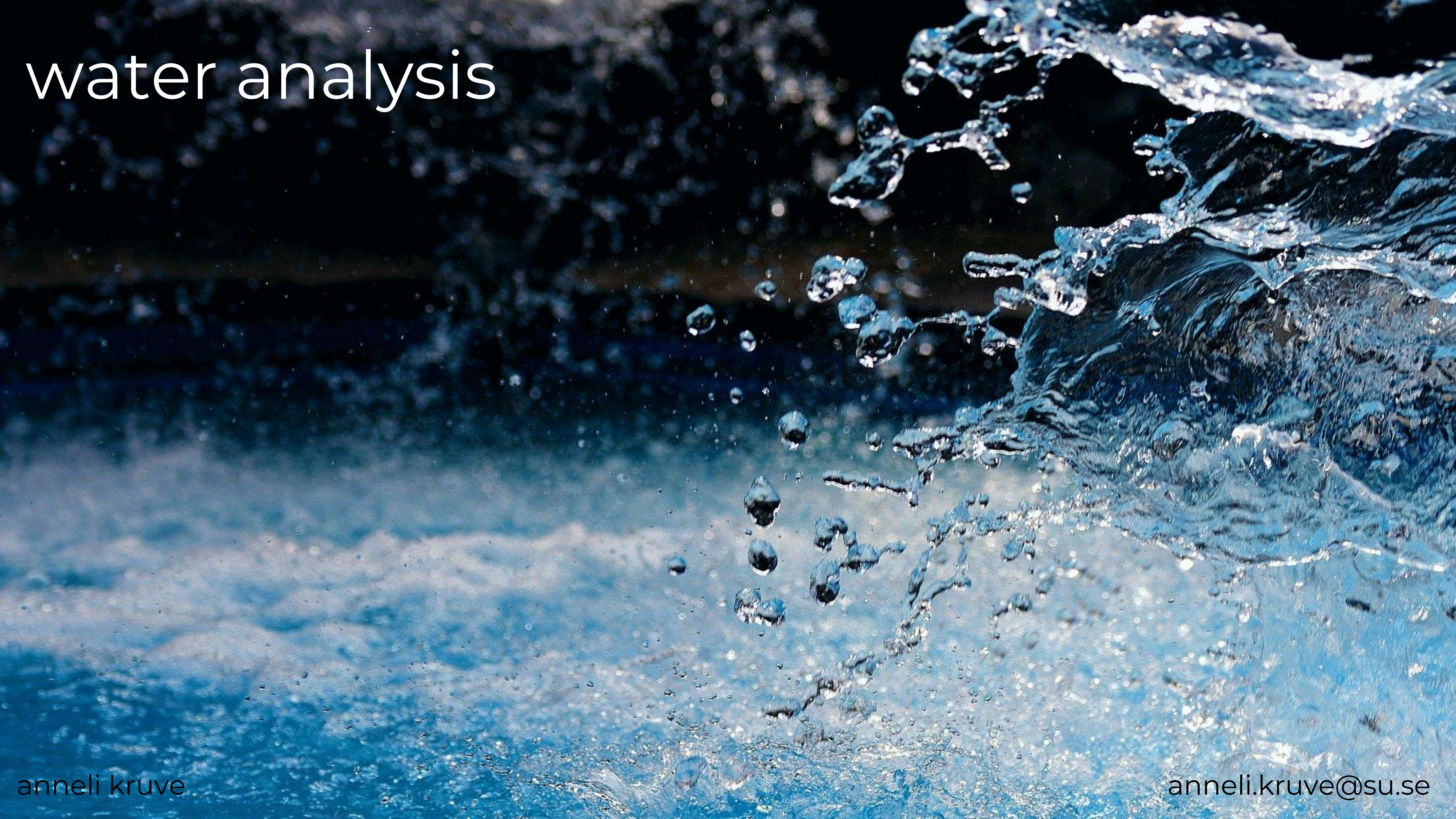
MS2Tox & MS2Quant: automated prediction of toxicity and concentration

