Making more of pesticide LC/HRMS data: machine learning for standard-free quantification and risk-based prioritisation

EPRW 2026, Rotterdam, 12.06.2026

Dr. Henrik Hupatz

Researcher

Stockholm University
Department of Chemistry

henrik.hupatz@su.se

Kruve Lab

Stockholm University

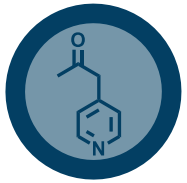
SUCCeSS

INTRODUCTION





Anneli Kruve



Structural annotation



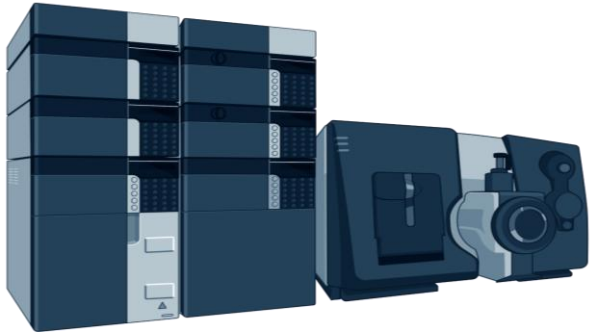
Semi-quantification



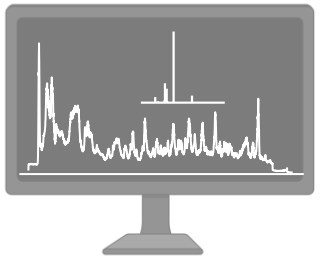
Hazard prediction



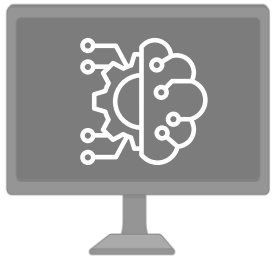
Diverse samples



LC/HRMS analysis



Data processing



Machine learning

H. Hupatz, I. Rahu, W.-C. Wang et al. *Anal. Bioanal. Chem*. **2025**, 473; L. Malm, E. Palm, A. Souihi et al. *Molecules.* **2021**, 3524.

INTRODUCTION – PESTICIDE ANALYSIS

?

