27^{èmes} Journées Scientifiques

MS2Tox & MS2Quant Automated detection and identification of **TOXIC and HIGH-RISK chemicals**

Villars-sur-Ollon 2024

Kruve lab

IDA RAHU Postdoctoral Fellow































time



time





















NORMAN SusDat (n = 55793)

T. Hulleman et al., Environ. Sci. Tech. (2023).



NORMAN SusDat (n = 55793)

detected structures (n = 1606)

level 1 to 2b identification confidence

T. Hulleman et al., Environ. Sci. Tech. (2023). E. L. Schymanski et al., Environ. Sci. Tech. (2014).





NORMAN SusDat (n = 55793)

detected structures (n = 1606)level 1 to 2b identification confidence

T. Hulleman et al., Environ. Sci. Tech. (2023). E. L. Schymanski et al., Environ. Sci. Tech. (2014).







in silico tools





in silico tools

MOLECULAR REPRESENTATIONS

• fixed representations

1D

- number of atoms
- number of bonds
- molecular weight
- atomic properties

MW = 321.8936 (g/mol)

deep representations

LINEAR NOTATIONS

SMILES STRINGS

C1 = C2C(=CC(=C1CI)CI)OC3 = CC(=C(C=C3O2)CI)CI

2D

- physiological properties topological properties
- electrotopological properties
- 2D fingerprints









3D

- geometrical properties
- surface properties
- grid properties

 $TPSA = 18.5 (Å^2)$

VECTOR EMBEDDING

0.8	0.1	0.3	• • •	0.2	0.2
-----	-----	-----	-------	-----	-----





 \sim



0.7

		N	DD	; 1I 3	5 NDI	EX							
		1	2	3	4				BO	ND) TY	Έ	
×	1		1	\cap	0		S S		_ _ _	2	3 (arom	
E			-	1	0	•••	<u></u>	(1, 2)	0	0	0	1	
Ż	Z		0	I	0	•••		(2,3)	0	0	0	1	
ш.	3	0	1	0	1	•••		(-, -)		0	0		
0	4	0	0	1	0	•••	L L L L	(3, 4)		0	0		
ž	••••	•••	•••	•••	•••	•••	ED ((,)	L	•••	•••		
2	dia	- cer		ma	trix	· (Δ)	ec	lge fea	atur	ſes	ma	trix (E	:)

0

0

0

...

		С	0
EX	1	1	0
	2	1	0
<u> </u>	3	1	0
10 D	4	1	0
2	•••		•••

node features matrix (X)

••• ••• ••• •••

0 ... 0 1 0 1

... 0 1 0 1

... 0 1 0 0

0 0

0 0

0 0

••• •••

0

0

0 1 0 1 0 0 0



in silico tools

• fixed representations

1D

- number of atoms
- number of bonds
- molecular weight
- atomic properties

• deep representations

LINEAR NOTATIONS

- physiological properties
- topological properties
- electrotopological properties
- 2D fingerprints

MACCS keys, ECFP, etc.