anneli kruve

anneli.kruve@su.se

kruvelab.com

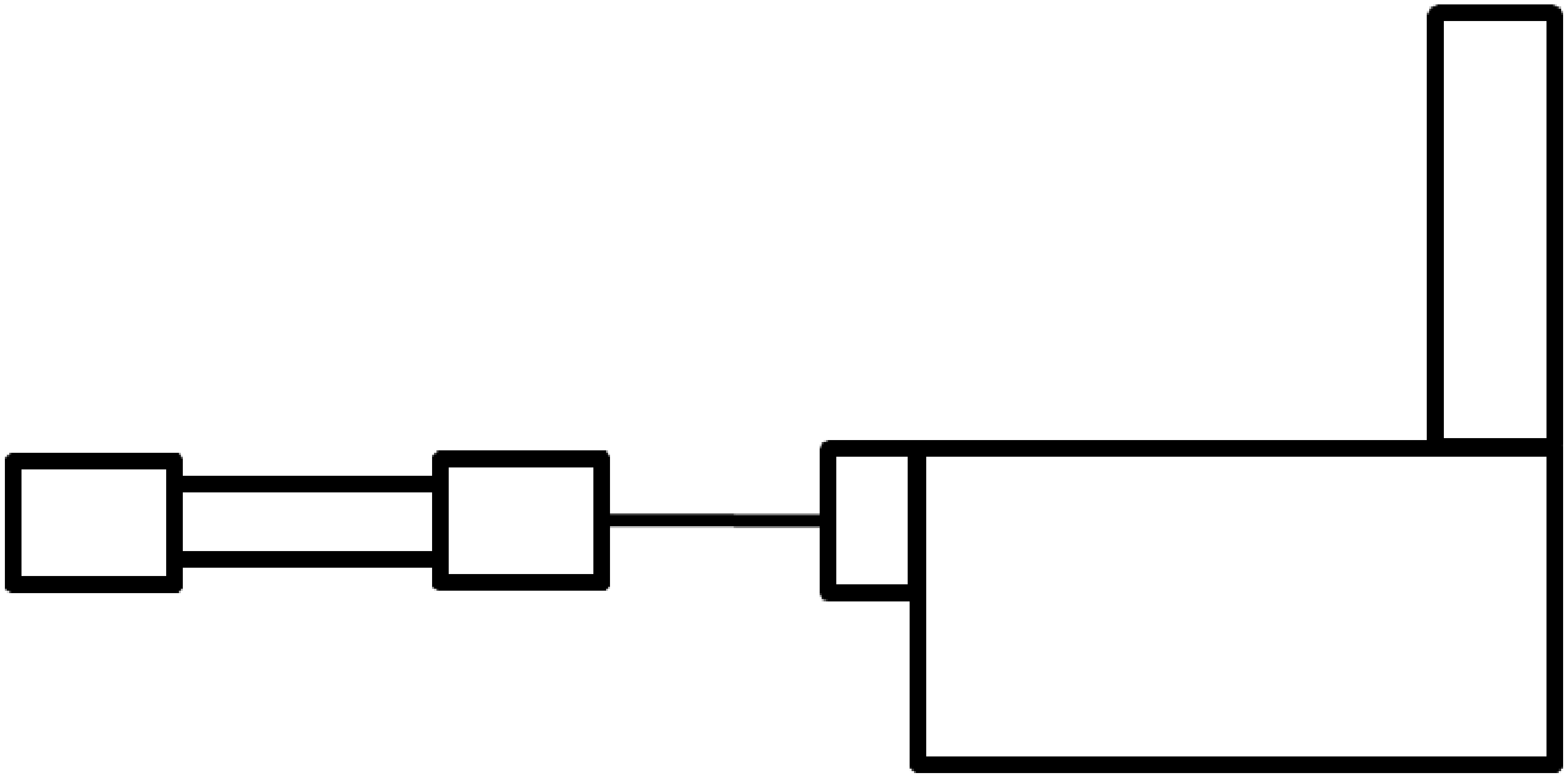
water analysis



anneli kruve

anneli.kruve@su.se

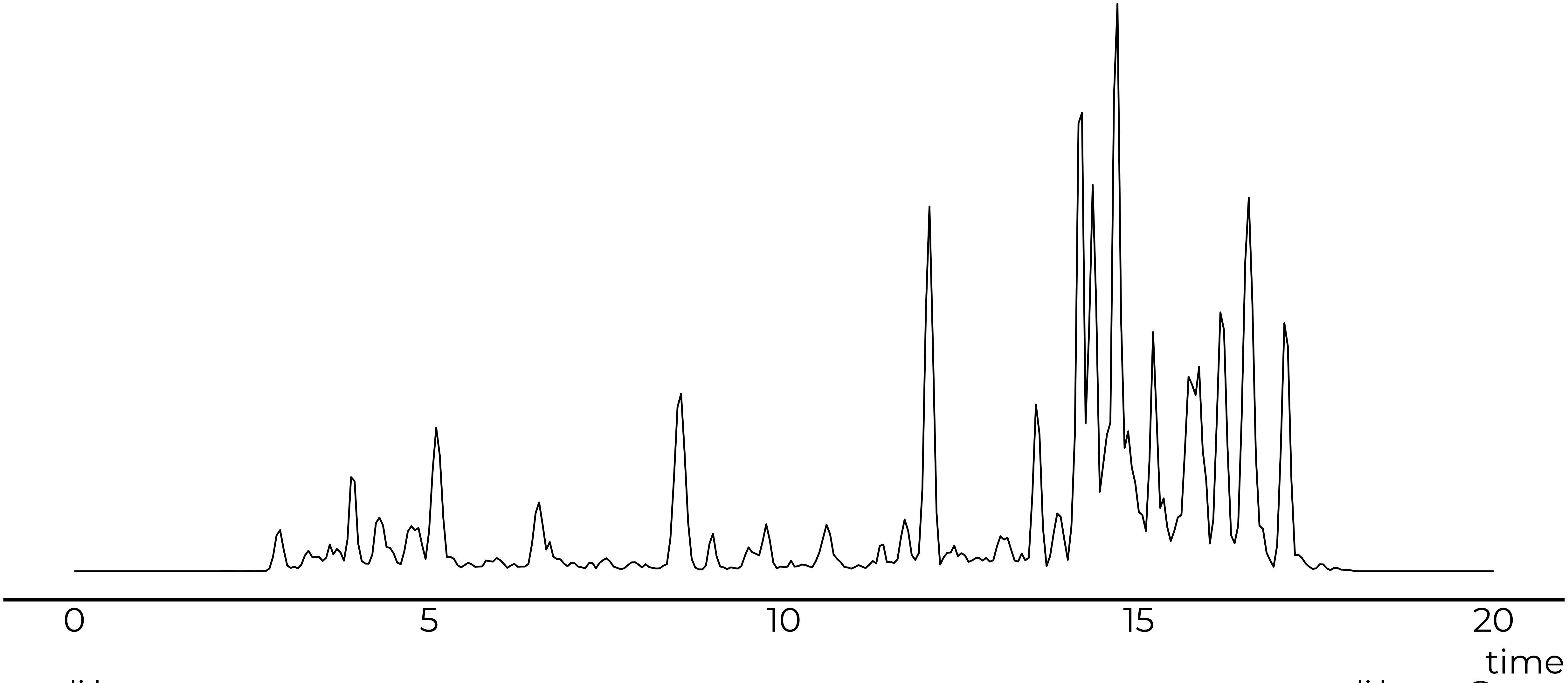
nontarget screening with LC/HRMS



anneli kruve

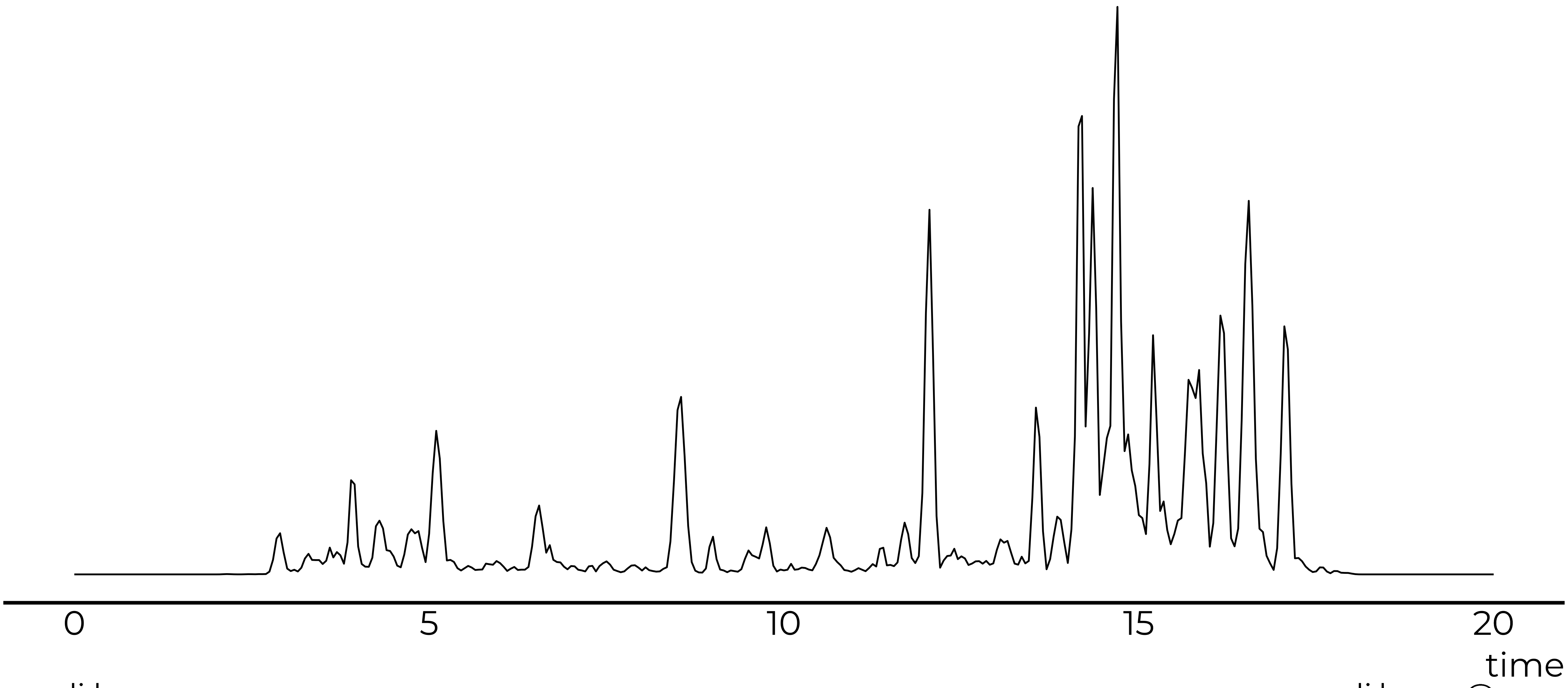
anneli.kruve@su.se

nontarget screening with LC/HRMS



anneli kruve
anneli.kruve@su.se

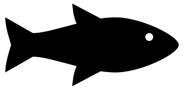
what next?



anneli kruve

anneli.kruve@su.se

prioritization