Targeted analysis

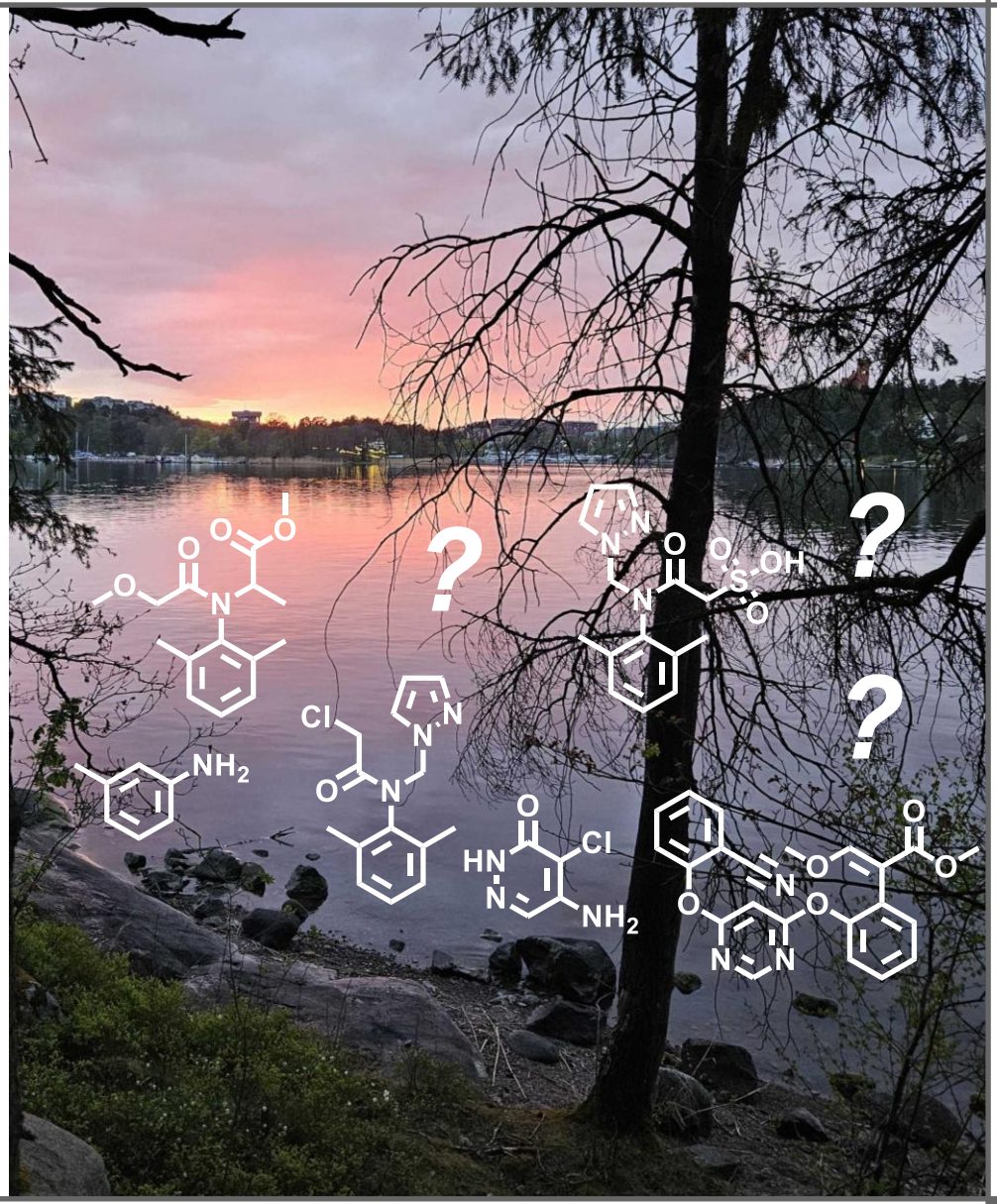
55/142 detected and **quantified**

Suspect screening

38/258 suspects flagged and **26/38** had **standards** available

Non-target screening

Unknown transformation products are missed



F. Menger, G. Böström, O. Jonsson et al. *Environ. Sci. Technol*. **2021**, 10343.

INTRODUCTION – PESTICIDE ANALYSIS

Semi-quantification

Hazard prediction

Targeted analysis

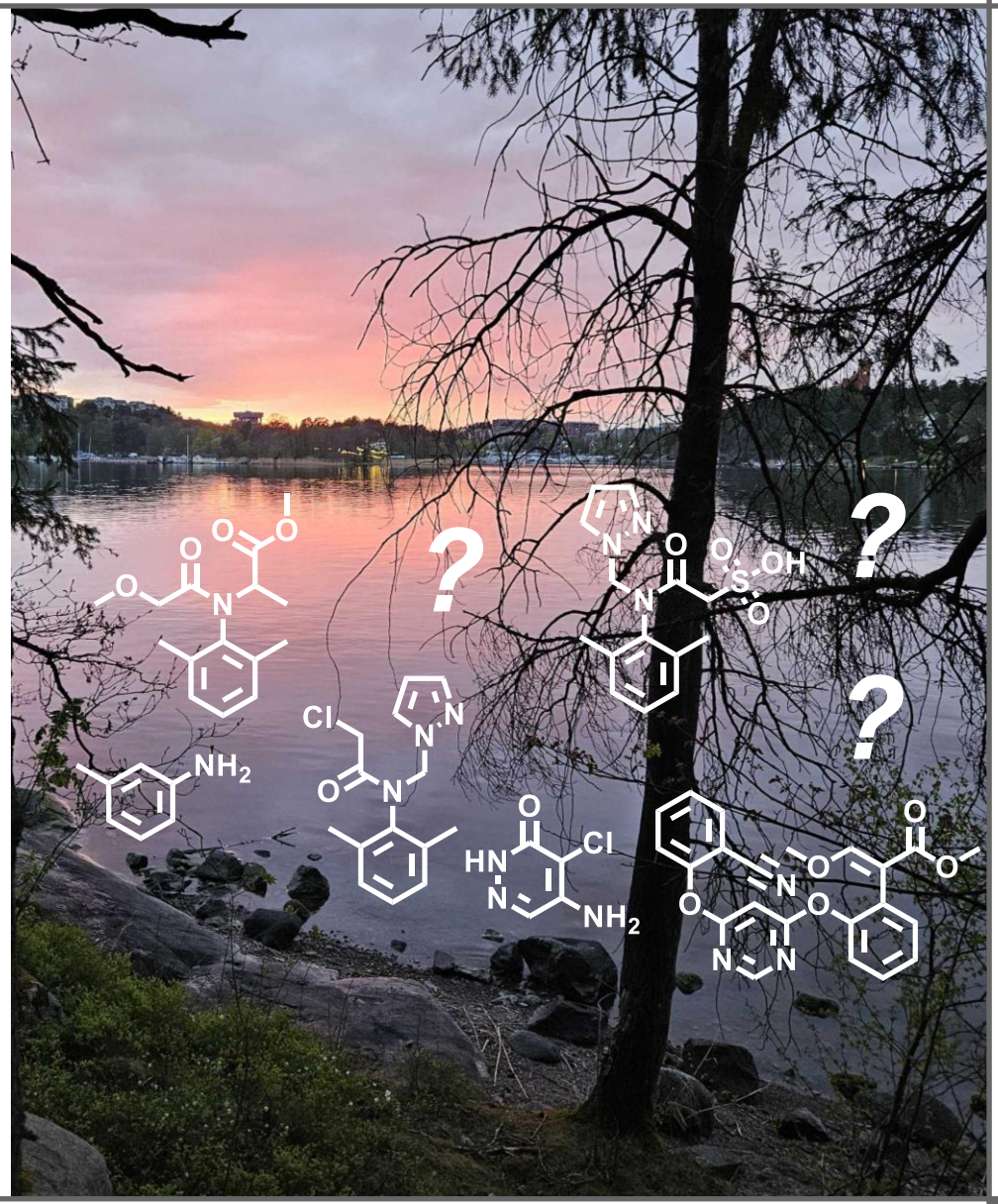
55/142 detected and **quantified**

Suspect screening

38/258 suspects flagged and **26/38** had **standards** available

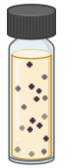
Non-target screening

Unknown transformation products are missed

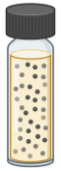


F. Menger, G. Böström, O. Jonsson et al. *Environ. Sci. Technol*. **2021**, 10343.

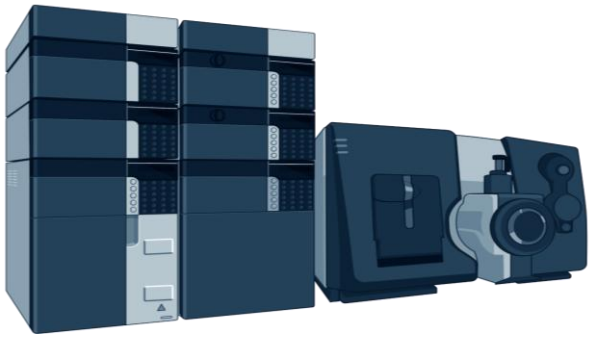
LC/HRMS WORKFLOW



Spiked standard mix



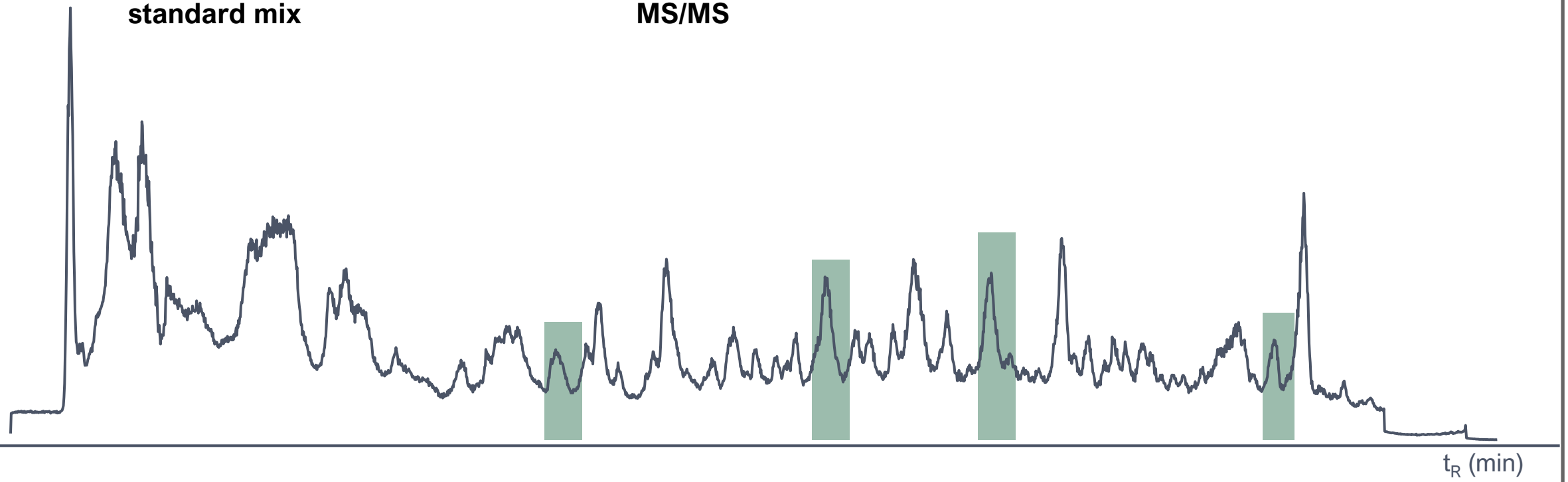
Sample



High-resolution LC–ESI-MS/MS



Data processing



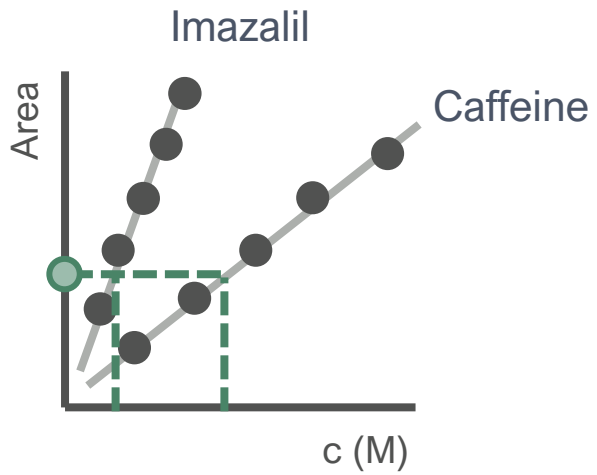
No experimental data.

LC/HRMS WORKFLOW



No experimental data. MS2 data taken from MassBankEU.

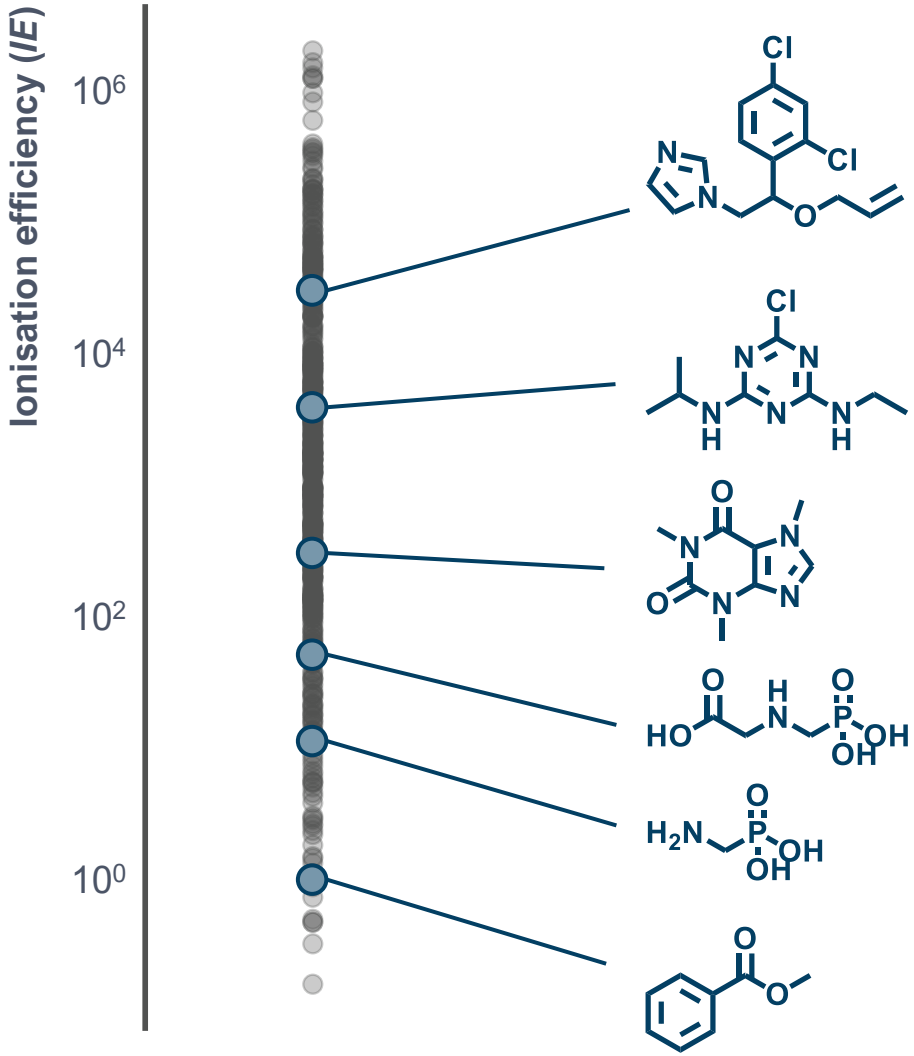
IONISATION EFFICIENCY





$$RF = \frac{\text{Area}}{c} \qquad rIE = \frac{RF1}{RF2}$$

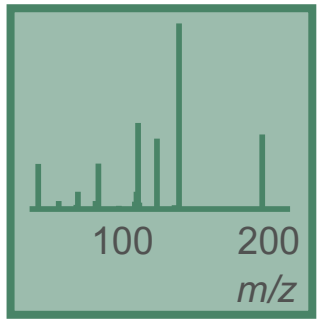
Mobile phase** and **instrument-specific



J. Liigand, T. Wang, J. Kellogg et al. *Sci. Reports* **2020**, 5808.

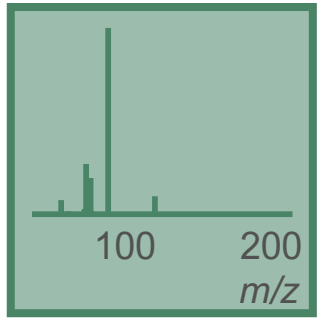
LC/HRMS WORKFLOW

Caffeine



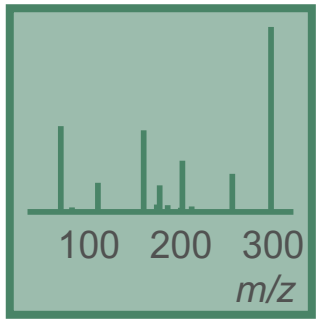
c = 25 nM

Thiram



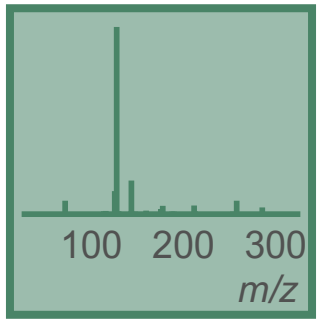
How to determine RF?