0	0	1	•••	1	0	0



- geometrical properties
- surface properties
- grid properties

0.8	0.1	0.3	 0.2	0.2



MOLECULAR GRAPHS





	1		
1			
2			
			•••
•••		 	



-1						
2		•••				
		•••				



in silico tools

BINARY MOLECULAR FINGERPRINT





K. Dührkop et al., Nat. Methods. (2019).



K. Dührkop et al., Nat. Methods. (2019).







K. Dührkop et al., Nat. Methods. (2019).



$C_{12}H_4CI_4O_2$

C₁₀H₃Cl₃O₂





K. Dührkop et al., Nat. Methods. (2019).







K. Dührkop et al., Nat. Methods. (2019).





K. Dührkop et al., Nat. Methods. (2019).



K. Dührkop et al., Nat. Methods. (2019).



K. Dührkop et al., Nat. Methods. (2019).

PROBABILISTIC MOLECULAR FINGERPRINT

0.45	0.99	•••	0.71	0.27	0.08
------	------	-----	------	------	------



K. Dührkop et al., Nat. Methods. (2019).

BINARY MOLECULAR FINGERPRINT

0	1	•••	1	0	0
---	---	-----	---	---	---

PROBABILISTIC MOLECULAR FINGERPRINT

0.45	0.99	•••	0.71	0.27	0.08
------	------	-----	------	------	------




AFE



13







I. Rahu et al., J. Chem. Inf. Model. (2024).

ENDOCRINE DISRUPTION

Classification task







MS2Tox: DATA



ENDOCRINE DISRUPTION

CHEMICAL 12 BIOASSAYS

TOX21 11764 instances

ENDPOINT

DATASET

CHEMICAL

1. Data cleaning



P. Peets et al., Environ. Sci. Tech. (2022).





P. Peets et al., Environ. Sci. Tech. (2022).



P. Peets et al., Environ. Sci. Tech. (2022).



- 4. Train/intermediate test set data preprocessing



P. Peets et al., Environ. Sci. Tech. (2022).

















- 4. Train/intermediate test set data preprocessing



P. Peets et al., Environ. Sci. Tech. (2022).

I. Rahu et al., J. Chem. Inf. Model. (2024).



FINGERPRINT FEATURES

SELECTING THE MODELS FOR FINAL EVALUATION

 $RMSE = 0.79...1.12 \log - mM$ OREAL-LIFEC50 LC_{50} fish water flea (static)























PROBABILISTIC MOLECULAR FINGERPRINT

0.14 0.45 0.99	•••	0.71	0.27	0.08
----------------	-----	------	------	------

P. Peets et al., Environ. Sci. Tech. (2022).



PROBABILISTIC MOLECULAR FINGERPRINT



P. Peets et al., Environ. Sci. Tech. (2022).









5. HRMS data preprocessing



RMSE = 0.79...1.26 log-mM EC_{50} **EC**₅₀ water flea algae

FINGERPRINT FINAL **HRMS DATA** FEATURES

© REAL-LIFE

PROBABILISTIC MOLECULAR FINGERPRINT

9 0.71 0.27 0.0



PROBABILISTIC MOLECULAR FINGERPRINT





PROBABILISTIC MOLECULAR FINGERPRINT



I. Rahu et al., J. Chem. Inf. Model. (2024).

IN prediction i 1 N FINAL PREDICTION 0.77



FPR at 90% recall = 0.25...0.85**AHR** ER



27



SAFE

MS2Quant: ESI



J. Liigand et al., *Sci. Rep.* (2020). L. Malm et al., *Molecules*. (2021).







J. Liigand et al., *Sci. Rep.* (2020). L. Malm et al., *Molecules*. (2021).



J. Liigand et al., *Sci. Rep.* (2020). L. Malm et al., *Molecules*. (2021).







J. Liigand et al., *Sci. Rep.* (2020). L. Malm et al., *Molecules*. (2021).



 $logRF_{pred} = slope \cdot log/E_{pred} + intercept$



J. Liigand et al., *Sci. Rep.* (2020). L. Malm et al., *Molecules*. (2021).



$$c = \frac{\text{peak area}}{10}$$

MS2Quant



H. Sepman et al., Anal. Chem. (2023).



MS2Quant: WORKFLOW





CHEMICAL

log/E

1191 unique chemicals 13 different instruments



H. Sepman et al., Anal. Chem. (2023).

logRF_{pred} = slope · logIE_{pred} + intercept

 $c = \frac{\text{peak area}}{10^{\text{logRF}_{\text{pred}}}}$







H. Sepman et al., Anal. Chem. (2023).

32


FIELD OF APPLICATION





INTERPRETABILITY



FINGERPRINT FEATURES CONTAINING N



FIELD OF APPLICATION





CASE STUDY: RECYCLED TEXTILES

priority score = -



D. Szabo et al., Anal. Chem. (2024).

predicted concentration (mM) predicted aquatic LC₅₀ (mM)

CASE STUDY: RECYCLED TEXTILES

priority score = -

D. Szabo et al., Anal. Chem. (2024).

predicted concentration (mM) predicted aquatic LC₅₀ (mM)

MS2Risk in TOP 10 by priority score tris(2-ethylhexl)phosphate 37

CASE STUDY: INTERLABORATORY COMPARISON

I. Rahu, G. Sandberg et al., in preparation

38

CASE STUDY: INTERLABORATORY COMPARISON

labeled as active

I. Rahu, G. Sandberg et al., in preparation

FUTURE PERSPECTIVE: UPDATING MODELS

I. Rahu et al., in preparation

39

FUTURE PERSPECTIVE: MS2Quant ESI-

40

FUTURE PERSPECTIVE: MOLECULAR NETWORKS

Y. Kreutzer et al., in preparation

FUTURE PERSPECTIVE: MOLECULAR NETWORKS

Y. Kreutzer et al., in preparation

FUTURE PERSPECTIVE: MOLECULAR NETWORKS

Y. Kreutzer et al., in preparation

Swedish Research Council

CARL TRYGGERS STIFTELSE

FÖR VETENSKAPLIG FORSKNING

FORMAS

Wenner-Gren Foundations Wenner-Gren Stiftelserna

Wallenberg Initiative **Materials Science** for Sustainability

Exploring the research space...

the second second

MS2Tox & MS2Quant

Automated detection and identification of TOXIC and HIGH-RISK chemicals