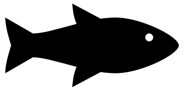


toxicity

anneli kruve

anneli.kruve@su.se

prioritization



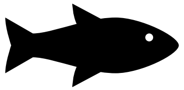
toxicity

concentration

anneli kruve

anneli.kruve@su.se

prioritization



toxicity

concentration

risk

anneli kruve

anneli.kruve@su.se

prioritization

toxicity

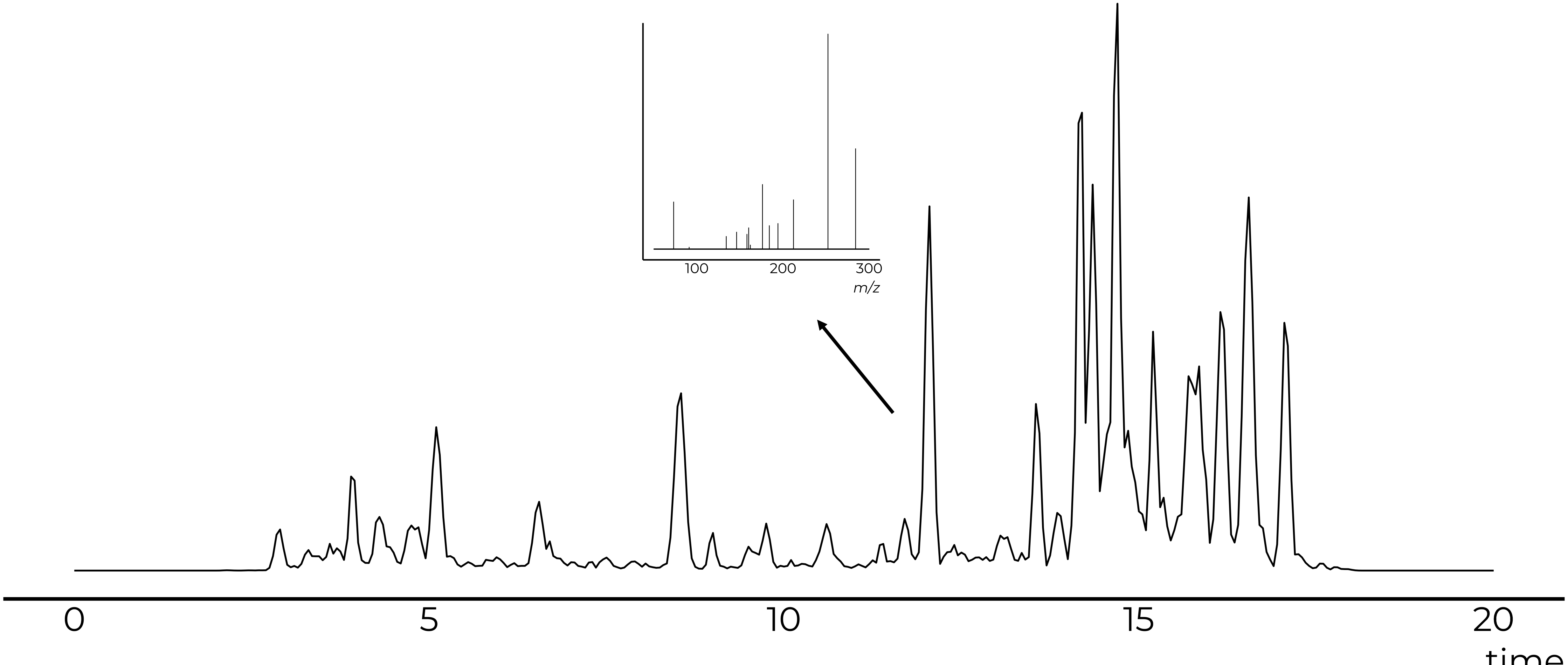
concentration

risk $$\mathrm{PriorityScore} = \frac{c_{\mathrm{predicted}}}{AC_{50}^{\mathrm{5th\ percentile}}}$$