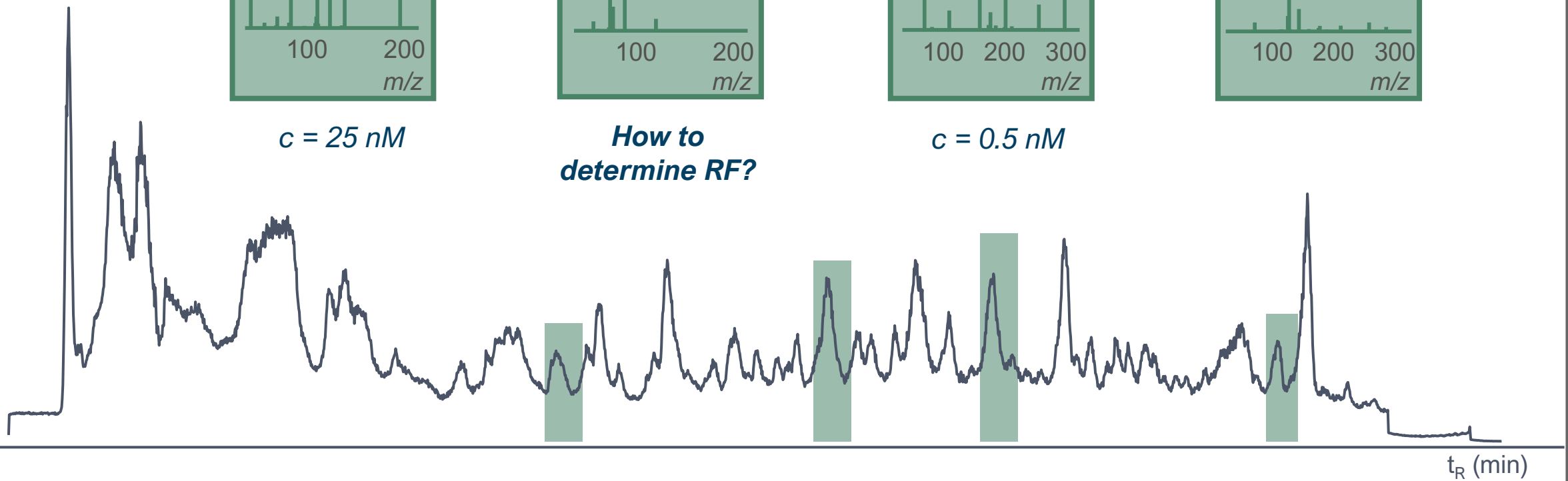
Imazalil



c = 0.5 nM

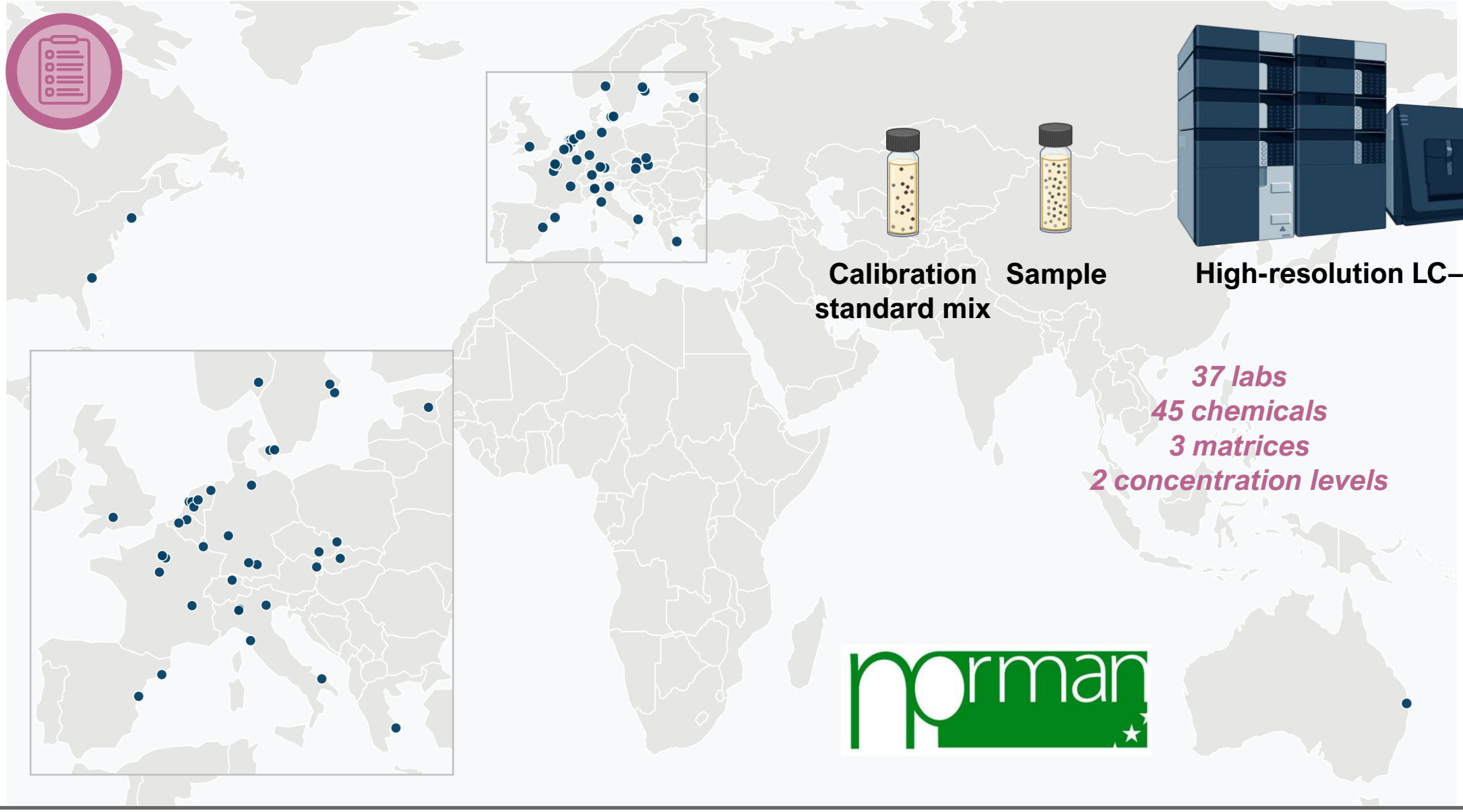
Unknown





No experimental data. MS2 data taken from MassBankEU.

SEMI-QUANTIFICATION INTERLABORATORY STUDY



Calibration standard mix

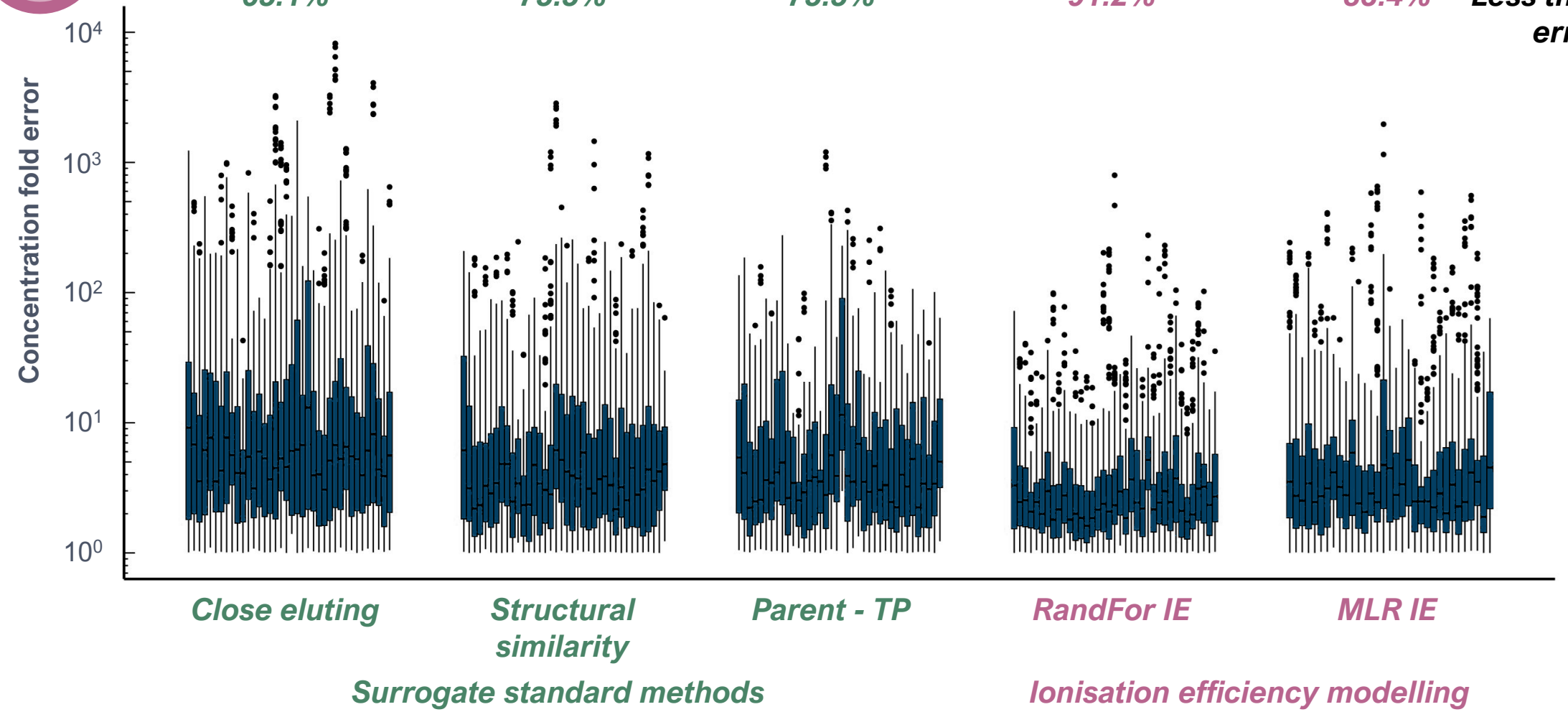
Sample

High-resolution LC–ESI-MS

37 labs
45 chemicals
3 matrices
2 concentration levels

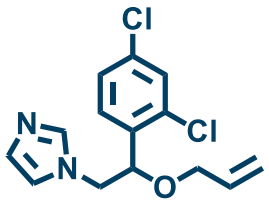
L. Malm, J. Liigand, R. Aalizadeh et al. *Anal. Chem.* **2024**, 16215.

