# Predicting the biochemical activities of UNIDENTIFIED CHEMICALS from MS² SPECTRA to pinpoint potential TOXIC AGENTS


**Ida Rahu**[a,b], Meelis Kull[a] and Anneli Kruve[b,c]

ida.rahu@mmk.su.se

a) Institute of Computer Science, University of Tartu, Narva mnt 18, 51009, Tartu, Estonia;
b) Department of Materials and Environmental Chemistry, Stockholm University, Svante Arrhenius Väg 16, SE-106 91 Stockholm, Sweden;
c) Department of Environmental Science, Stockholm University, Svante Arrhenius Väg 16, SE-106 91 Stockholm, Sweden


University of Tartu 1632 · Stockholm University · Kruve Lab

## BACKGROUND

1 in 6 premature deaths worldwide is reported to be caused by pollution.[1]

Endocrine disruptors (EDs) interfere with hormone action and are now a major focus in global risk assessment and management strategies.[2]

Nontarget LC/ESI/HRMS enables the simultaneous detection of numerous chemicals, but their identification remains limited (<5%), leaving gaps in toxicity assessment.[3-5]

The molecule's toxicity is associated with specific structural patterns,[6] which can be extracted as molecular fingerprint features from MS² spectra using SIRIUS+CSI:FingerID.

**We investigated whether these features could be used to predict the biochemical activity of unidentified chemicals to flag those warranting further testing due to potential harmful effects.**

## WORKFLOW

**Tox21 10K DATASET**

| # | CHEMICAL | 12 TOXICITY ASSAYS | | | | |
|---|---|---|---|---|---|---|
| 1 | BrC(Br)Br | 0 | 1 | NaN | ... | 0 |
| ... | ... | ... | ... | ... | ... | ... |
| 8043 | Nc1cc(Cl)ccc1O | 1 | NaN | 0 | ... | 0 |

1 - active; 0 - non-active

**TRAINING DATA**

SMILES STRINGS

Clc1cc2c(cc1Cl)Oc1cc(Cl)c(Cl)cc1O2

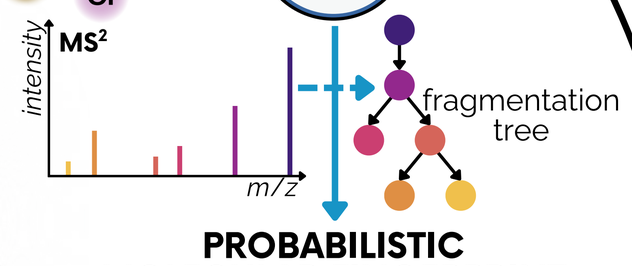
**BINARY MOLECULAR FINGERPRINT**

| | | | | | | |
|---|---|---|---|---|---|---|
| 0 | 0 | 1 | ... | 1 | 0 | 0 |

TRAINING → single- and multi-output models

**REAL-LIFE TEST DATA**

MS² SPECTRA




**PROBABILISTIC MOLECULAR FINGERPRINT**

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.14 | 0.45 | 0.99 | ... | 0.71 | 0.27 | 0.08 |

