

Cosmic reionization is likely driven by UV starlight emanating from the first generations of galaxies. A galaxy's UV escape fraction, or the fraction of photons escaping from the galaxy, is useful to quantify its contribution to reionization. However, the UV escape fraction is notoriously difficult to predict due to local environment dependency and variability over time. Using data from the Renaissance Simulations, we attempt to make predictions about the impact of the first stars and galaxies on their environments. We present a time-independent classification model using a general artificial neural network architecture to predict the UV escape fraction given other galaxy properties—namely halo mass, stellar mass, redshift, star formation rate, lookback time, and gas fraction. We find our validation accuracy to be approximately 50%–65%, depending on the data set size from each zoom-in region of the Renaissance Simulations.

← **Previous** article in issue

Next article in issue →

1. Introduction

The Epoch of Reionization refers to an early period when the universe was in a cold, neutral state and was gradually ionized and heated from the first stars and galaxies. These first galaxies are at the edge of JWSTs detection capabilities and were previously unobservable. Hence, many simulations have been developed to understand how high-redshift galaxies evolve over cosmological timescales (Vogelsberger et al. 2019).

The Renaissance Simulations (O'Shea et al. 2015) are one such example: they are zoom-in cosmological simulations that model high-redshift galaxy formation with a resolution sufficient to capture their entire star formation history, starting with metal-free stars in minihalos of mass $\sim 10^6$ $M_\odot$. They were run with the adaptive mesh refinement code, ENZO (Bryan et al. 2014), which evolves cosmological volumes considering various physical processes, including metal-free and metal-enriched stars. Within these simulations, we examine three zoom-in regions in a $(40\ \mathrm{Mpc})^3$ comoving volume. The three selections consist of a high-density region (Rarepeak), an average-density region (Normal), and a low-density region (Void) that evolve to $z = (15, 12.5, 8)$, respectively.

2. Methods

We use ROCKSTAR and CONSISTENT-TREES (Behroozi et al. 2012, 2013) to find dark matter halos and construct their merger trees. For this initial study, we only consider the branches that follow the most massive progenitor in each merger tree. We then compute the total halo mass, gas fraction, star formation rate (SFR), stellar mass, lookback time, and distance to the nearest massive galaxy at each redshift for each halo with YT (Turk et al. 2010). The evolution of these halo/galaxy properties, along with their redshifts, are used as the inputs (features) for the machine learning model. The UV escape fractions, originally presented in Xu et al. (2016), are used as outputs (labels) to check the model's accuracy during validation. Due to an abundance of extremely low escape fractions, given the small size of these high-redshift galaxies, we balanced the data by filtering out halos with escape fractions below 10^{-5}, as their contributions to intergalactic medium ionization are negligible (Wise & Cen 2009). Our final data set contained 168 Rarepeak, 296 Normal, and 89 Void halos, corresponding to 47,040, 82,880, and 24,920 data points respectively when incorporating their evolution over time.

The model itself utilizes a time-independent general artificial neural network architecture that, when given the halo properties, predicts one of four groupings for an instantaneous escape fraction: <1%, 1%–10%, 10%–25%, and 25%–100%.

We trained a separate model for each region, using Cross-Entropy loss and the AdamW optimizer. To prevent overfitting in this relatively small data set, we cut off model training once test loss began to stagnate or increase. The most optimal hyperparameters, including number of hidden layers,

nodes per layer, dropout rates, learning rate, and weight decay for each model were determined using the OPTUNA package (Akiba et al. 2019). Full model details and specific hyperparameters are available on Zenodo:10.5281/zenodo.8415551 .

3. Results

Figure 1 depicts the results of the training process. The train and test losses are calculated during the training process, whereas the validation loss is calculated when the final model is used on unseen data. As shown in Figure 1, our loss curves exhibit favorable trends, generally exponentially decreasing with each epoch, indicating our models are working and valid. The Normal model has the best performance, with its loss bottoming out at approximately 0.95. The Void model has the poorest performance, with a noisy loss curve that does not have a fully converging test loss. This behavior persisted despite increasing the number of epochs and changing other hyperparameters during model testing. The Rarepeak and Normal loss curves are much smoother than those of the Void data set, most likely from the larger number of halos. We find that the classification models have accuracies of 56%, 63%, and 53% for the Rarepeak, Normal, and Void regions, respectively. We note that the accuracies are directly proportional to the number of halos in each catalog.

