Discovering Hidden Galaxy Populations in TNG100-1 With Unsupervised Autoencoder Learning

Bai, Isabella (School: North Carolina School of Science and Mathematics)

Galaxy classification often relies on human-defined categories, which can introduce observer bias, miss non-linear relationships, and fail to scale to new billion-galaxy datasets. This project uses an autoencoder, an unsupervised machine learning model, to discover galaxy populations in cosmological simulations. Inputting 5,000 galaxies from the IllustrisTNG TNG100-1 cosmological simulation, each described by 10 physical properties (including stellar mass, gas mass, etc.), the neural network compressed 10-dimensional data into a 3D latent space, while preserving 96% of the original information (MSE=0.037). The autoencoder's weak Pearson correlations ($|r| < 0.04$) between latent dimensions and input features indicate that the model captured the non-linear relationships in galaxy physics that linear methods cannot. Using reconstruction error, 250 anomalous galaxies were identified, and they were characterized by unusual feature combinations, not single extremes. From K-means clustering in the latent space, four distinct galaxy populations emerged: massive quenched (31%), dwarf quenched (51%), massive star-forming (5%), and dark matter-depleted compact systems (13%, n=639). These 639 galaxies with near-zero dark matter have a 12.8% prevalence, which is 6-13 times higher than observational estimates of 1-2%. Their ultra-compact and quenched properties align with tidal dwarf predictions, which suggests these galaxies are undercounted for in surveys. These results demonstrate the power of unsupervised deep learning in uncovering hidden structure and rare systems in large-scale cosmological datasets and offer a scalable approach in upcoming astronomical surveys.

Awards Won: