Predicting the biochemical activities of UNIDENTIFIED CHEMICALS from MS^2 SPECTRA to pinpoint potential TOXIC AGENTS

UNIVERSITY OF TARTU

Stockholm University

Kruve lab

Ida Rahu[1,2] Meelis Kull[1] and Anneli Kruve[2,3]

ida.rahu@mmk.su.se

[1] Institute of Computer Science, University of Tartu, Narva mnt 18, 51009, Tartu, Estonia;
[2] Department of Materials and Environmental Chemistry, Stockholm University, Svante Arrhenius Väg 16, SE-106 91 Stockholm, Sweden;
[3] Department of Environmental Science, Stockholm University, Svante Arrhenius Väg 16, SE-106 91 Stockholm, Sweden

BACKGROUND

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

Nontarget LC/ESI/HRMS enables the simultaneous detection of numerous chemicals, but their identification remains limited (<5%), leaving gaps in toxicity assessment.[2-4]

The molecule's toxicity is associated with specific structural patterns,[5] which can be extracted as molecular fingerprint features from MS^2 spectra using SIRIUS+CSI:FingerID.

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

REFERENCES

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

MS2Tox

Tox21 10K dataset

#	CHEMICAL	12 TOXICITY ASSAYS				
1	BrC(Br)Br	0	1	NaN	...	0
...	...	...	...	...	...	...
11764	Nc1cc(Cl)ccc1O	1	NaN	0	...	0