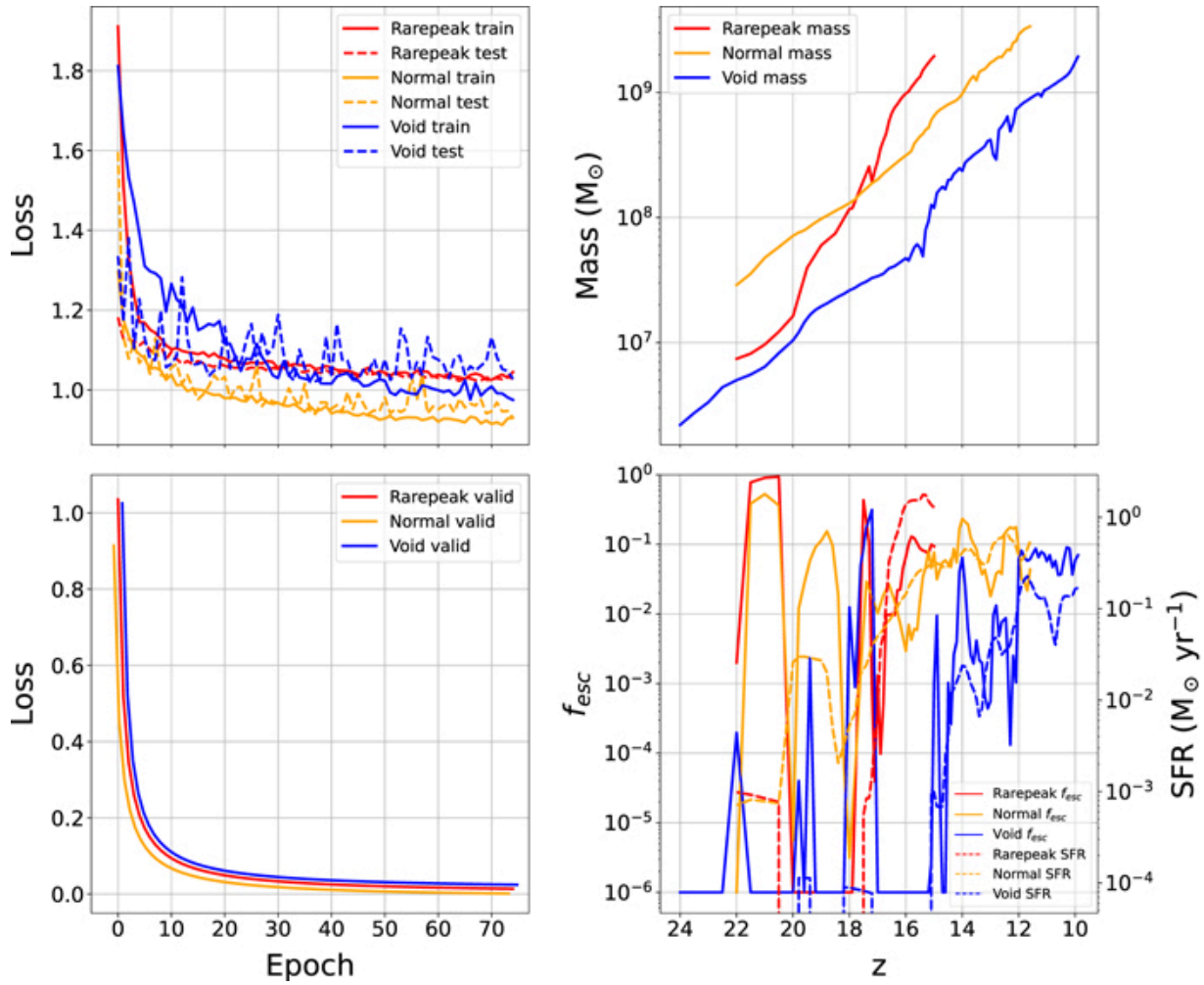

Figure 1. The upper left panel shows the training and testing loss curves and the lower left panel shows the validation curves. The upper right panel depicts the evolution of the mass of the most massive halo in each region and the lower right panel depicts the evolution of star formation rate and escape fraction for these same three halos. We note a delay between the peak in f_{esc} and SFR as discussed in Kimm & Cen (2014).

4. Discussion

The Renaissance Simulations provide us with a rich data set to investigate the role of the first galaxies in cosmic reionization, but the small number of halos with a non-negligible UV escape fraction led to relatively low accuracies. With additional data, either more simulations, features, or augmented data, we are confident that our model will improve. The highly variable nature of the UV

escape fraction and its unpredictability can also lead to low accuracies. We note that random guessing would lead to an accuracy of 25%, thus our models are performing at a significantly higher rate.

Further improvements to the model could include incorporating the half-mass (total and stellar) radius as another feature. We also expect that the best option for augmenting the data would be to interpolate data between discrete redshifts, providing smoother data curves.

In the future, we aim to create a time-dependent regression model that predicts the UV escape fraction at an arbitrary point in their evolution given a time series of halo/galaxy properties. The ideal model will likely have a recurrent neural network architecture that considers the full time series for each halo's properties. We anticipate its accuracy will be higher than the classification model because using time series as features in the model will innately capture the nature of galactic evolution.

Acknowledgments

B.C.S. acknowledges support from the the Georgia Tech Physics REU and the usage of Georgia Tech's PACE HPC resources. S.S.S. is supported by the NASA FINESST fellowship award 80NSSC22K1589. This work is supported by NSF grants OAC-1835213, AST-2108020, and DMR-2244423, and NASA grants 80NSSC20K0520 and 80NSSC21K1053.

IOP Publishing open access policy

Accessibility