Mass spectrum to structure: Calculation of compound structural fingerprints based on high-resolution tandem mass spectrometry data

Project: ExpectMINE: Elucidation of toxicity components in wastewater by integration of HRMS analysis and data mining

Eawag: Kasia Arturi, Juliane Hollender
SDSC, ETH Zürich: Eliza Harris, Lilian Gasser, Fernando Perez-Cruz

With the rapid development of high-resolution mass spectrometry and non-targeted screening workflows, it is now possible to detect and process tens of thousands of unknown and potentially harmful environmental pollutants in aquatic samples. However, since even the most sophisticated workflows require manual verification for tentative compound identification and quantification, compounds must be prioritized based on an initial estimate of potential toxicity. Such prioritization strategies are often based on peak areas or frequencies of occurrence in measurement spectra, which lack the toxicological relevance so critical in the context of environmental pollution. To append toxicological relevance to HR-MS/MS analysis, machine learning methods (ML) are being developed to predict the toxicity of unknown compounds directly based on their MS/MS spectra, and thus prioritize investigation of compounds with high potential toxic effect.

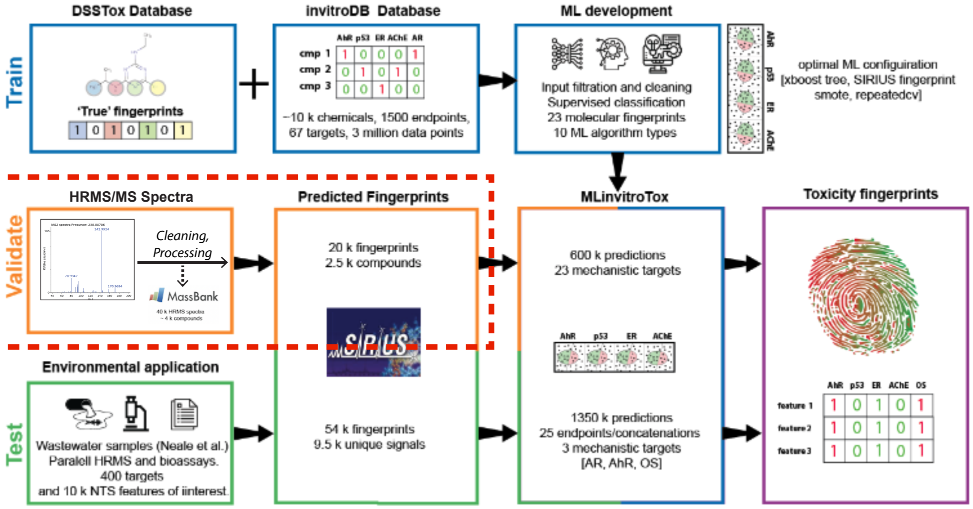

Figure 1: The ExpectMINE project involves several stages: Training a machine learning model to relate compound structure to toxicity (blue), pipeline development to calculate structural fingerprints from HR-MS/MS spectra, and consequently estimate toxicity of unknown compounds (orange), and application and testing of the pipeline on non-targeted environmental analyses to prioritise potentially toxic compounds (green). The focus of this MSc project is highlighted in red.

This MSc project focusses on the calculation of structural fingerprints based on HRMS/MS spectra (Figure 1). Mass spectra first need to be cleaned and preprocessed which can be done using a variety of commercial and open source programs (eg. mzMine, MS-DIAL, Compound Discoverer, enviMass, SLAW, patRoon). Structural fingerprints for processed spectra can then be calculated using the 'Sirius' spectral annotation software (https://bio.informatik.uni-jena.de/software/sirius/); moreover, processed spectra can be contributed to the MassBank depository. The calculated structural fingerprints are then passed to the machine learning algorithm developed within ExpectMINE in parallel, in order to construct an estimate of potential toxicity. The focus of this project will be to develop an efficient, automated pipeline to clean and preprocess HRMS/MS data to calculate structural fingerprints, in particular:

- to consider different HRMS/MS data preprocessing methods and packages, and develop and optimize a pipeline for automated, fast, and robust processing,
- to apply SIRIUS and other fingerprint estimate methods in an automated pipeline with streamlined input and outputs,
- to consider how different data processing methods and fingerprint formats could affect toxicity prediction,
- to implement the pipeline as an efficient workflow using SDSC's Renku platform.

This MSc project will form a key part of the ExpectMINE project (Figure 1), which aims to develop a full pipeline to estimate compound toxicity based on non-targed HRMS/MS screening of environmental samples.

Additional information

- **What will you learn?**
  - Construction of efficient and robust data analysis workflows
  - Workflow management tools, particularly within Renku
  - Metholodologies for HRMS/MS data processing and molecular structure representations
- **Requirements:**
  - Excellent knowledge of Python and/or R
  - Experience with multistage analysis pipelines is an advantage
  - A strong interest in chemistry and bioinformatics
- **Supervisors and collaborators:**
  - Eawag: Kasia Arturi, Juliane Hollender
  - ETHZ: Eliza Harris, Lilian Gasser, Fernando Perez-Cruz
- Please contact Eliza Harris (eliza.harris@sdsc.ethz.ch) or Kasia Arturi (kasia.arturi@eawag.ch) for further information