SEMI-QUANTIFICATION INTERLABORATORY STUDY

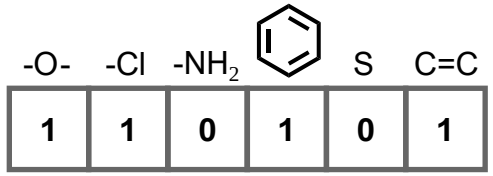


L. Malm, J. Liigand, R. Aalizadeh et al. *Anal. Chem.* **2024**, 16215. RandFor IE by Quantem.

IONISATION EFFICIENCY MODELS



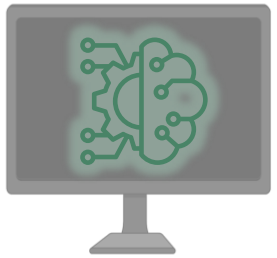
Structure



Molecular fingerprint

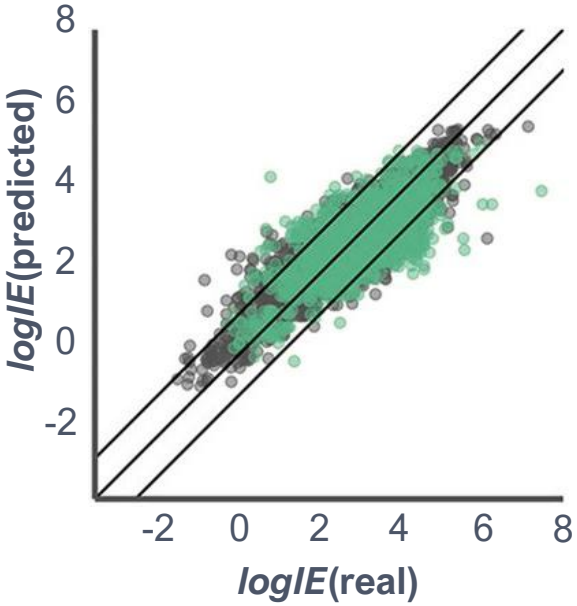


Mobile-phase descriptors

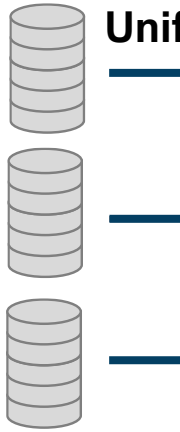


MS2Quant
xgbTree

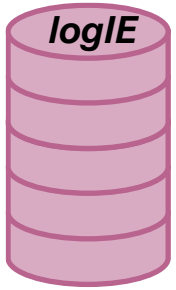
logIE



Mean fold error on test set: ***6.3x***
train set: 3.5x

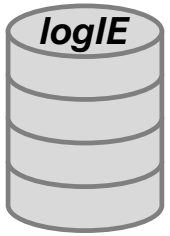


Unification



1191 chemicals
6049 datapoints

Splitting

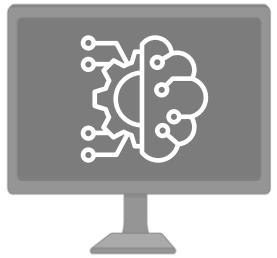


80% Train set



20% Test set

Training

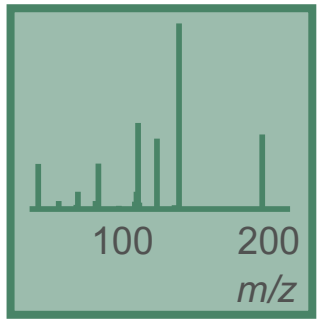


ML algorithms

H. Sepman, L. Malm, P. Peets et al. *Anal. Chem.* **2023**, 12329.

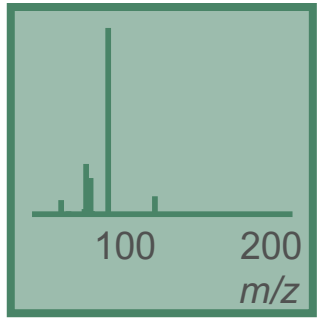
LC/HRMS WORKFLOW

Caffeine



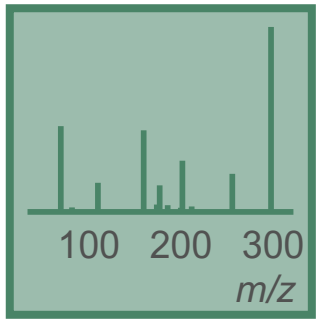
c = 25 nM

Thiram



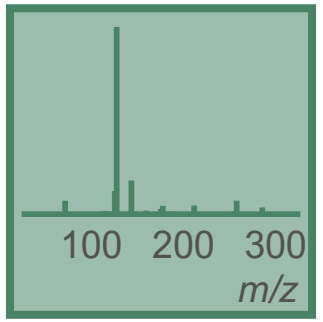
How to determine RF?

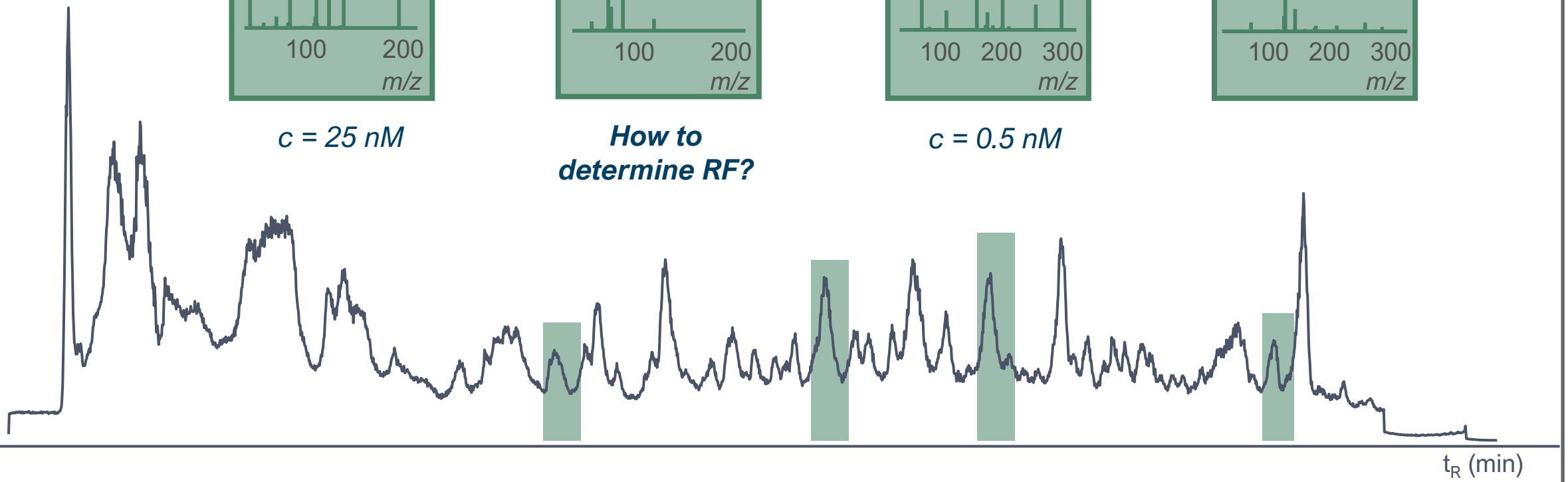
Imazalil



c = 0.5 nM

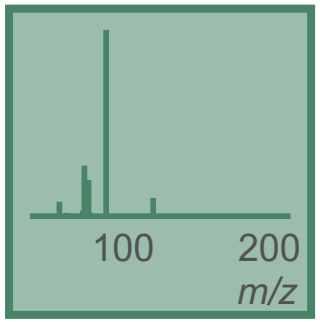
Unknown



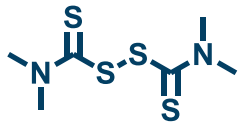


No experimental data. MS2 data taken from MassBankEU.

