Machine learning spectral assignments

The ExoMol project (www.exomol.com) provides laboratory data (largely computed) on the spectroscopic properties of molecules for studies of exoplanets and other hot astronomical atmospheres. Spectroscopic studies contain a wealth of information on the species being studied but more importantly on the (astronomical) environment in which they are being observed. However to extract this information it is necessary to have fully assigned spectra where assignment means that each transition is uniquely linked with initial and final states of the molecule. Assigning complicated observed spectra can be a long and tedious process. For example, our high profile assignment of the spectrum of water vapour in the Sun (actually in sunspots)[1] only actually assigned about 20% of the actual lines observed and no further assignments have been made in the subsequent 25+ years.

The aim of the project will be to develop ML algorithms to ideally assign spectra or perhaps aid the assignment of spectra. We have a wealth of datasets both for suitable training purposes (the entire ExoMol database for a start!) and that would benefit from this methodology. For example, besides the sunspot example (and water is not the only molecule which has a partially assigned solar spectrum), (a) our recently completed analysis of laboratory methane spectra extracted about 80 000 transitions from 101 separate papers but found transitions lacking assignments in about 40 papers; (b) the situation is similar for ammonia where a well studied absorption feature in Jupiter red region of the visible lies in a region where there are laboratory spectra but none of them are assigned. Our calculations are not accurate enough to bridge this gap (indeed the current ExoMolHD project is all about cannibalizing assigned experimental spectra to improve the accuracy of our line lists) . There are a wealth of other astronomically important spectral data that such a procedure could be applied to.

I am not aware of any attempts to use ML to make systematic line assignments although Meerts and co-workers very successfully pioneered the use of an evolutionary model for this process[2] but their algorithm relied on the successful characterisation of the spectrum using effective Hamiltonian models. These models only work well at low energies/temperatures and are not suitable for the spectra we wish to analyse. We will replace this with a ML algorithm based on the Ritz principle that each level is quantised so that transitions occur between fixed points and therefore spectra form a graph (with energy levels as nodes and transitions as edges). The plan will still be to use an evolutionary (or bootstrap) model where starting from partial assignments will allow further assignments to be made on the basis of the ML algorithm allowing a new a new ML model to be constructed and further assignments to be made. As assigned levels have to fit into the graph, successful assignments should be confirmed by multiple transitions linking to the each level (a process called combination differences in the spectroscopic community). Initial work will used know assigned spectral datasets to explore different ML algorithms (and feature sets) to find the most effective methodology.

ML methods are significantly underused for spectroscopic analysis. Indeed at the recent main European conference of High Resolution Molecular Spectroscopy, my group was the only one who presented an ML based study!

Jonathan Tennyson Oct 2023

A final comment: I currently have a DIS CDT student (my first PhD student on this scheme) but she is largely paid for by my ERC which is now too near to its end to support any further students.

1 O.L. Polyansky, N.F. Zobov, S. Viti, J. Tennyson, P.F. Bernath and L. Wallace, Water in the Sun: line assignments based on variational calculations, *Science*, **277**, 346-349 (1997).

2 J. van Wijngaarden, D. Desmond, W.L Meerts, Analysis of high resolution FTIR spectra from synchrotron sources using evolutionary algorithms, *J. Molecular Spectroscopy*, **315**, 107-113 (2015)