8043 chemicals with varying activity

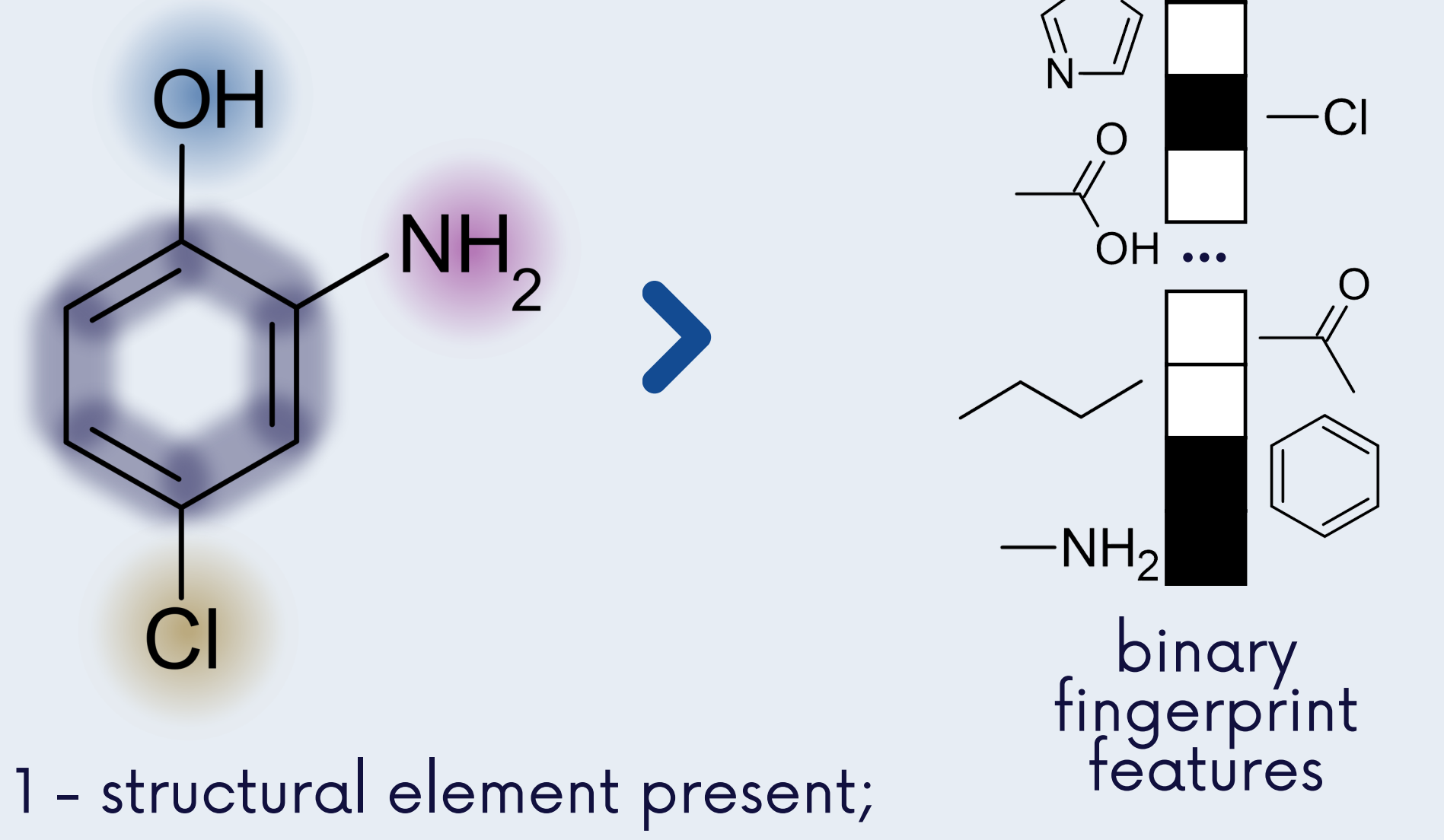
1 - active; 0 - non-active

DATA FOR TRAINING THE MODELS