anneli kruve

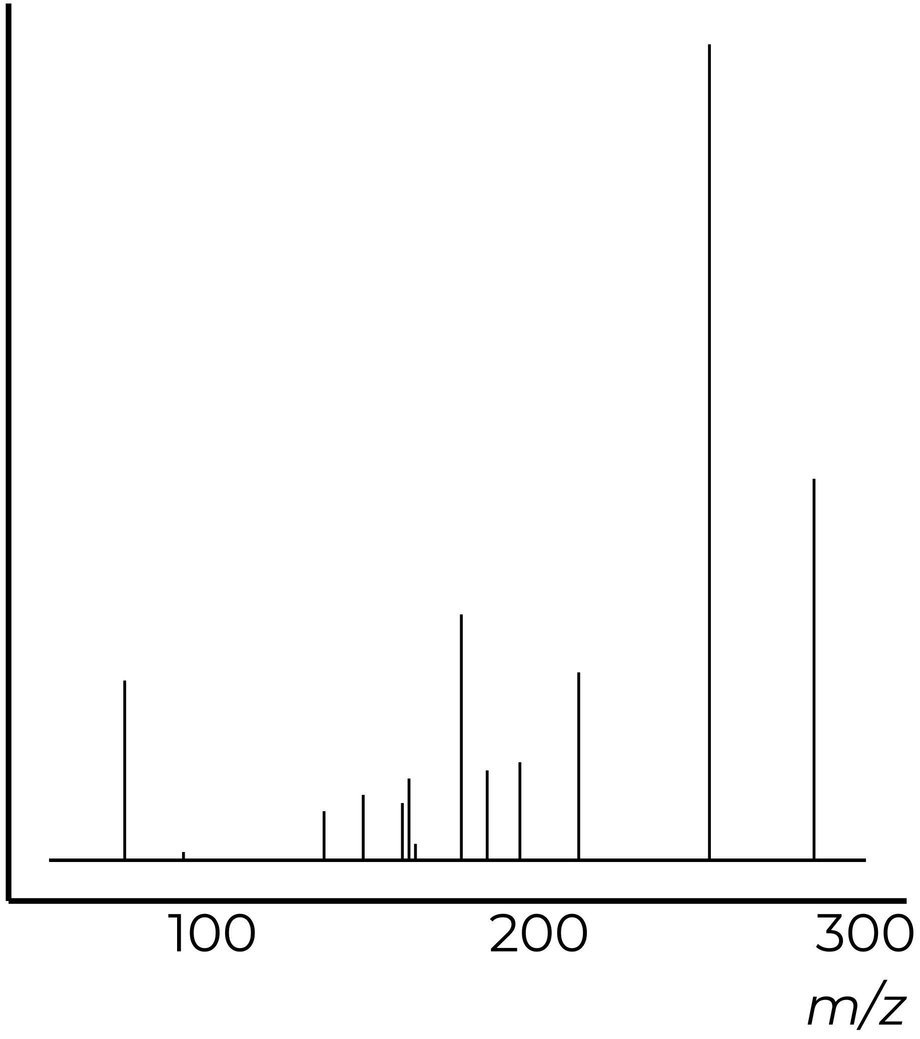
anneli.kruve@su.se

nontarget screening with LC/HRMS

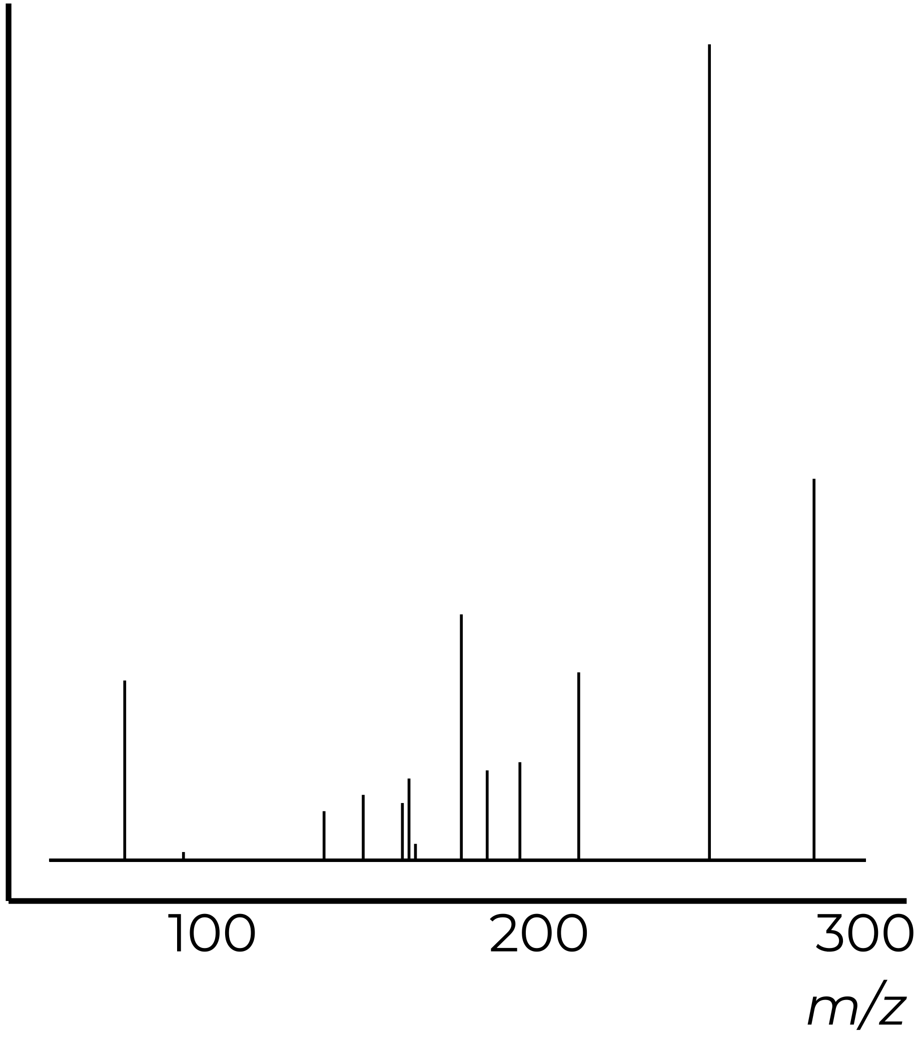


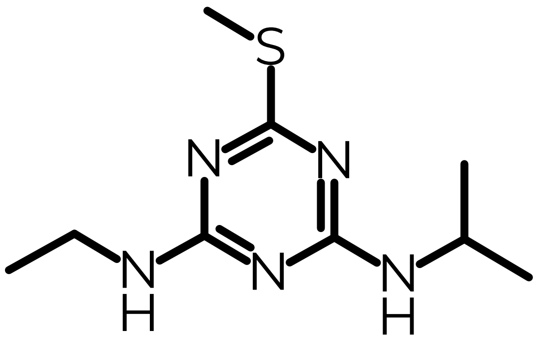
anneli kruve
anneli.kruve@su.se

toxicity assessment



