Machine learning for guided analysis of laboratory spectra to study life-precursor molecules in space

Co-directors: Adam WALTERS (Pr), MICMAC group, Institut de Recherche en Astrophysique et Planétologie, Université de Toulouse / CNRS / CNES. Leszek SIWIK (Associate Professor in Computer Science), Faculty of Space Technologies, AGH University of Krakow.

By leveraging artificial intelligence to streamline and accelerate the spectral analysis of laboratory spectra, this doctoral project aims to enhance our understanding of the molecular complexity present in the interstellar medium (ISM). This could subsequently provide critical insights into astrobiology, the origins of prebiotic molecules and their potential role in the emergence of life on Earth.

The ISM is a vast and complex environment filled with gas and dust; radioastronomy has played a pivotal role in identifying a diverse array of complex organic molecules in this medium many of which are crucial for understanding the origins of life. The radio-astronomical spectra of star-forming regions, exhibit a rich tapestry of spectral lines, with thousands of overlapping lines, some of which have been identified while others remain enigmatic. Key molecules identified in the ISM include: (1) Formaldehyde (H_2CO), one of the simplest organic molecules, that plays a role in the formation of amino acids; (2) Methanol (CH_3OH) that is a building block for more complex organic molecules. Its presence suggests that the chemical processes leading to life's building blocks are occurring in space; (3) A variety of Complex Organic Molecules (COM) including aldehydes and alcohols, that are essential for understanding the chemical pathways that could lead to the formation of life; (4); Glycolaldehyde ($HOCH_2CHO$) the simplest sugar-related molecule; (5) Glycolamide ($HOCH_2CONH_2$) a chemical cousin (isomer) to the simplest amino acid glycine. The molecules found in the ISM can contribute to the chemical inventory of forming planets. Glycine itself has been identified in comets as well as ribose and other bio-essential sugars including arabinose and xylose. This leads to the hypothesis that molecules created in the ISM and incorporated or later formed in comets could have delivered essential building blocks of life to Earth during its early history. If comets can deliver organic compounds to planets, similar processes may occur on exoplanets, increasing the likelihood of finding life elsewhere in the cosmos.

To accurately characterize these spectra and identify the molecular species present, it is essential to first measure the spectra of candidate molecules in the laboratory. However, direct application of these laboratory measurements to the ISM is complicated by the vastly different conditions present in space, including temperature, pressure, and density. To bridge this gap, a quantum mechanical analysis is necessary to derive a set of molecular parameters that can predict not only the frequencies of spectral lines but also their intensities in the ISM.

The spectra of prebiotic molecules are often highly intricate due to several factors, including molecular asymmetry, the presence of numerous low-lying vibrational states, and the interactions between torsional, vibrational, and rotational motion. While modern laboratory techniques allow for rapid measurement of spectra, the subsequent quantum-mechanical analysis is labor-intensive and requires acquired expertise. The process also involves repetitive tasks that could be significantly expedited through the application of machine learning trained in the expertise required.

The project will focus on automating the spectral analysis process by applying sequence models such as LSTMs or Transformers to identify patterns in spectral data, classify spectral lines, and predict their shifts relative to theoretical values. Learning-based optimization methods (e.g., Bayesian optimization or reinforcement learning) can support iterative refinement of molecular parameters by minimizing RMS deviations. Additionally, unsupervised learning techniques (e.g., clustering, autoencoders) may assist in recognizing structures within complex spectra, especially in the presence of overlapping lines and noise.
The process will be verified using a molecule like formic anhydride ($(HCO)_2O$)) for which human expert analysis has already begun, thus providing a benchmark to guide and evaluate the procedure before directing the analysis to new molecules of astrophysical interest.
The thesis will start by an acquisition of the expertise in spectral analysis at Toulouse. It will be continued by the application of machine learning in Krakow, Poland. Then finalized by the application of the technique for new research.

An overview of the Spectral Analysis Process is given below:

1. **Measure the Spectra in the Laboratory:** Conduct precise measurements of the molecular spectra under controlled laboratory conditions.
2. **Start with the Vibrational Ground State:** Begin the analysis by focusing on the vibrational ground state of the molecule, which serves as the baseline for further analysis.
3. **Obtain Initial Molecular Parameters:** Estimate the molecular parameters (using ab-initio calculations for example) and use them to predict the rotational spectrum.
4. **Identify the best set of transitions (lines) for the present stage of the analysis:** Select a set of transitions (lines) for initial comparisons between measured and theoretical spectra. The choice depends on the present accuracy of the predictions, the tolerance for the line to be displaced in frequency from the prediction and the strength of the lines to be observed.
5. **Predict Frequencies of Selected Lines:** Calculate the frequencies of chosen spectral lines, labeling them with their corresponding quantum numbers for clarity.
6. **Identify Lines in Measured Spectra:** Analyze the measured spectra to determine the center frequencies of identified lines, while estimating measurement uncertainties. Recognize that lines may be displaced from predictions due to various factors, including accuracy of the prediction, noise and overlapping spectra (vibrational states and contaminant molecules). Patterns in the spectra may provide better identification than absolute frequency values.
7. **Select Lines for Analysis:** Choose the most suitable lines for detailed analysis. Refine the list of molecular parameters based on this selection by means of a least squares analysis, making new predictions and comparing them with the measurements. This iterative process may require multiple refinements to achieve the best quality fit (lowest RMS deviation between predictions and measurements).
8. **Incorporate Lines of higher quantum number, energy and of lower intensity:** Progressively include additional lines into the analysis. Return to step 4 and repeat until all relevant lines have been identified within established quality criteria.

9. **Progress to Higher Vibrational States:** Once the analysis of the vibrational state is complete, move to the next lowest vibrational state and repeat the process starting from step 3. Continue as long as the vibrational state is expected to be populated in the ISM.

The project will focus on automating steps 4 to 9 of the spectral analysis process, initially using a molecule for which human expert analysis has already been completed (for example, formic anhydride ($(HCO)_2O$)). This will provide a benchmark to guide and evaluate the procedure. The machine-learning analysis will then be applied to new molecules relevant to the study of the origins of life in space.

(In the diagram below only a small portion of the spectrum is shown, typically hundreds of lines are to be analysed)