REAL-LIFE TEST SET
MassBank and MoNA

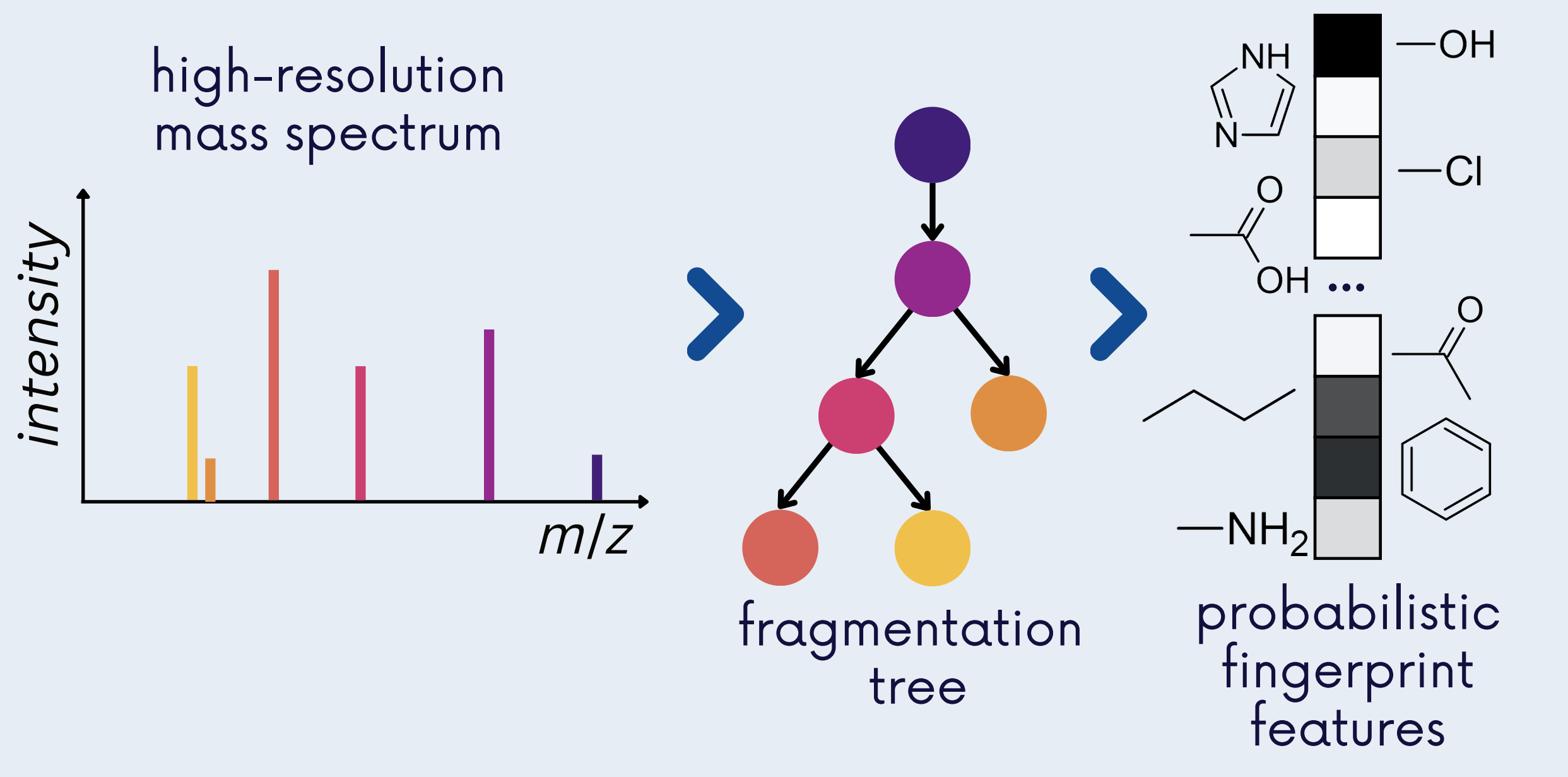
R package "rcdk" fingerprints from SMILES