EVALUATION

SELECTING → CUSTOM METRIC · MONTE CARLO SAMPLING

## FIELD OF APPLICATION

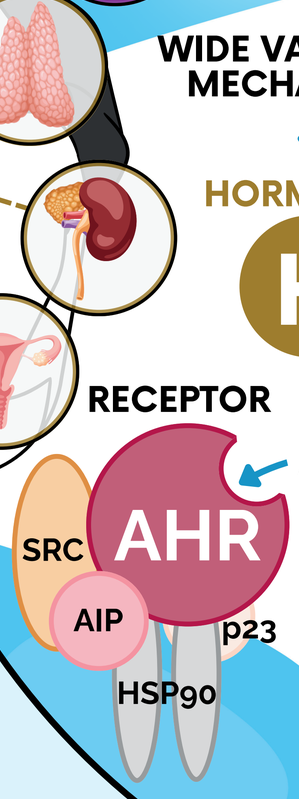

WIDE VARIETY OF MECHANISMS

HORMONES

H

RECEPTOR

AGONIST



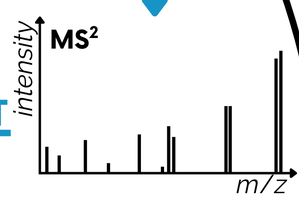


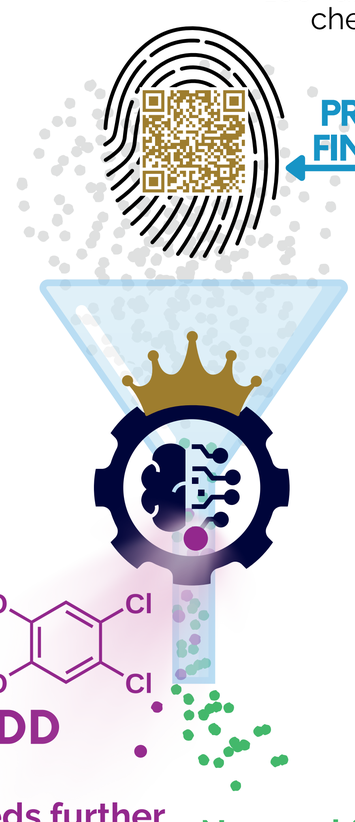
Needs further *in vivo* and/or *in vitro* testing

**REAL-WORLD SAMPLES**

100+ unidentified chemicals

LC/HRMS → PREDICTING FINGERPRINT



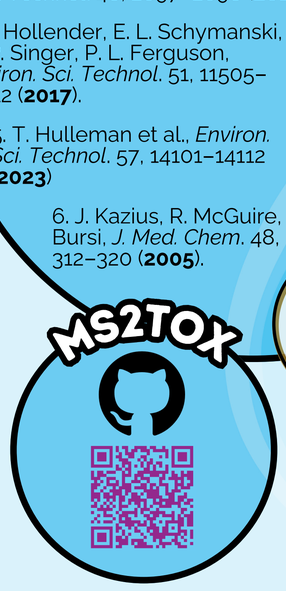
rapid
no animal testing

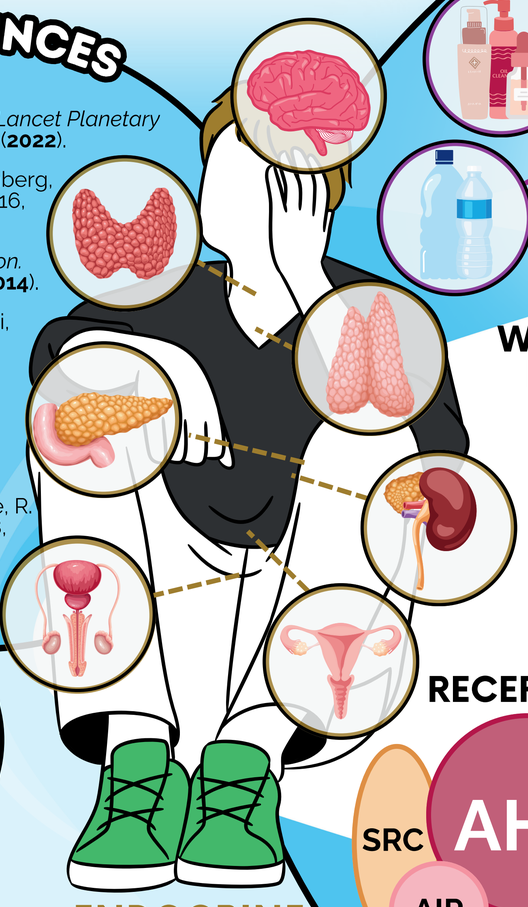
No need for further testing

## REFERENCES

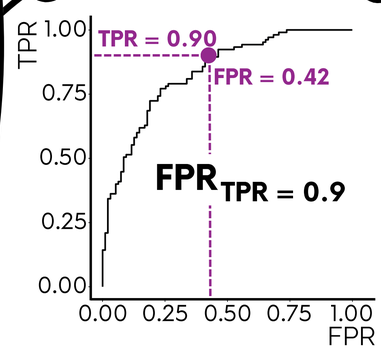

1. R. Fuller et al., *The Lancet Planetary Health*. 6, e535–e547 (**2022**).
2. A. Beronius, L. N. Vandenberg, *Rev. Endocr. Metab. Disord.* 16, 273–287 (**2015**).
3. E. L. Schymanski et al., *Environ. Sci. Technol.* 48, 2097–2098 (**2014**).
4. J. Hollender, E. L. Schymanski, H. P. Singer, P. L. Ferguson, *Environ. Sci. Technol.* 51, 11505–11512 (**2017**).
5. T. Hulleman et al., *Environ. Sci. Technol.* 57, 14101–14112 (**2023**)
6. J. Kazius, R. McGuire, R. Bursi, *J. Med. Chem.* 48, 312–320 (**2005**).


