MS2Tox: Assessing Toxicity of unidentified chemicals through MS^2 spectra embeddings

Yvonne Kreutzer[a], Anneli Kruve[a,b]

[a]Department of Chemistry, Stockholm University, Svante Arrhenius väg 16, 106 91 Stockholm, Sweden, [b]Department of Environmental Science, Stockholm University, Svante Arrhenius väg 8, 106 91 Stockholm, Sweden

Contact: yvonne.kreutzer@su.se

Machine learning models for **offline prioritization** of **unidentified** potential **endocrine active features** improves by using fragmentation tree based MS^2 representations

Background and Aim

Discovering endocrine disruptors (EDs) in complex mixtures is challenging as <10% of the chemicals are structurally annotated in non-target screenings [1].

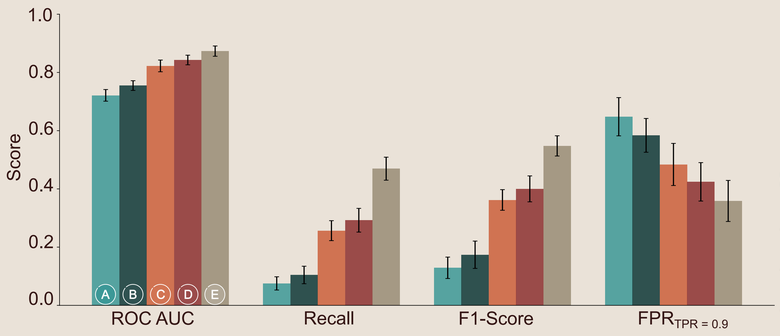
Machine learning models, from the MS2Tox toolbox, can predict toxicity from MS^2 spectra, [2,3] but are dependent on SIRIUS+CSI:FingerID fingerprints as intermediate.

This study assesses various MS^2 spectra processing methods for pinpointing EDs with a special focus on the Aryl hydrocarbon receptor (AhR) from MS^2 data.

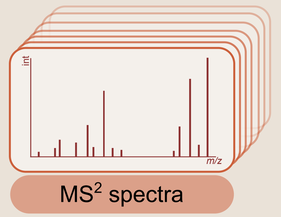
Results and Discussion

XGBoost repeated 5-fold cross-validation

SIRIUS+CSI:FingerID [5] (v.6) fingerprints obtained best performance (C+D). A reference model was trained on absolute fingerprints obtained from SMILES of the same chemicals.



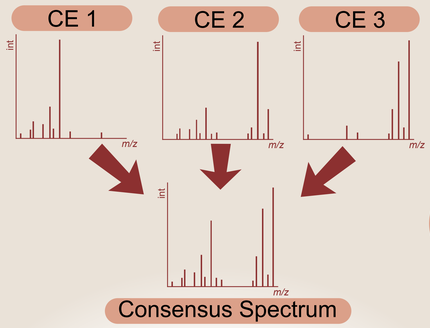
Methods



MS^2 processing

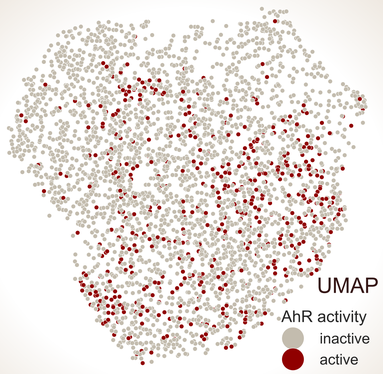
Database experimental spectra from MassBank, MoNA, GNPS, NIST23, MSnLib, Spectaverse were gathered and cleaned:

- adduct filtering --> only keep $[M+H]^+$
- precursor *m/z* max 1,000 Da
- remove peaks outside 10.0 to 1,000.0 Da
- remove noise (peaks lower 0.001 real. int.)
- remove spectra > 1,000 peaks
- remove spectra with < 4 remaining high peaks
- intensity amplitude (max int. /min int. <= 5)



Same InChIkey14
different Collision Energy

3525 Consensus Spectra
of Unique InChIKey14



Chemical Space of Training Data

3525 Consensus spectra with available AhR activity information from the Tox21 Data Challenge were obtained after processing and cleaning.
UMAP representation of DreaMS [4] embeddings.

Generating Machine Learning Input

- A: DreaMS embeddings
- B: DreaMS embeddings + XlogP
- C: SIRIUS 1st ranked formula
- D: SIRIUS True InChIKey14 matched
- E: Baseline model from structural fingerprints

Future Outlook

- Expanding model training on other MS^2 representations
- Including more relevant endpoints from the Tox21 Data Challenge
- Real life evaluation of model by applying prioritization on wastewater samples

The authors declare no compeing financial interest.
The project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement No. 101124488, LearningStructurE: Machine Learning and Mass Spectrometry for Structural Elucidation of Novel Toxic Chemicals.

[1] Papazian, S.; et al. Commun. Earth Environ. 2022, 3 (1), 1–14.
[2] Rahu, I.; et.al. J. Chem. Inf. Model. 2024, 64 (8), 3093–3104.
[3] Peets, P.; et al. Environ. Sci. Technol. 2022, 56 (22), 15508–15517.
[4] Bushuiev, R.; et al. Nat. Biotechnol. 2025, 1–11.
[5] Dührkop, K.; et al. Nat. Methods. 2019, 16 (4), 299–302.

Kruve Lab

Stockholm University

Vetenskapsrådet VR
Swedish Research Council 3R2022-01353

Funded by the European Union

European Research Council
Established by the European Commission