SMILES: Nc1cc(Cl)ccc1O

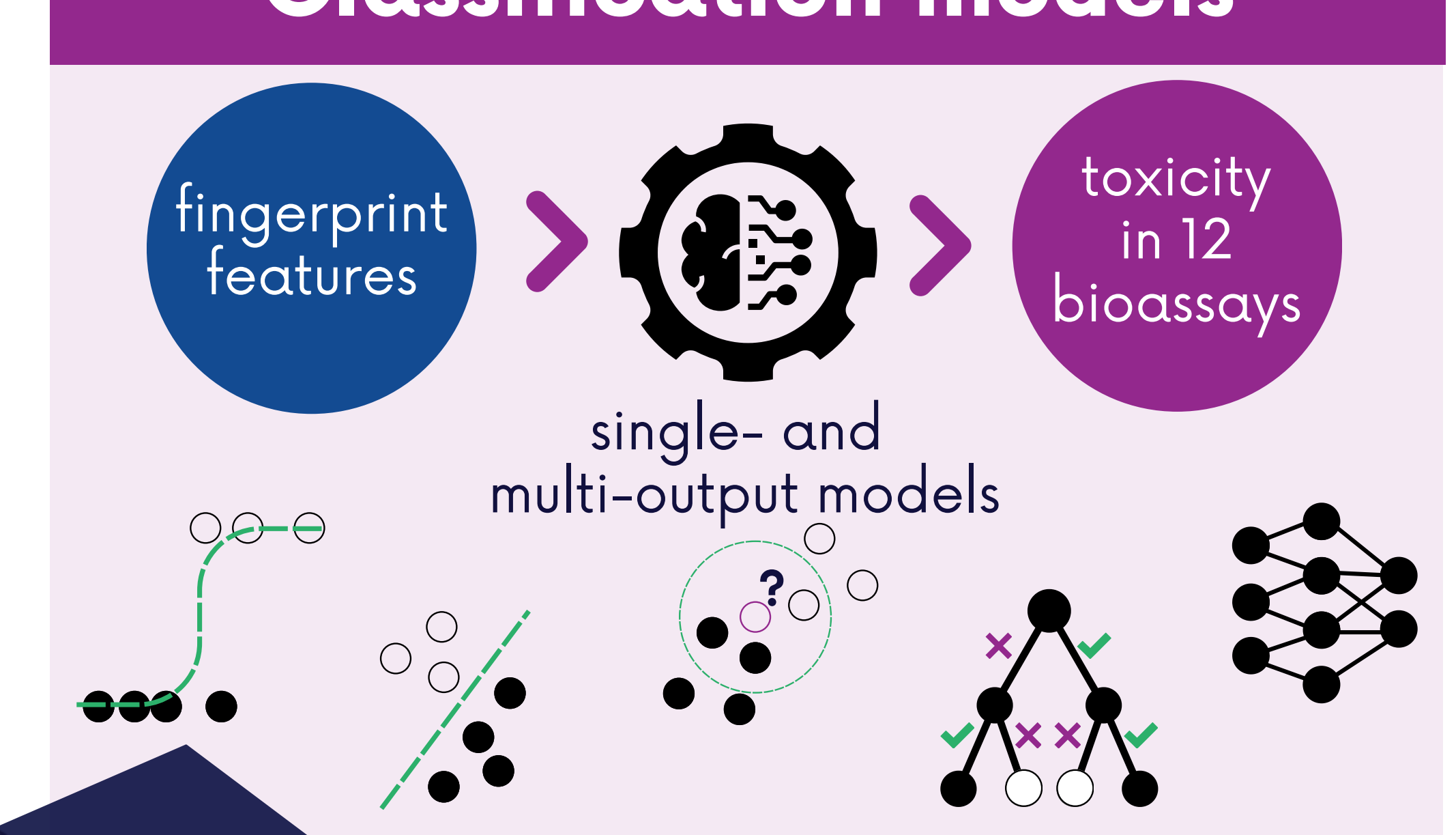


■ = 1 - structural element present;
□ = 0 - structural element missing

SIRIUS + CSI:FingerID fingerprints from MS

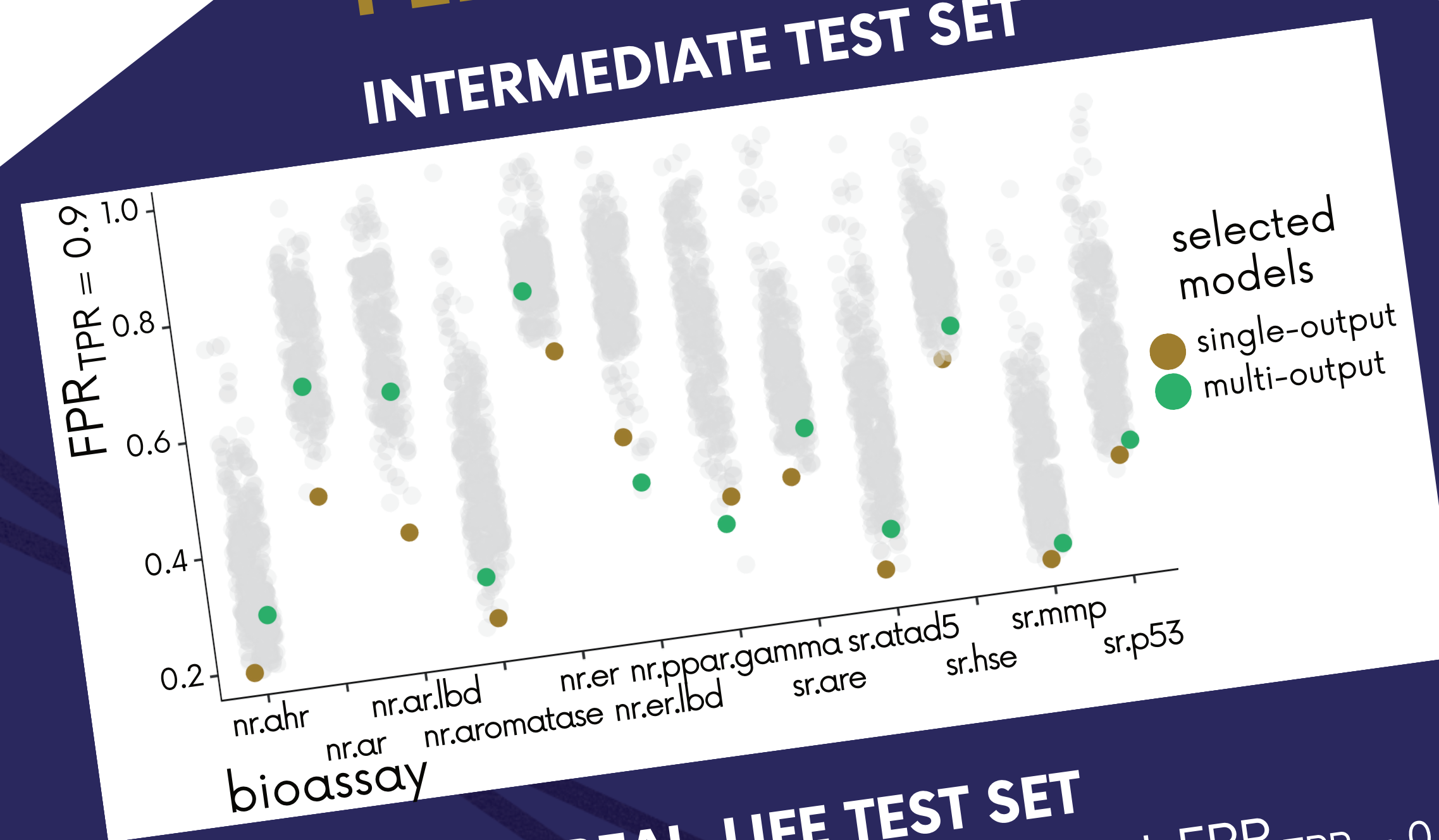

high-resolution mass spectrum
intensity
m/z
fragmentation tree
NH
N
OH
Cl
O
OH ...
NH_2
probabilistic fingerprint features

Classification models