SEMI-QUANTIFICATION WORKFLOW



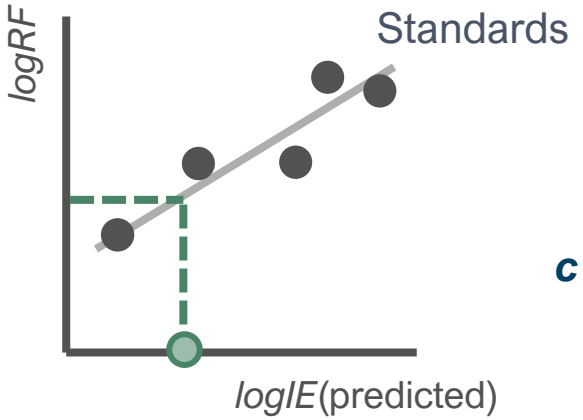
MS² data



Identification



Predicted by MS2Quant



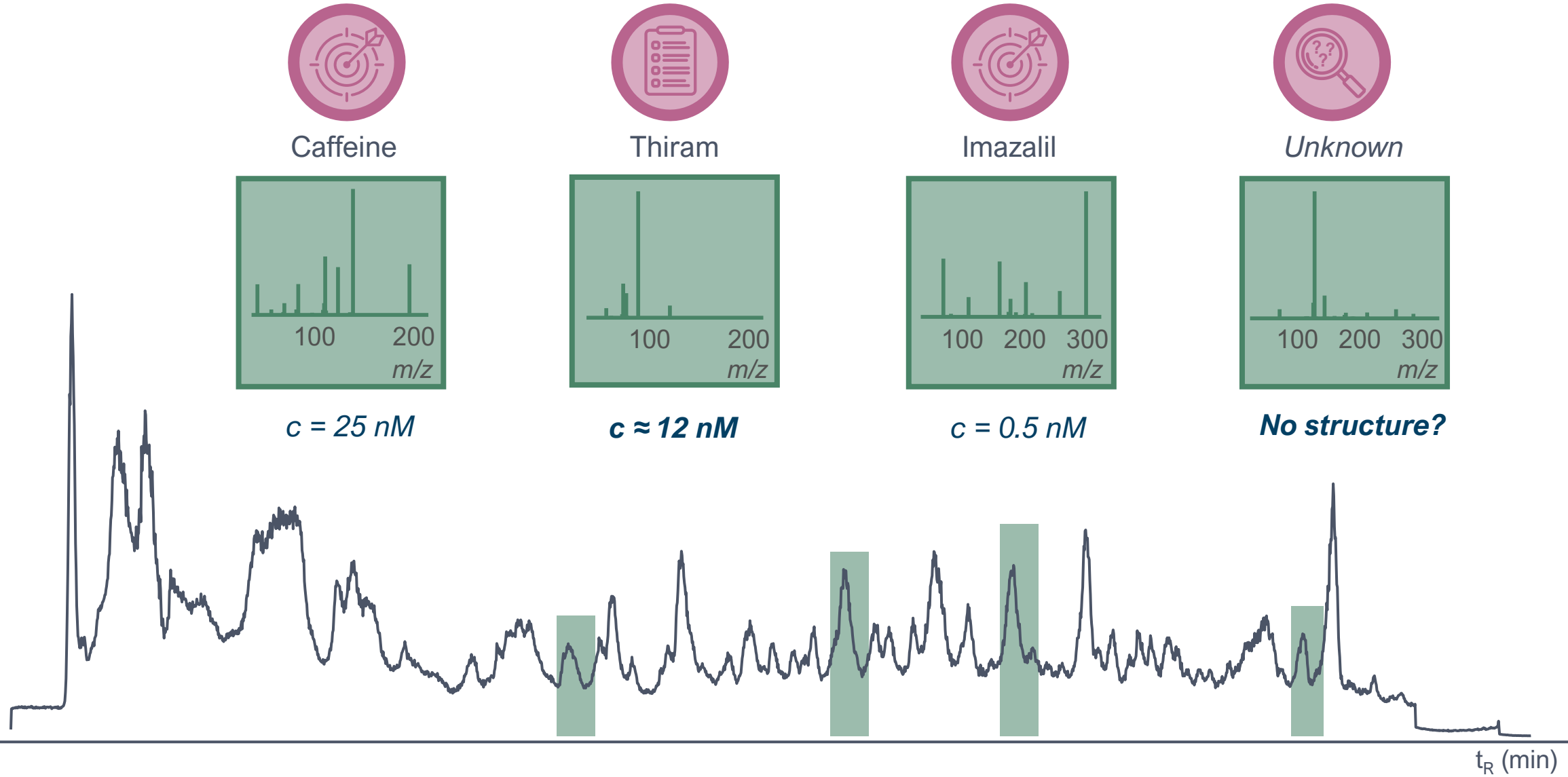
Calibration

$$c\,(suspect) = \frac{Area}{RF(suspect)}$$

Concentration

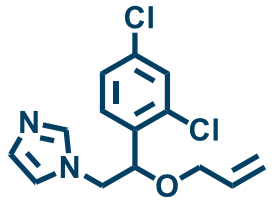
$$logRF(suspect) = slope \cdot logIE + intercept$$

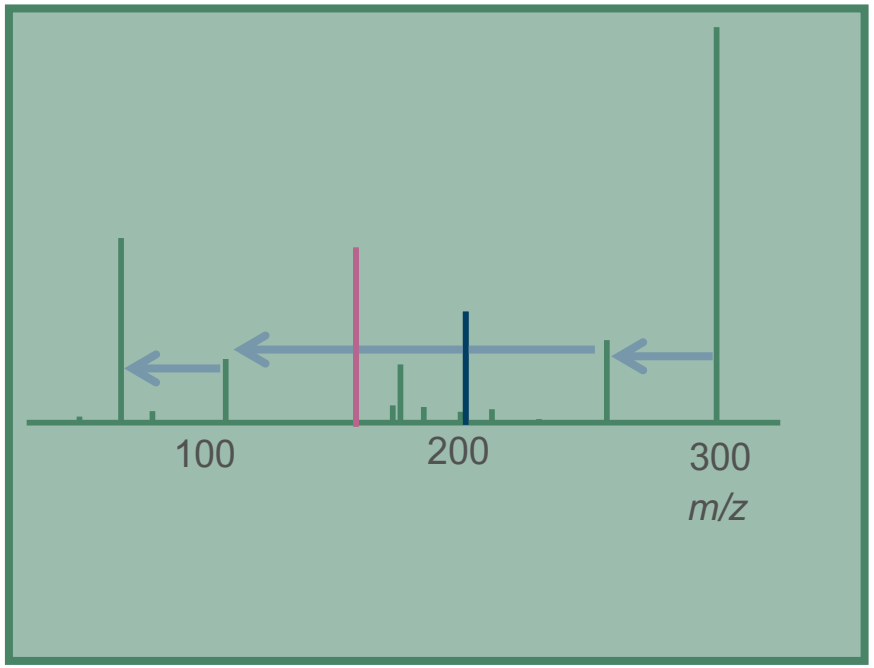
LC/HRMS WORKFLOW



No experimental data. MS2 data taken from MassBankEU.

