Investigating the molecular richness in star-forming gas with machine learning

○Nina KESSLER[1] Patricio SANHUEZA[1] Timea CSENGERI[2] David CORNU[3] Sylvain BONTEMPS[2] Laure BOUSCASSE[4]
(1.Department of Astronomy, The University of Tokyo 2.Laboratoire d'Astrophysique de Bordeaux, Université de Bordeaux 3.LUX, Observatoire de Paris, PSL 4.IRAM)

Protostars, and more specifically their hot gaseous envelopes (hot cores, hot corinos), have been observed to be molecularly rich, particularly during their early formation stages. Interferometers (e.g. ALMA, NOEMA) and single dish telescopes (e.g. IRAM 30m, Yebes 40m) provide high-resolution spectroscopic data that can help us understand the origin of the chemical diversity that may develop in these regions, as well as how the molecular composition evolves in an environment influenced by one or multiple protostars. However, analyzing the wealth of data from star-forming regions requires iterative fitting of spectroscopic models, a process that is often time-consuming for molecularly rich objects. Besides, the sheer volume of broadband data from state-of-the-art telescopes further complicates this task, limiting our ability to perform large-scale statistical and systematic studies of these environments. To address this challenge, we have developed novel machine learning approaches to automate the analysis of large spectral data cubes. In this talk, I will present a method, described in Kessler et al. 2025, designed to accelerate the detection of molecules around protostars by leveraging their rotational spectra in millimeter data. This includes the creation of a fully synthetic training dataset, as well as the elaboration of a convolutional neural network (CNN) architecture to optimize the ability of the model to identify molecular signatures within rich and complex spectra. The resulting CNN model demonstrates reliable performance on synthetic data and generalizes effectively to millimeter observations, delivering a list of detected molecules for hot core and hot corino spectra within seconds. To enhance interpretability, we analyzed the CNN model's decision-making process, revealing the way it uses specific spectral features to confirm molecular detections.