MODELS' PERFORMANCE

INTERMEDIATE TEST SET



REAL-LIFE TEST SET

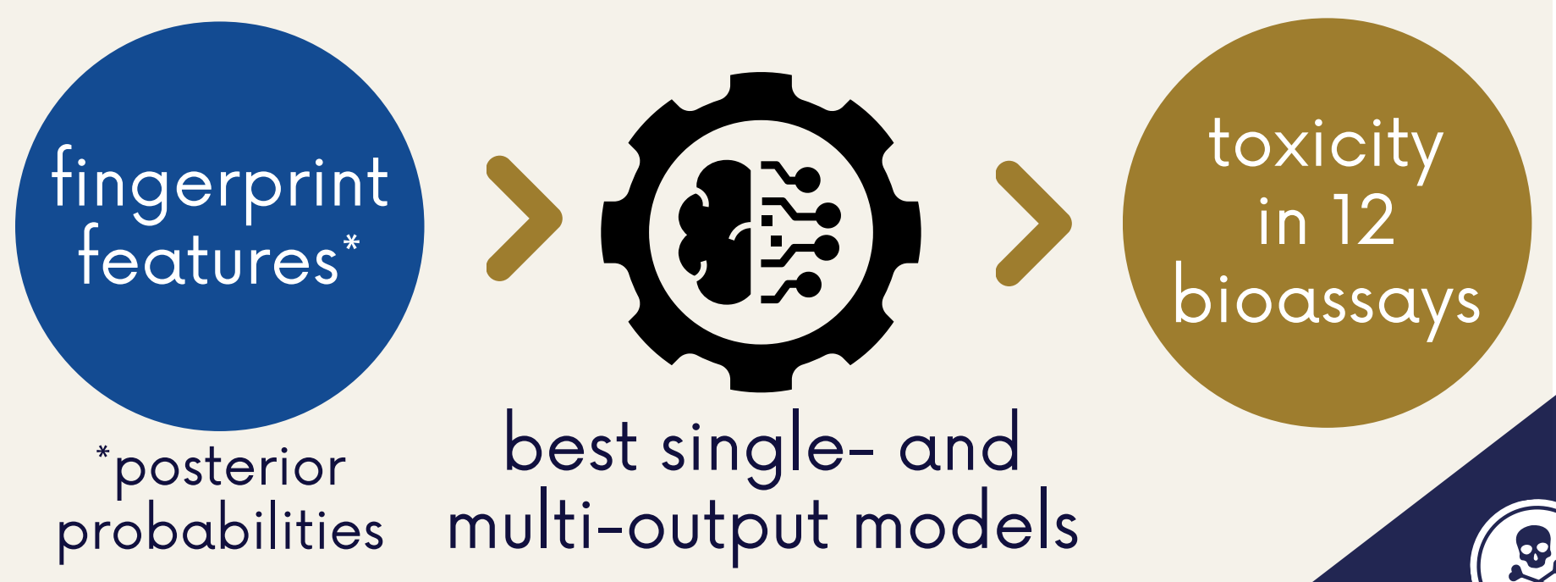
Depending on the bioassay, the lowest $FPR_{TPR = 0.9}$ ranged from 0.251 (sr.mmp) to 0.824 (nr.ar), consistent with the trends observed in the Tox21 Data Challenge, implying a potential reduction of up to 75% in the post-processing workload for nontarget HRMS.