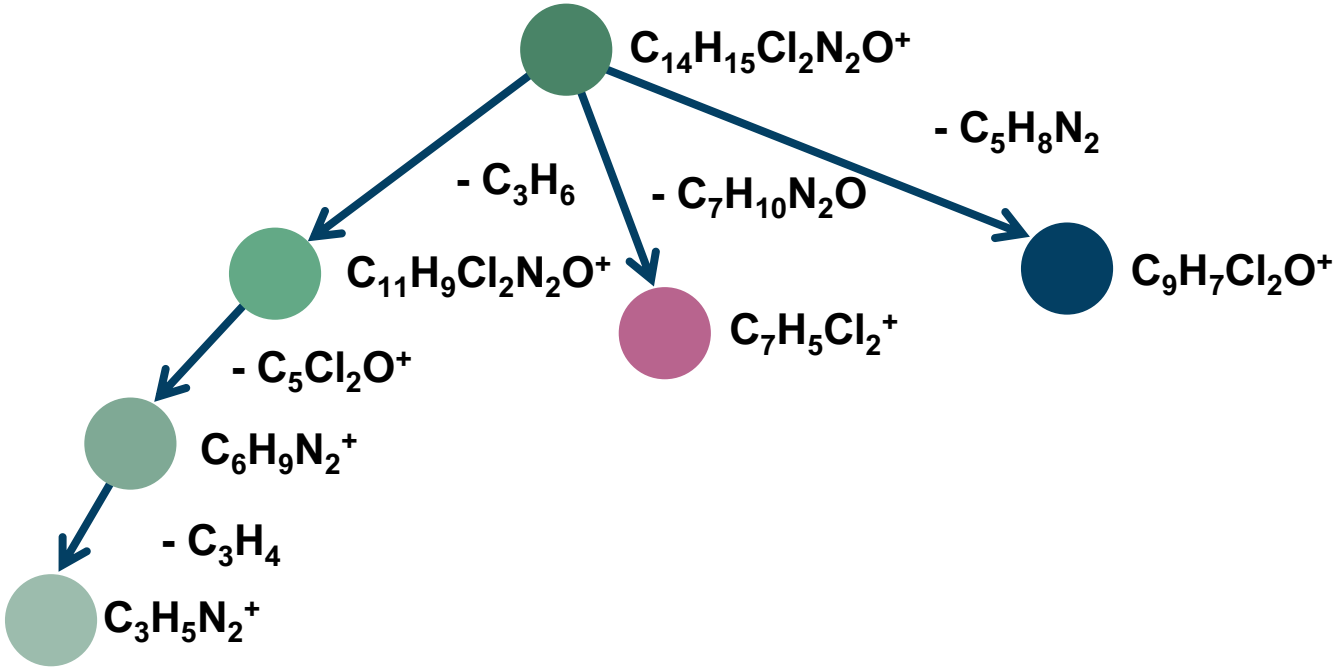
MOLECULAR FINGERPRINTS FROM MS2 DATA





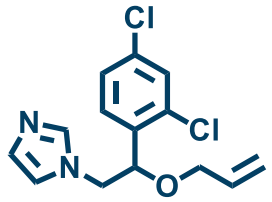
MS² data

Fragmentation tree

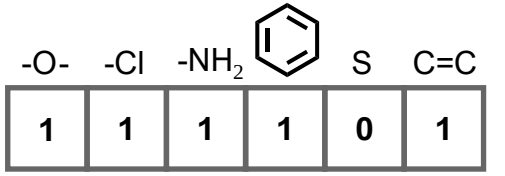


K. Dührkop, M. Fleischauer, M. Ludwig et al. *Nature Methods* **2019**, 299. MS2 data taken from MassBankEU.

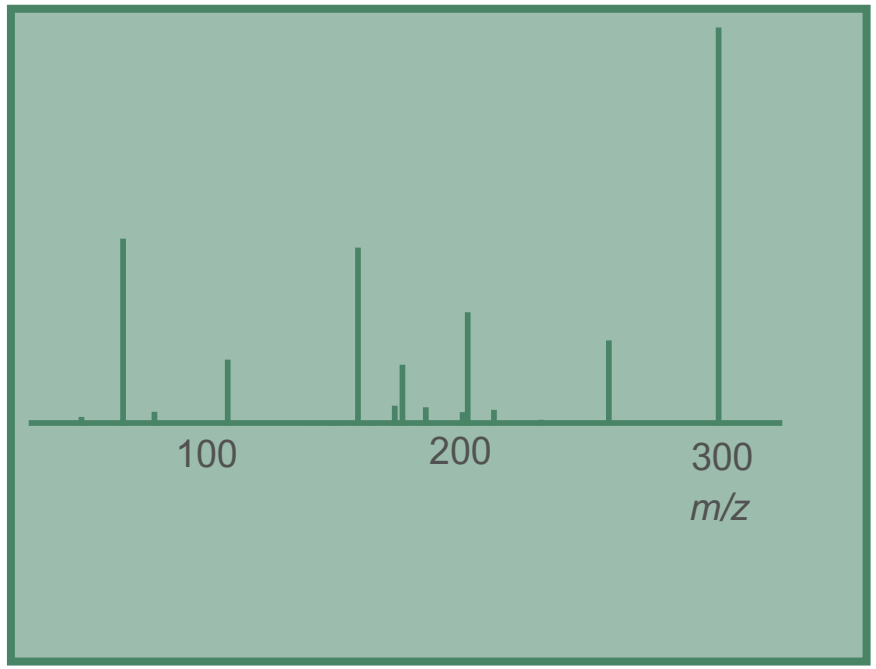
MOLECULAR FINGERPRINTS FROM MS2 DATA



Structure

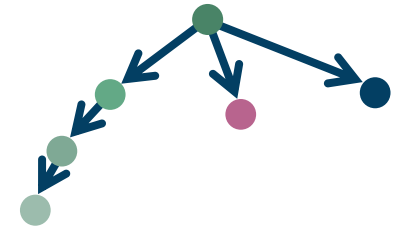


Molecular fingerprint

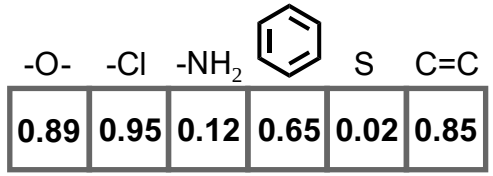


MS² data

Fragmentation tree



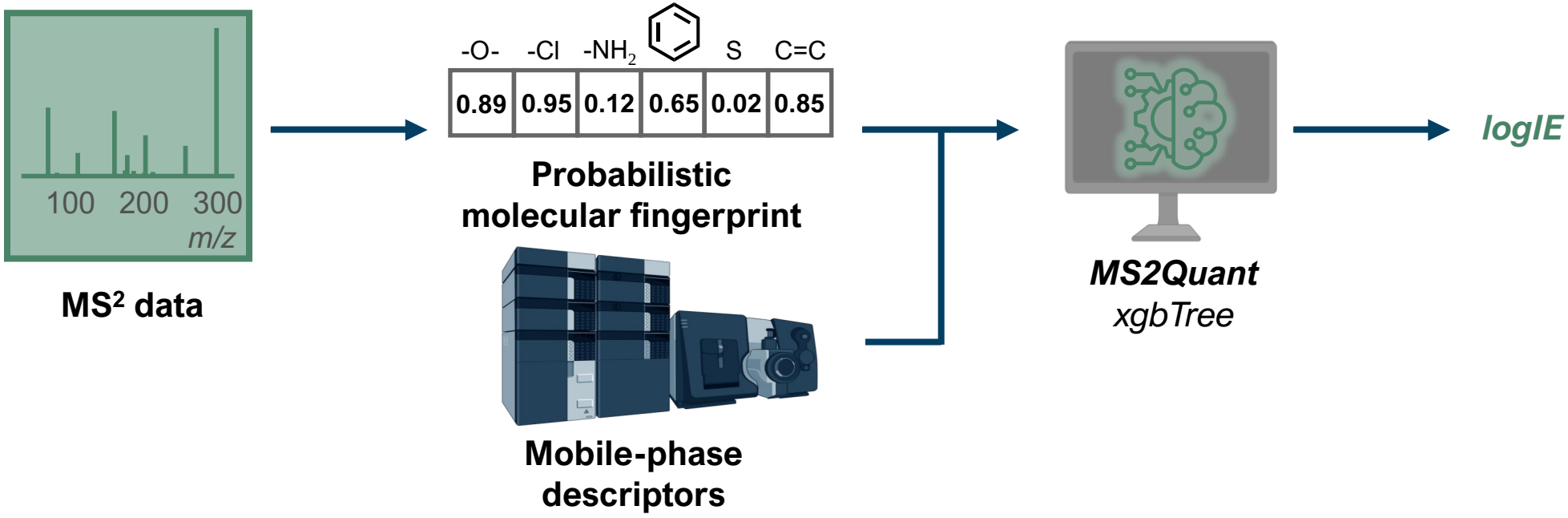
SIRIUS+CSI:FingerID



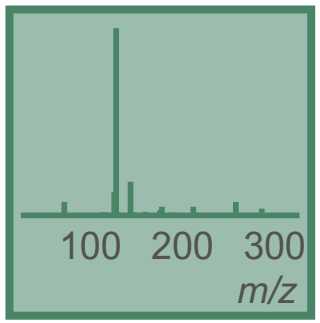
Probabilistic molecular fingerprint

SIRIUS

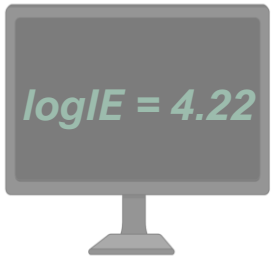
MS2QUANT SPECTRAL MODE



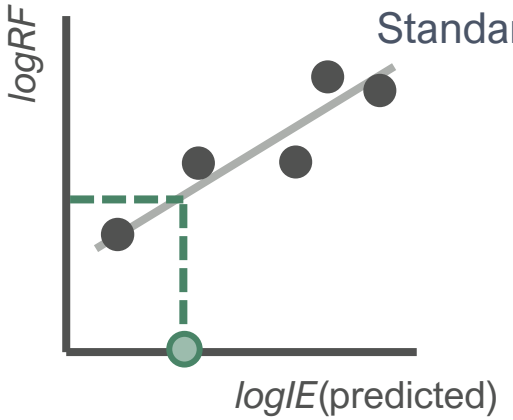
SEMI-QUANTIFICATION WORKFLOW



MS² data



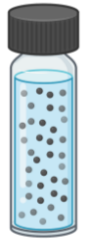
Predicted by MS2Quant



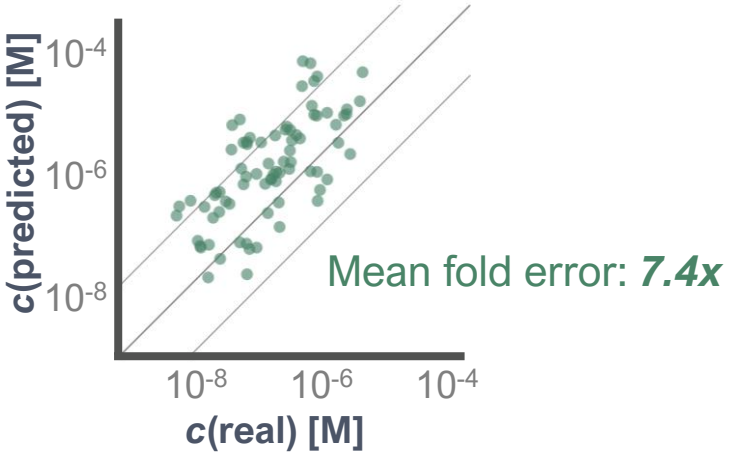
Calibration

$$c\,(suspect) = \frac{Area}{RF(suspect)}$$

Concentration