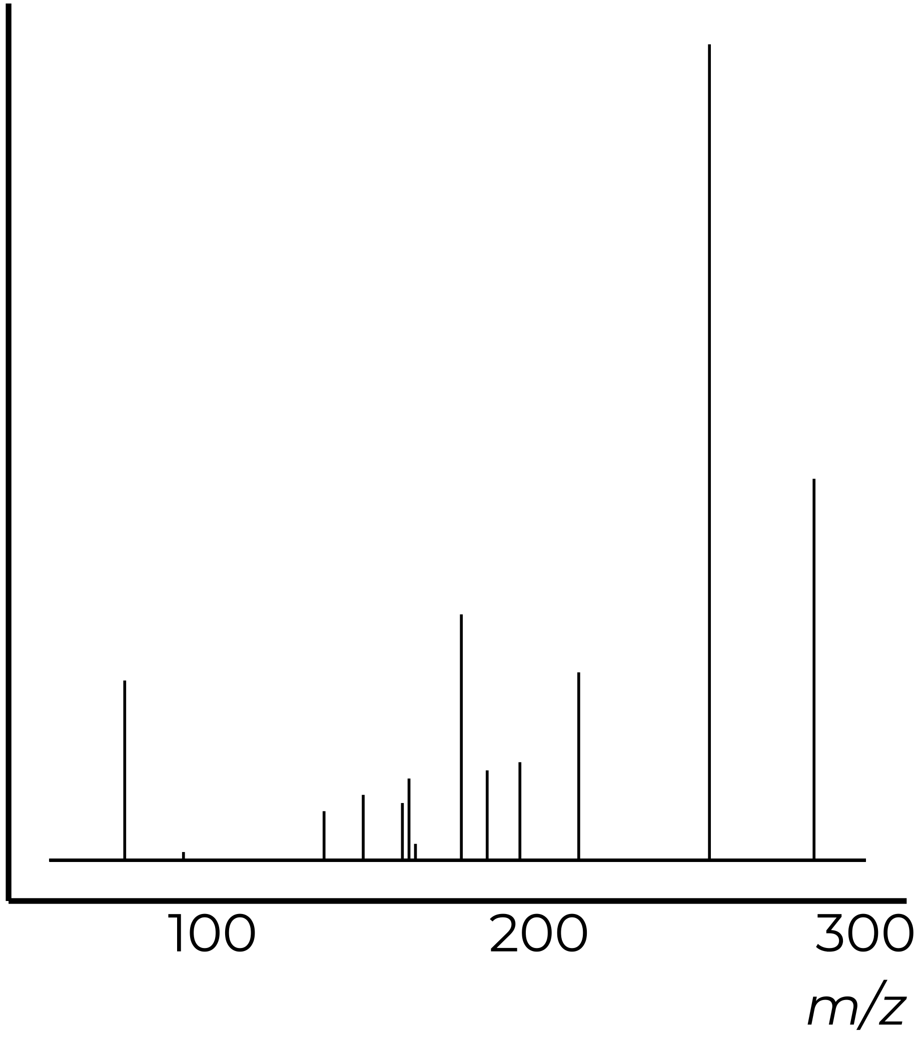
toxicity assessment

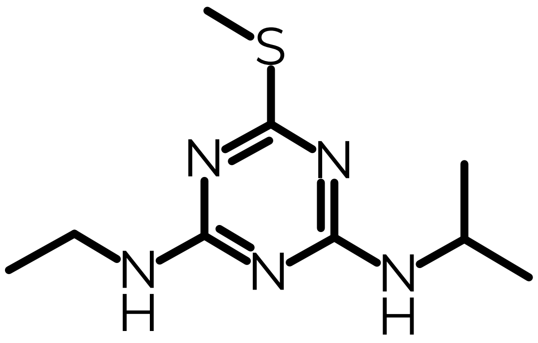




anneli.kruve@su.se

toxicity assessment



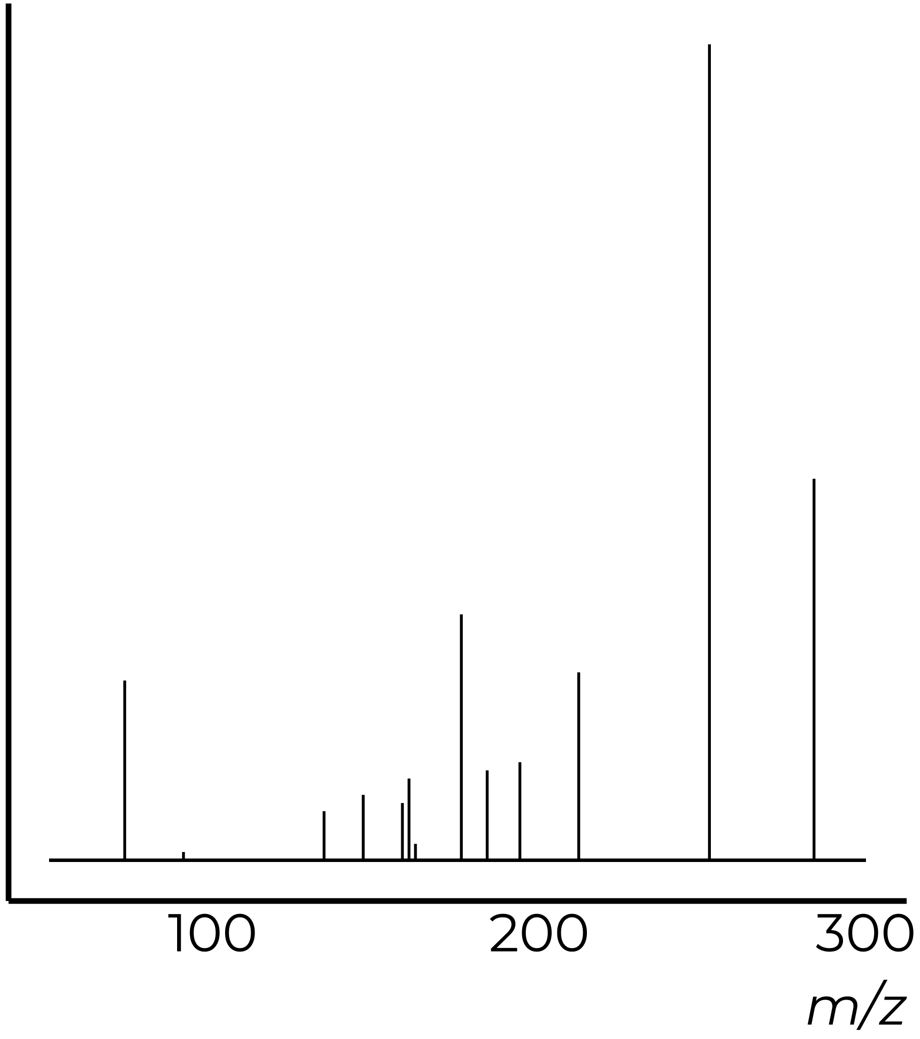


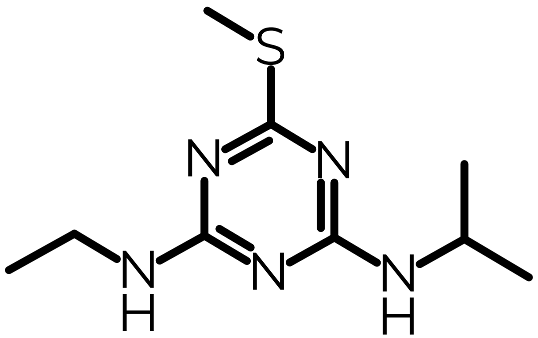
LC_{50} = 9.3 mg/L

anneli kruve