METRIC

$FPR_{TPR = 0.9}$

aiming for high recall to detect the majority of active chemicals while minimizing the workload associated with misclassified ones

Evaluated models

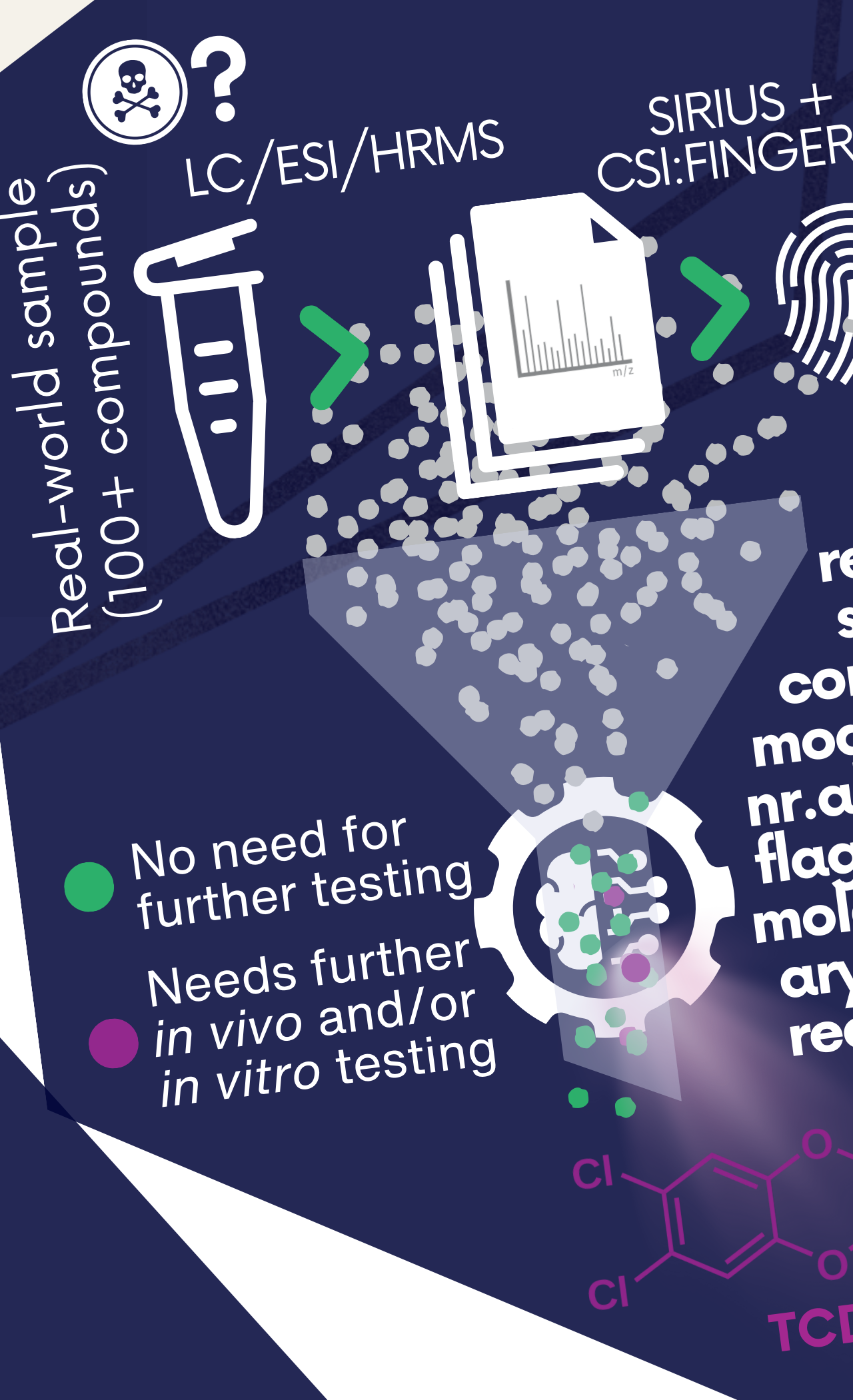


*posterior probabilities

		TRUE CLASS	
		active	non-active
PREDICTED CLASS	active	TP	FP
	non-active	FN	TN

FPR at 90% of recall

FIELD OF APPLICATION

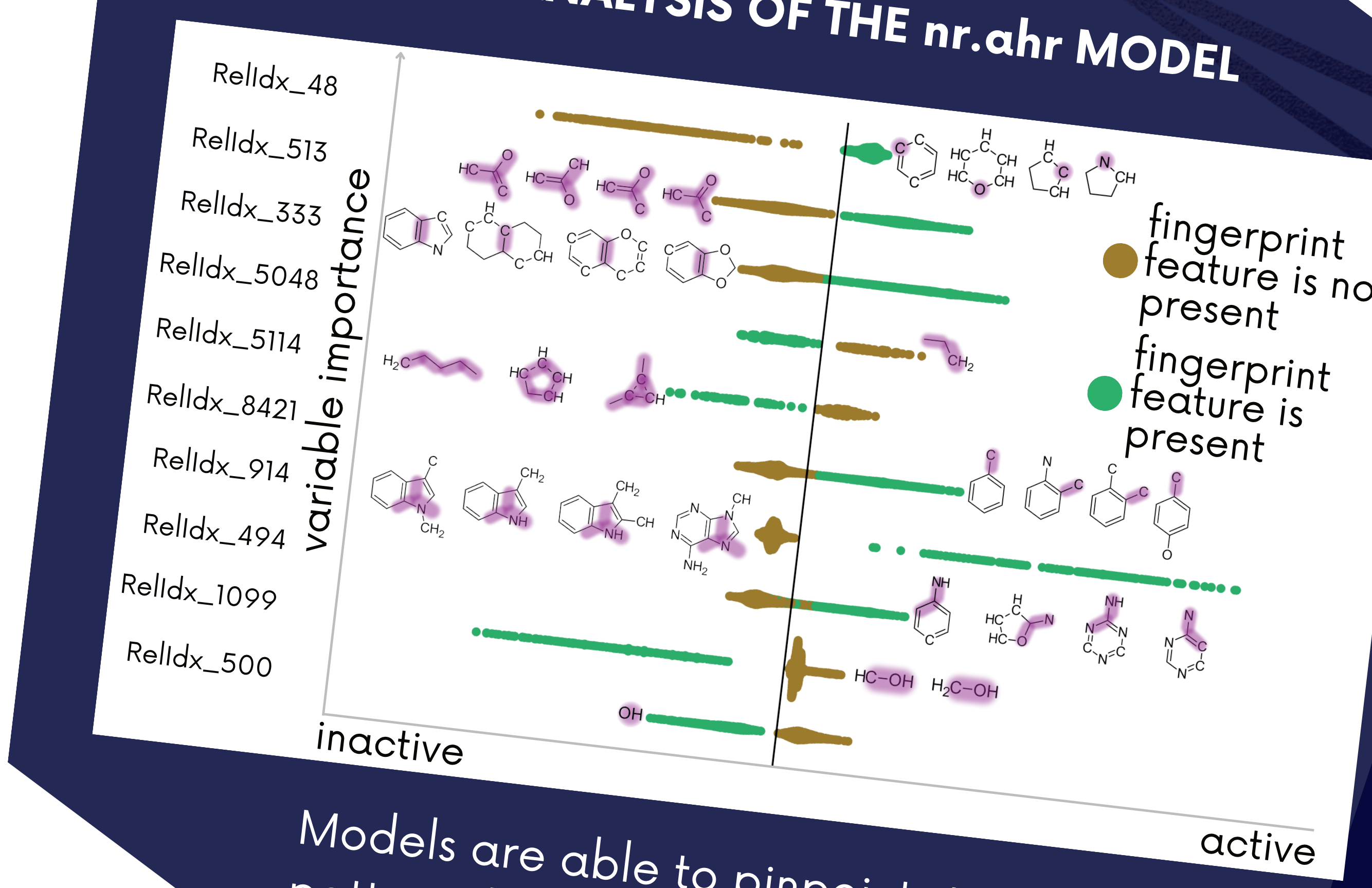

?
LC/ESI/HRMS
SIRIUS + CSI:FINGERID
Real-world sample (100+ compounds)
m/z
No need for further testing
Needs further in vivo and/or in vitro testing
Cl
Cl
O
O
Cl
Cl
TCDD

✓ **rapid**
✓ **cost-effective**
✓ **no animal testing**

Under real-world screening conditions, the model predicting nr.ahr successfully flagged the TCDD molecule, a known aryl hydrocarbon receptor agonist.

