Relating compound toxicity to molecular structure using machine learning

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.

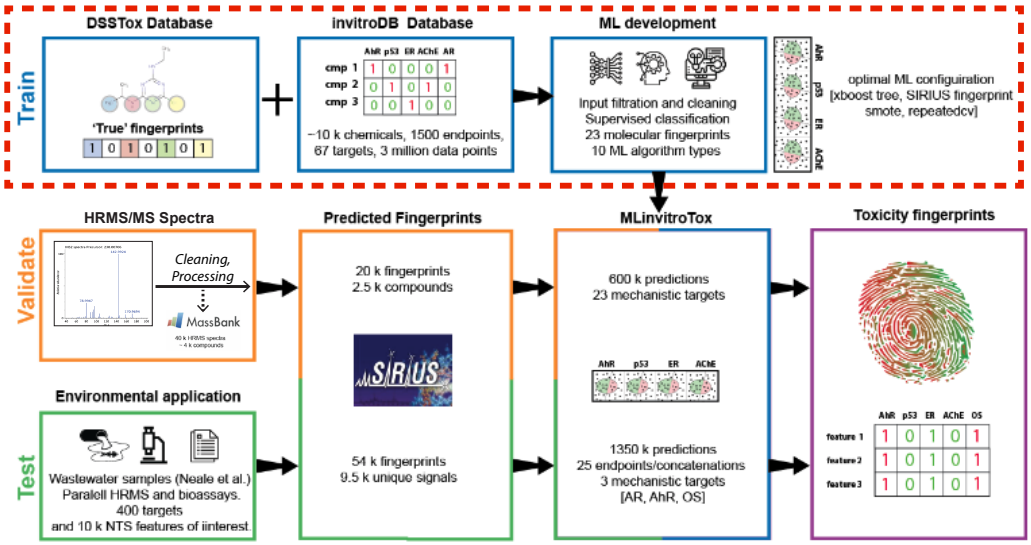

Figure 1: The ExpectMINE project involves several stages: Training a machine learning model to relate compound structure to toxicity (blue), pipeline development to calculate structural fingerprints from HR-MS/MS spectra, and consequently estimate toxicity of unknown compounds (orange), and application 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 centers on the relationship between compound structure and toxicity, which is a key element in the prediction of toxicity during non-targeted screening (Figure 1). Compound toxicities are taken from the 'invitroDB' database and used to construct a *toxicity fingerprint* representing toxicity according to different mechanistic endpoints. Compound structure is represented using a *molecular fingerprint* developed within the 'Sirius' framework (https://bio.informatik.uni-jena.de/software/sirius/), where fingerprint digits relate to different chemical functional groups and substructures. The focus of this project will be to investigate different models to predict toxicity fingerprints based on structural information, in particular:

- to apply and compare different machine learning models to predict toxicity fingerprints based on 'Sirius' structural fingerprints,
- to consider which chemical groups and substructures are relevant for the prediction of different toxicity endpoints and mechanistic targets, and relate these to known toxicity mechanisms,
- to develop a streamlined workflow for structural fingerprint calculation, toxicity fingerprint estimate, and model evalulation.

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?**
  - Molecular structure representation for machine learning
  - Application, optimization and comparison of different classification and regression machine learning approaches
  - Application of data science approaches for bioinformatics, particularly pipeline construction and cloud computing using both R and Python
- **Requirements:**
  - Good knowledge of Python and/or R
  - Experience with classification and regression approaches 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