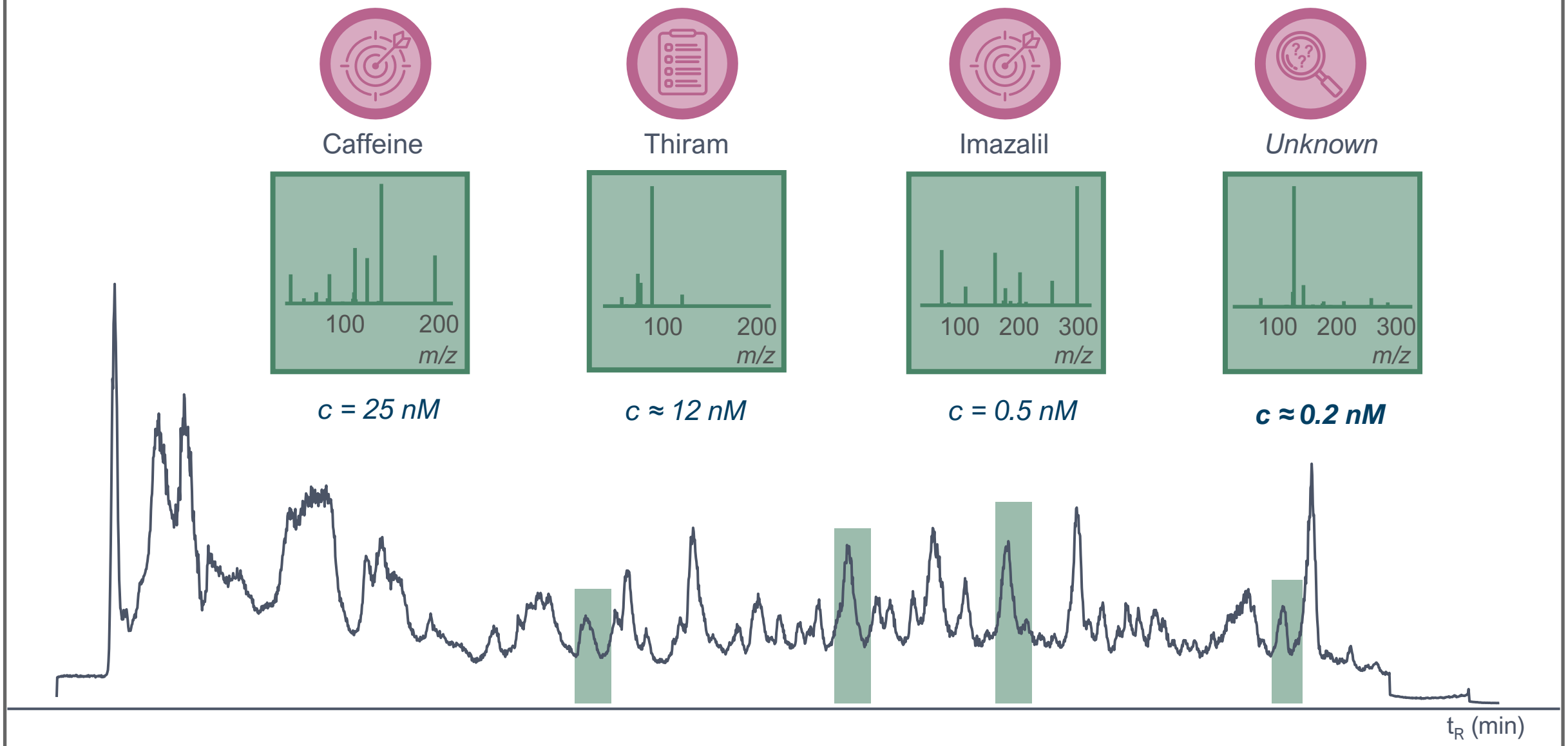


39 spiked standards



H. Sepman, L. Malm, P. Peets et al. *Anal. Chem.* **2023**, 12329.

LC/HRMS WORKFLOW



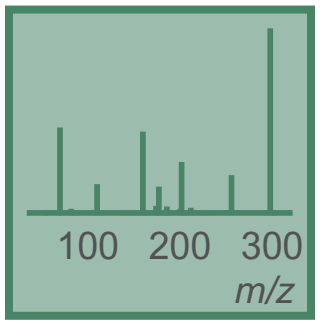
No experimental data. MS2 data taken from MassBankEU.

OUTLOOK MS2QUANT

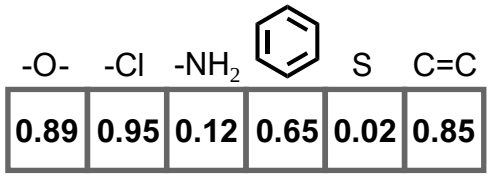
	MS2Quant	*MS2QuantV2*
Ionisation mode	*+*	*+ and -*
Unified datasets	*13*	*23 (20)*
Chemicals	*1191*	*1455 (899)*
In PPDB	*305*	*480 (240)*
Available in	R	R, Python, *WebApp*
Applicability domain	*No*	*Score*

H. Hupatz, L. Malm, F. Menger et al. *in preparation.*

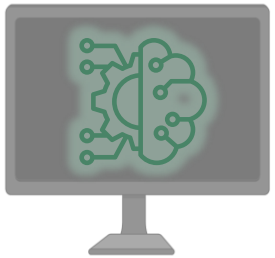
HAZARD PREDICTION FROM MS2 DATA



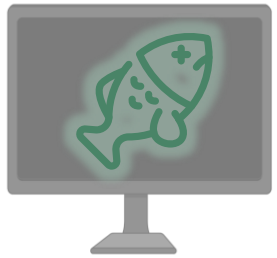
MS² data



Probabilistic molecular fingerprint

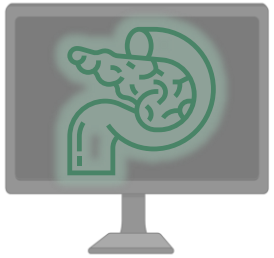


MS2Tox



Ecotoxicity

LC50 (mM)

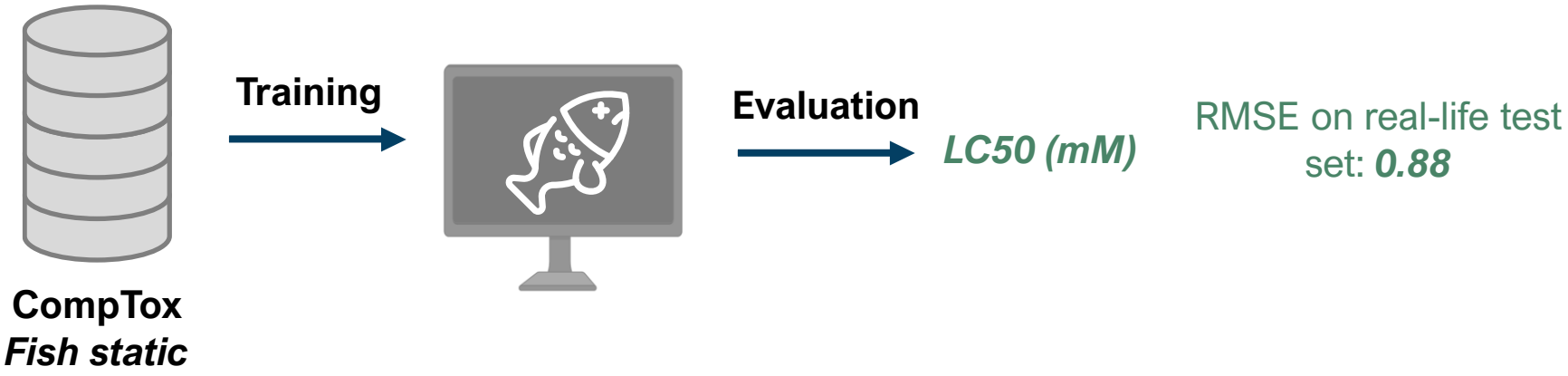


Endocrine disruption

Activity in 12 bioassays

I. Rahu, M. Kull, A. Kruve *J. Chem. Inf. Model.* **2024**, 3093. P. Peets, W.-C. Wang, M. MacLeod et al. *Environ. Sci. Technol.* **2022**, 12329.

