Automated Identification and Inference of Organic Molecular Structure and Relative Concentrations from Infrared Spectral Data

Jones, Cameron (School: The Harker School)

The discovery of complex organic molecules in space is critical to the understanding of the reaction pathways leading to biomolecules and the origins of life. Existing techniques for the analysis of astronomical spectra require knowledgeable researchers and often struggle to identify and differentiate complex spectral signatures, such as those of polycyclic aromatic hydrocarbons (PAHs). My project applies machine learning (convolutional neural networks) to the problem of identifying complex organic molecules in IR spectroscopic data and proposes a novel method for creating synthetic training data to tune models to specific astronomical environments. My project created: a) models to identify organic molecules from empirical IR spectroscopic data when trained on the approximate theoretical counterparts from NASA's PAHdb v2 and v3 databases and b), models to identify molecular compositions from spectra of random theoretical molecule mixtures with realistic noise. My principal findings are: a) network models trained on theoretical spectra can accurately identify empirical molecules with ~73% accuracy, and b) models trained on random mixtures of 3,139 theoretical PAH spectra can identify molecular concentrations with weight vector correlations of ~85% and can correctly identify the largest constituent ~67% of the time. In all cases, my models (the best being ResNet5 with ~200M parameters) dramatically outperform standard linear models. My convolutional network models can recognize complex spectral patterns and generalize across datasets with realistic noise. These models can significantly increase the scale and efficiency of analyzing astronomical IR spectra data and improve our understanding of the distribution of complex organic molecules in the universe.

Awards Won:
Fourth Award of $500