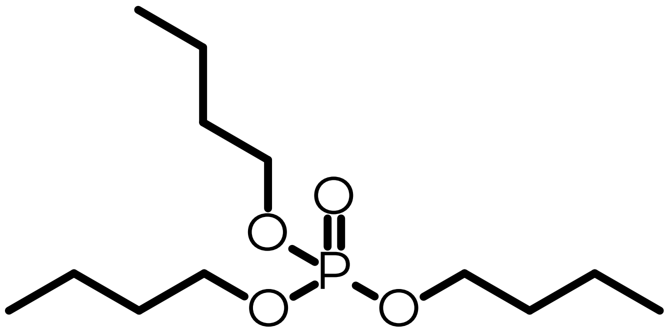
anneli.kruve@su.se

toxicity assessment





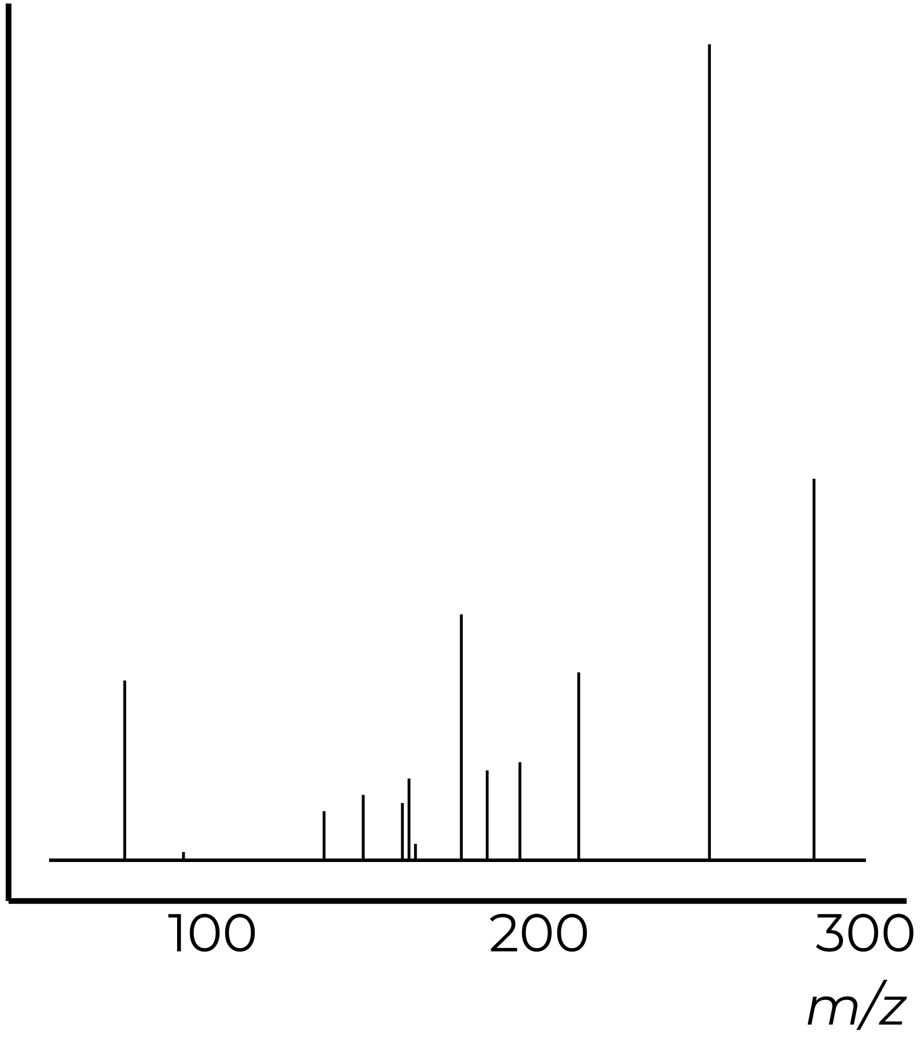
LC_{50} = 9.3 mg/L

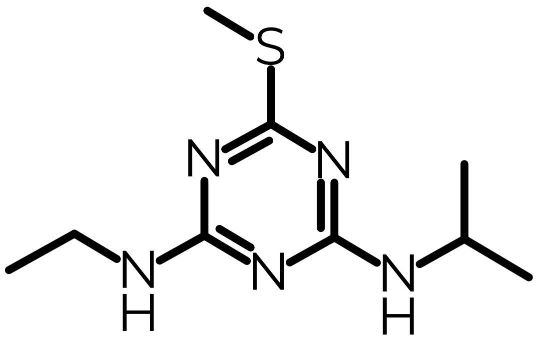


anneli kruve

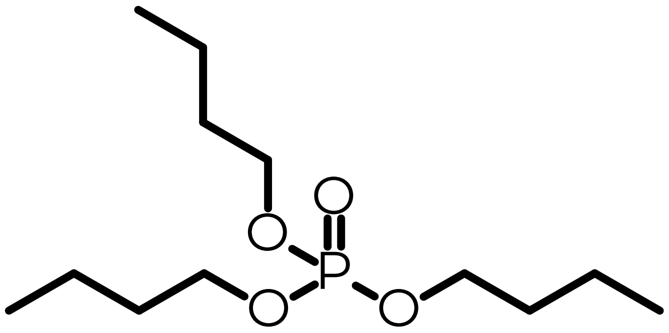
anneli.kruve@su.se

toxicity assessment





