Contribution ID: **9** Type: **Oral Presentation**

Mass Spectra Prediction and Analysis: Machine Learning and Quantum Computing Perspectives

Wednesday, 19 March 2025 11:00 (30 minutes)

Mass spectrometry (MS) is a compound identification technique used frequently in life- and environmenta sciences. The specific setup of mass spectrometry on electron ionization, coupled with gass chromatrography (GC-EI-MS) is appealing due to its relative simplicity and stability of the setup, while it is challenging computationally – the acquired data are strictly "flat", with no hierarchical structure which would reflect the chemical structure of the analyzed compounds provided by other experimental techniques (MS^n).

The community recognizes the importance of computational prediction of mass spectra of given compounds as well as elucidation of molecular formulae out of measured spectra. Traditional library-search approaches are limited by a quantitative gap. The number of "small" organic molecules is estimated up to 10^{60}, out of which approx. 1 billion is confirmed to exist (ZINC database). On the other hand, the state-of-the art spectral libraries (NIST, Willey) contain only about 500,000 records.

We introduce SpecTUS, our 354M-parameters transoformer-based ML model, which takes the mass spectrum on input, and produces molecular structural formulae (SMILES strings) on its output. SpecTUS was pre-trained on 2×4.7 million synthetic mass spectra (generated by state-of-the-art models NEIMS and RASSP), and fine tuned on 232,000 spectra of our training subset of the NIST spectral database. We evaluate the model on a~held-out testing subset of NIST (28,000 spectra), as well as spectra from other independent databases (1,640 SWDRUG, 5,015 MONA). When a single candidate is retrieved, the exact solution is returned in 40% cases, and the average precision (Morgan-Tanimoto similarity) is 0.66. With retrieving 10 candidates, the exact solution among them is in 62% cases, and the precision increases to 0.79. We also carried experiments of similarity comparison to legacy spectral database search methods to demonstrate that the model is able to generalize (not memorizing the trainig set only). There is no competing solution for the GC-EI-MS setup but our results superseed even the state-of-the-art MS^n approaches, which work with more structured information on input.

The other extremum on the scale of MS computations is the accurate prediction of mass spectra. These methods simulate the actual process of molecule fragmentation on the electron impact, using ab-initio energies calculated by solving time independend Schroedinger equation. Semi-empirical quantum-chemical methods are not sufficiently accurate in this case, and DFT-like methods easily become computationally unfeasible.

On the other hand, solving the Schroedinger equation is one of expected killer applications of quantum computers. We demonstrate integration of QCxMS software package (ab-initio spectra simulator) with Qiskit-based implementation of the energy calculation by Variational Quantum Eigensolver, a~method to approximate the ground state energy suitable for the current noisy quantum hardware. So far we carried simulated validation experiments only but we are heading to the use of the actual quantum hardware.

Primary authors: Dr KŘENEK, Aleš (Masaryk University); HÁJEK, Adam (Masaryk University); WAGNER, Michael (IBM)

Co-authors: PRICE, Elliott (Masaryk University); HECHT, Helge (Masaryk University); ROJAS, Wudmir (Masaryk University)

Presenter: Dr KŘENEK, Aleš (Masaryk University)

Session Classification: Health & Life Sciences

Track Classification: Track 2: Health & Life Sciences Applications