INTERPRETABILITY

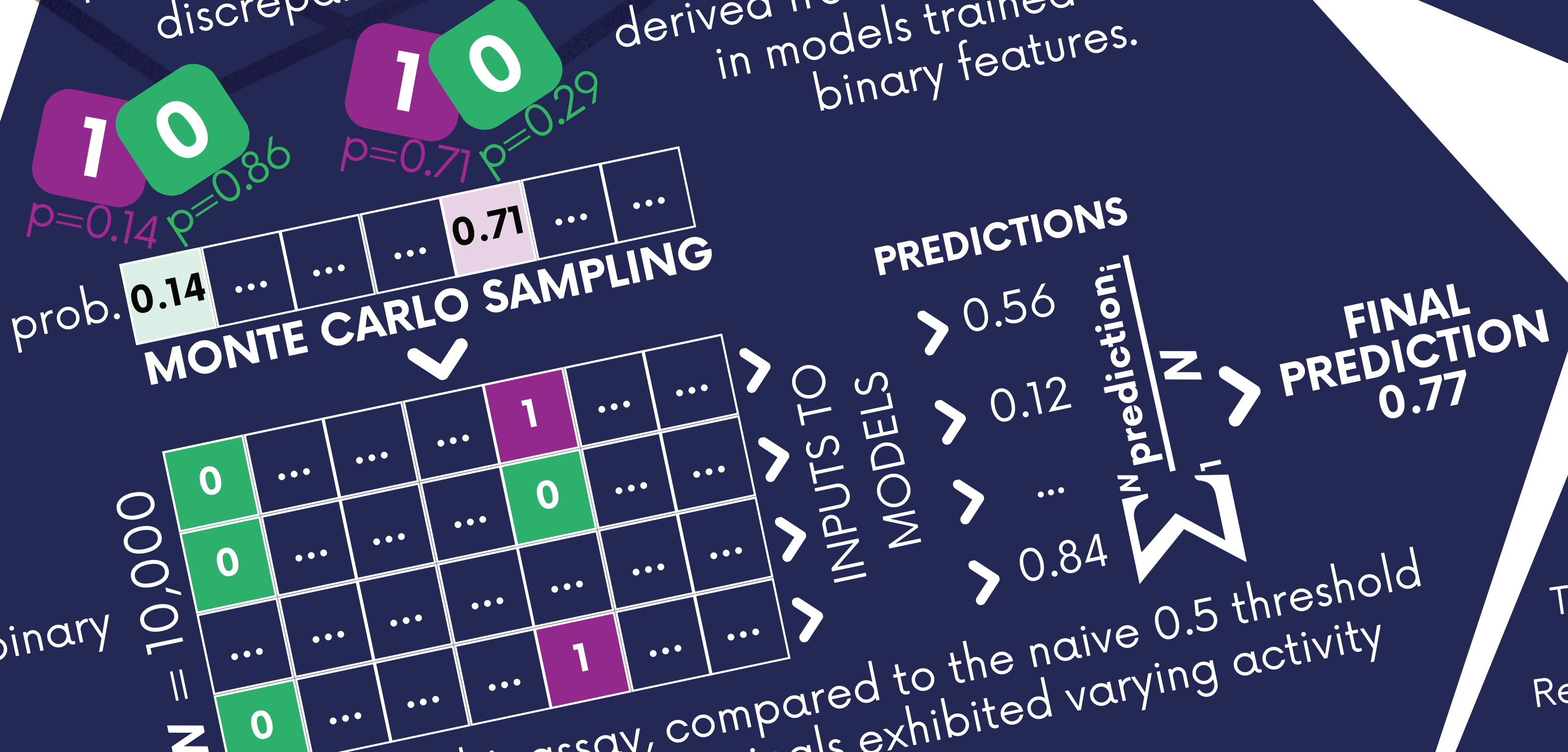
SHAP ANALYSIS OF THE nr.ahr MODEL



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

IMPLEMENTATION

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



Depending on the bioassay, compared to the naive 0.5 threshold approach, up to 20% of chemicals exhibited varying activity predictions.

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.