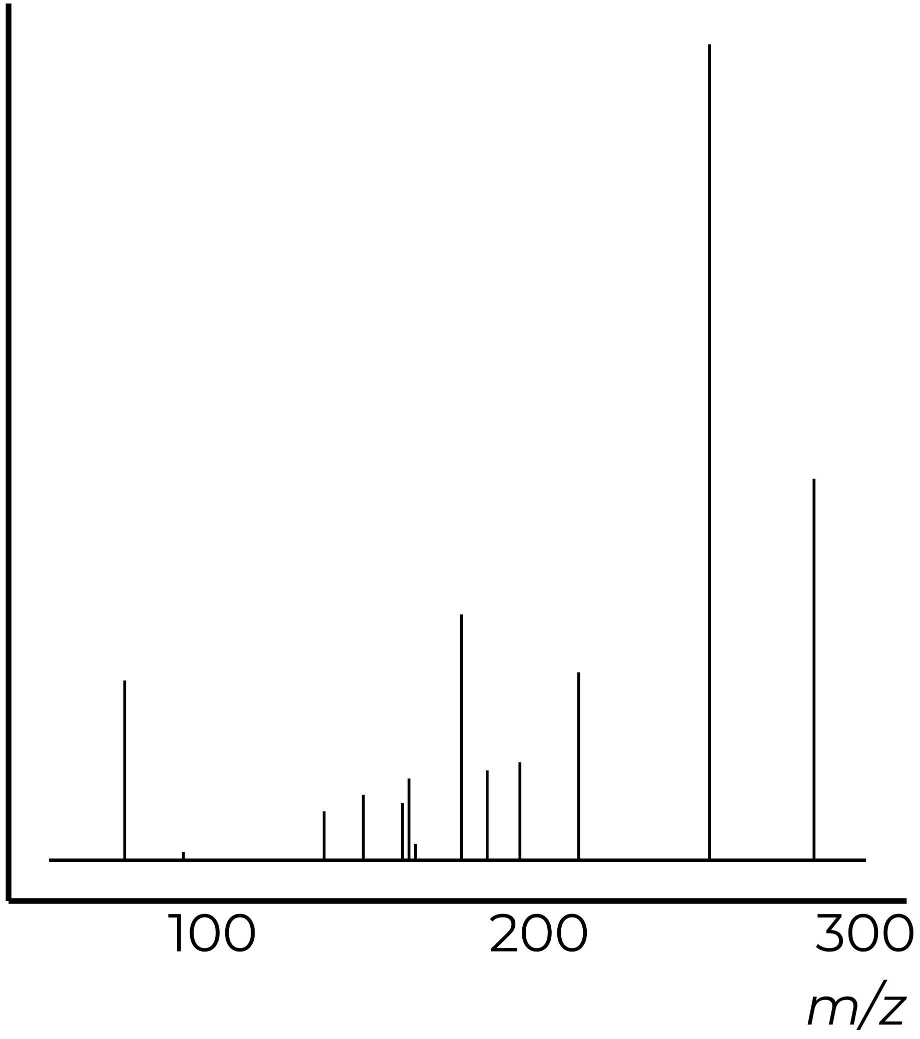
LC_{50} = 9.3 mg/L

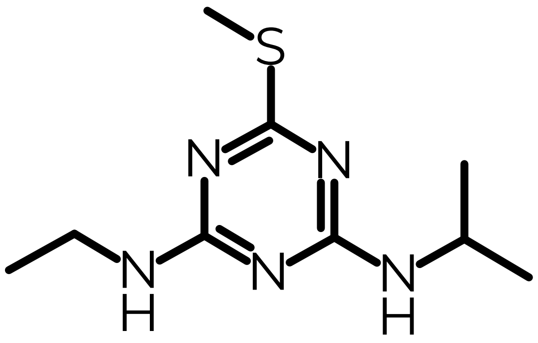


LC_{50} = ? mg/L

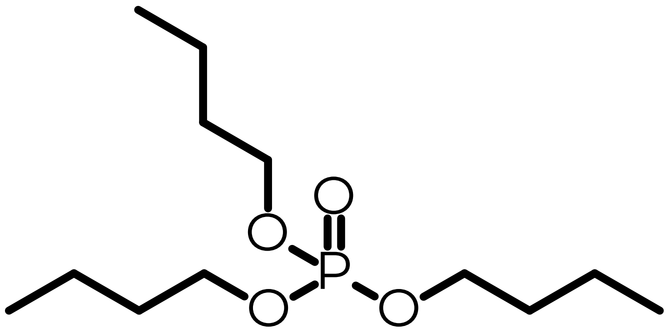
anneli.kruve@su.se

toxicity assessment





LC_{50} = 9.3 mg/L



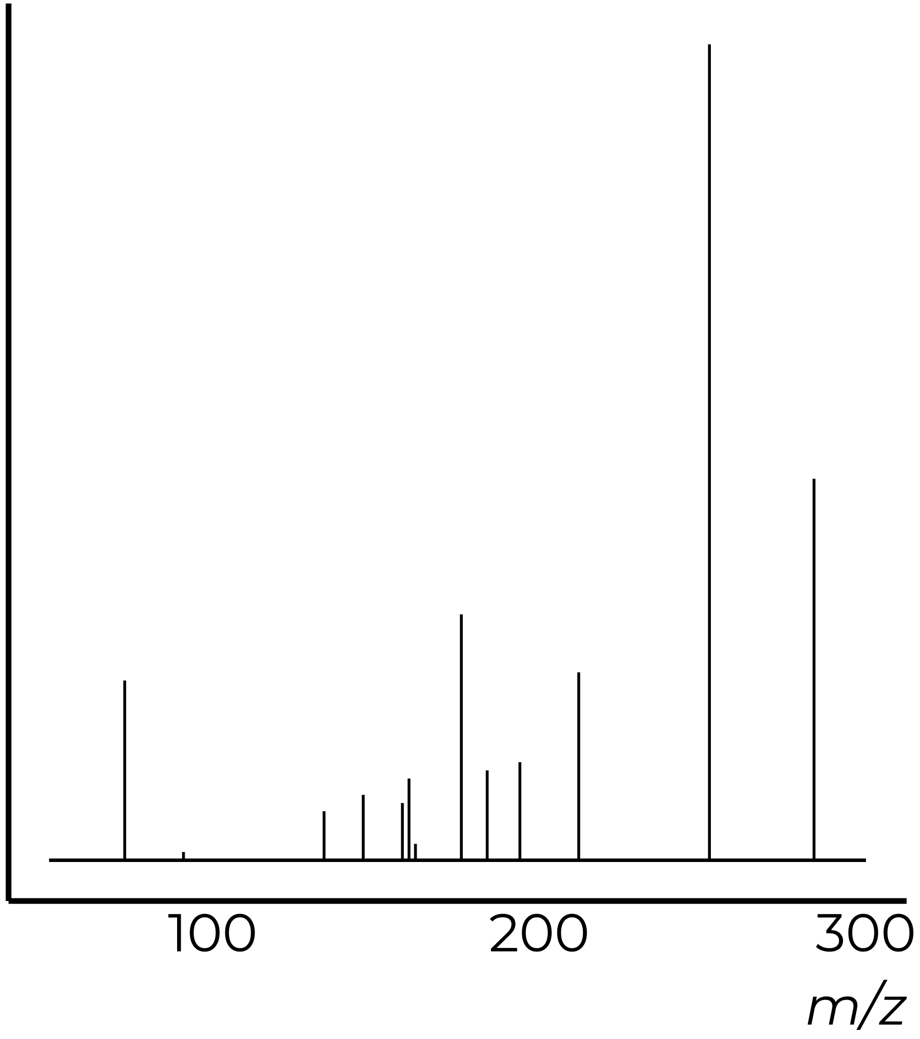