MS2TOX






## CUSTOM METRIC




**The metric aims for high recall to detect the majority of active chemicals while minimizing the workload associated with misclassified ones.**

## MONTE CARLO SAMPLING

Monte Carlo sampling was employed to mitigate discrepancies arising from using probabilistic fingerprint features derived from HRMS data in models trained on binary features.

**PROBABILISTIC MOLECULAR FINGERPRINT**

1 0 p=0.14 p=0.86 · 1 0 p=0.71 p=0.29

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.14 | ... | ... | ... | 0.71 | ... | ... |

MONTE CARLO SAMPLING

**BINARY MOLECULAR FINGERPRINT** (N = 10,000)

| | | | | | | |
|---|---|---|---|---|---|---|
| 0 | ... | ... | ... | 1 | ... | ... |
| 0 | ... | ... | ... | 0 | ... | ... |
| ... | ... | ... | ... | ... | ... | ... |
| 0 | ... | ... | ... | 1 | ... | ... |

INPUTS TO MODELS → **PREDICTIONS** 0.56, 0.12, ..., 0.84

$$\sum_{1}^{N} \frac{\text{prediction}_i}{N}$$

**FINAL PREDICTION 0.77**

Up to **20% of chemicals** showed **differences in activity predictions** compared to the naive 0.5 threshold method.

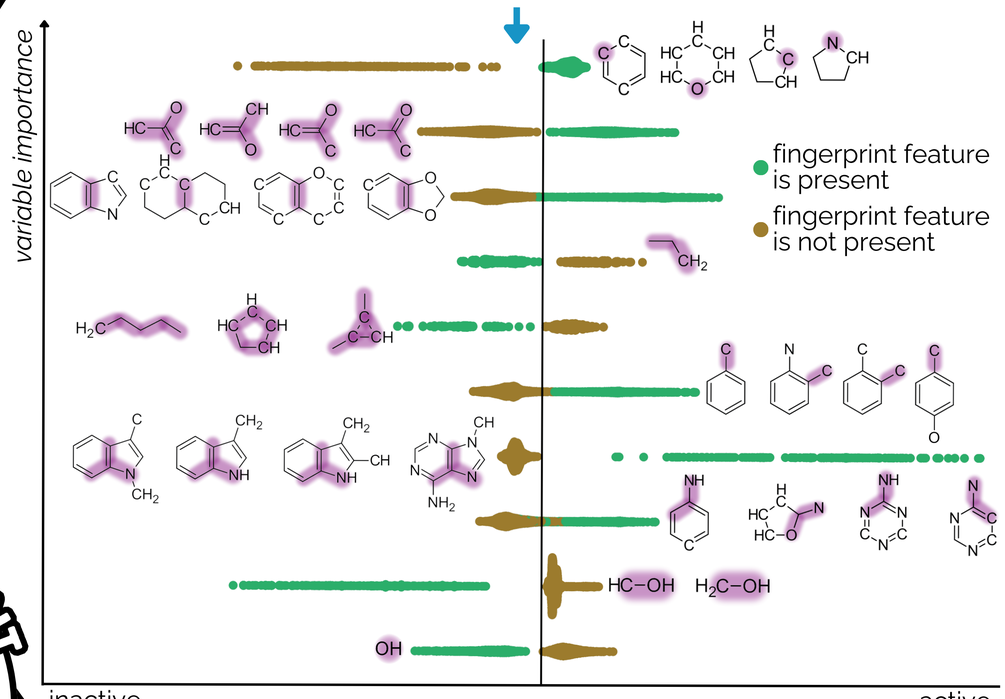
## MODELS' PERFORMANCE

Depending on the bioassay, the lowest $FPR_{TPR = 0.9}$ ranged from **0.251 (sr.mmp)** to **0.824 (nr.ar)**, consistent with the trends observed in the Tox21 Data Challenge, implying a **potential reduction of up to 75% in the post-processing workload** for nontarget HRMS.

## INTERPRETABILITY

nr.ahr

SHAP ANALYSIS




**Models are able to pinpoint structural patterns linked to the modes of action of active chemicals.**


## ACKNOWLEDGEMENT

This research received partial funding from the Estonian Research Council grant PRG1604 and the Swedish Research Council grant 2022-01353, with additional support from the Carl Trygger Foundation project 22:2336.