TRAINING OF MS2TOX MODELS



P. Peets, W.-C. Wang, M. MacLeod et al. *Environ. Sci. Technol.* **2022**, 12329.

TRAINING OF MS2TOX MODELS

		true label	
		active	non-active
prediction	active	TP	FP
	non-active	FN	TN

$$TPR = \frac{TP}{(TP + FN)} = 90\%$$

$$FPR = \frac{FP}{(FP + TN)}$$

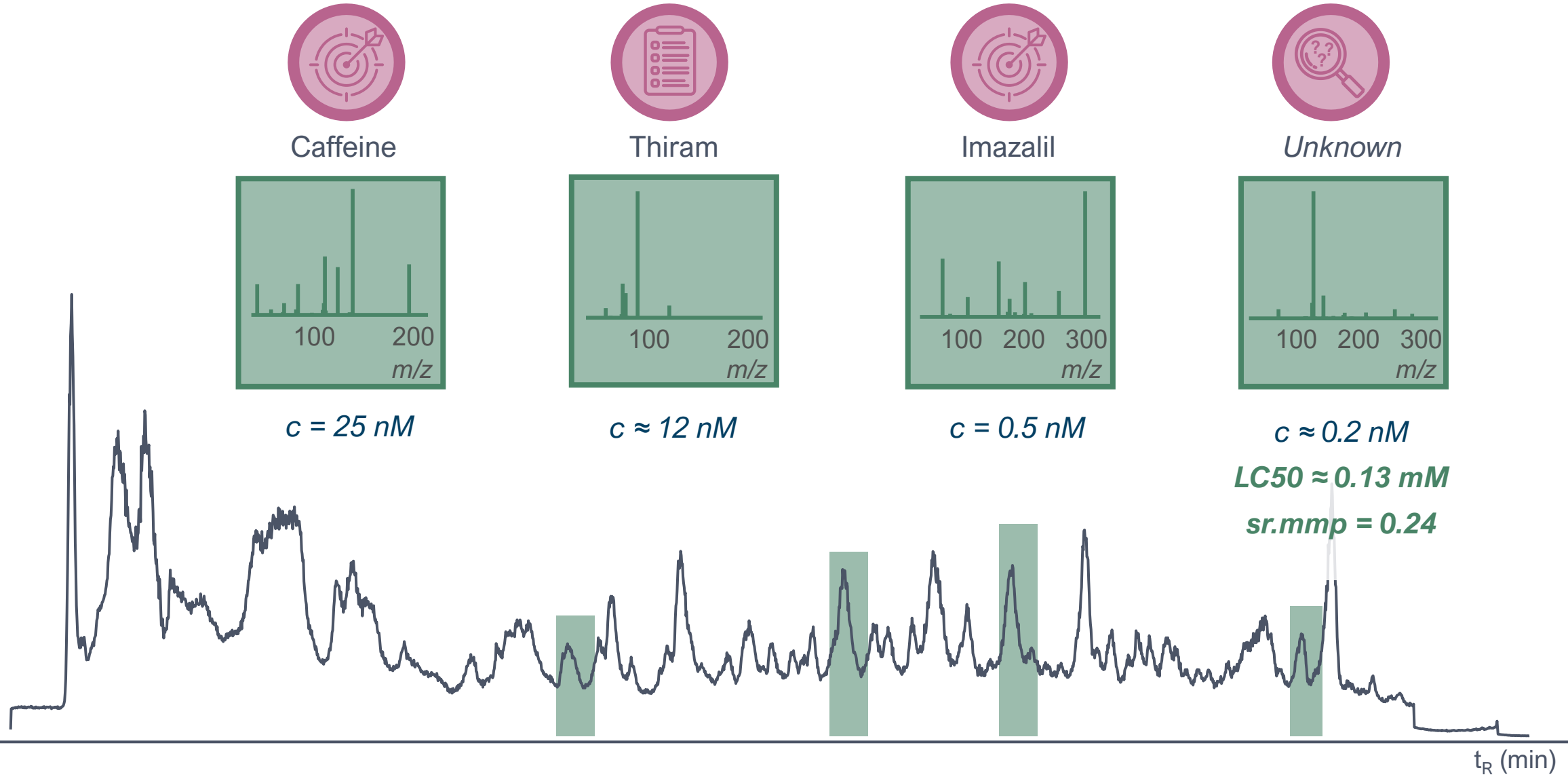


On real-life test set

bioassay	FPR(@TPR=90%)
sr.mmp	25.1%
sr.p53	25.4%
nr.ahr	41.8%
…	…
nr.ar	82.4%
nr.er	85.0%

I. Rahu, M. Kull, A. Kruve *J. Chem. Inf. Model.* **2024**, 3093.

LC/HRMS WORKFLOW



No experimental data. MS2 data taken from MassBankEU.

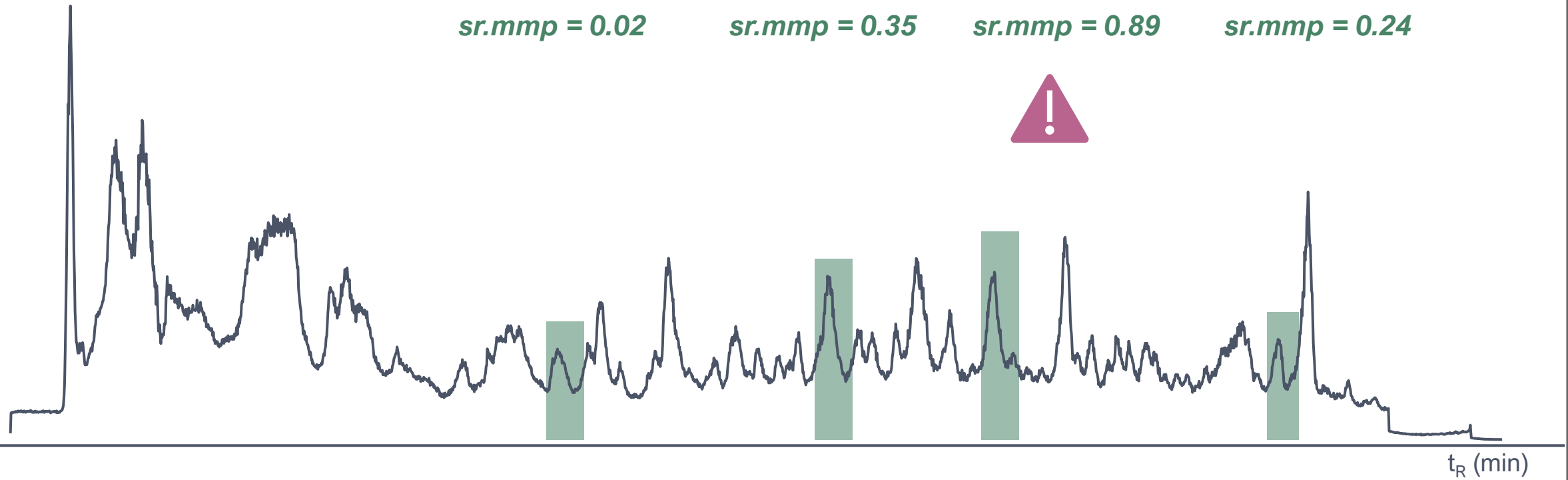
RISK-BASED PRIORITISATION

Unknown1
c ≈ 15 nM
LC50 ≈ 0.2 mM
sr.mmp = 0.02

Unknown2
c ≈ 12 nM
LC50 ≈ 0.02 mM
sr.mmp = 0.35

Unknown3
c ≈ 0.5 nM
LC50 ≈ 0.05 mM
sr.mmp = 0.89

Unknown4
c ≈ 20 nM
LC50 ≈ 0.13 mM
sr.mmp = 0.24



D. Szabo, S. Fischer, A. P. Mathew et al. *Anal. Chem*. **2024**, 14150; H. Sepman, L. Jonsson, M. Posselt et al. *Water Reserch, submitted.*

CONCLUSION AND OUTLOOK

Structural annotation

Semi-quantification

Hazard prediction

Suspect screening

ML models > surrogate models

Non-target screening

×

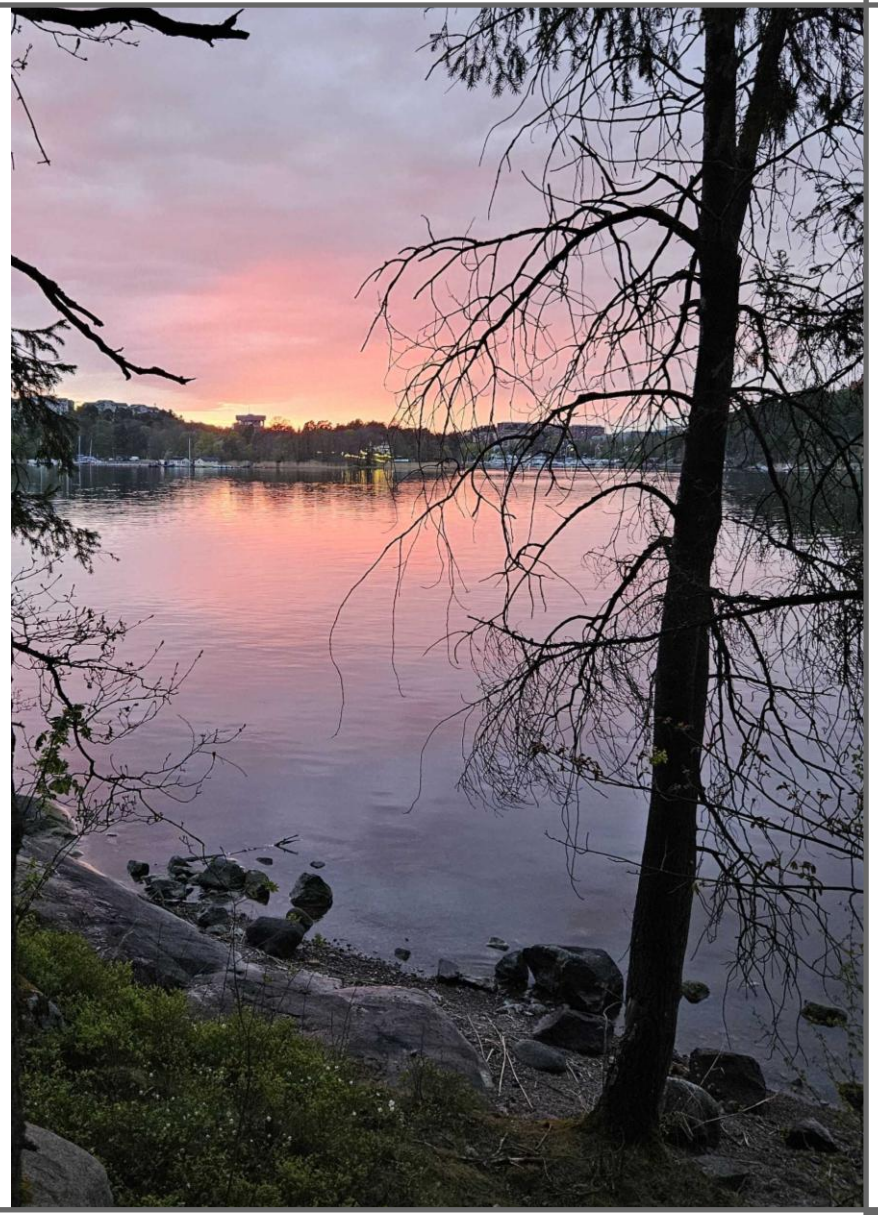
ML makes possible

ML makes possible

Risk-based prioritisation

Limitations

Data amount and quality



henrik.hupatz@su.se, anneli.kruve@su.se

ACKNOWLEDGEMENTS



Anneli Kruve

Kruve Lab



Helen Sepman



Louise Malm



Frank Menger



Jaanus Liigand



Ida Rahu



Some Figures were created with BioRender.com.