LC_{50} = ? mg/L

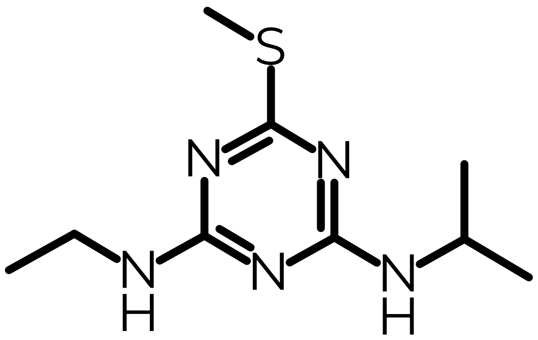
?

anneli kruve

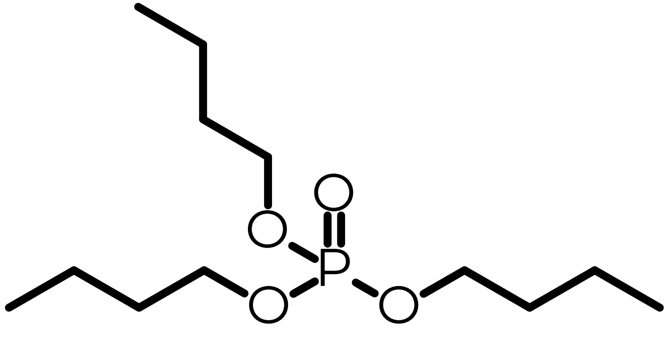
anneli.kruve@su.se

toxicity assessment





LC_{50} = 9.3 mg/L



LC_{50} = ? mg/L

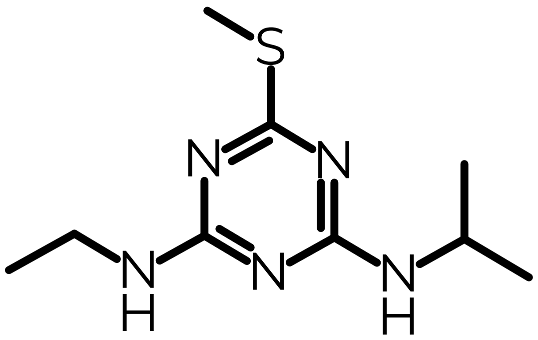
?

LC_{50} = ? mg/L

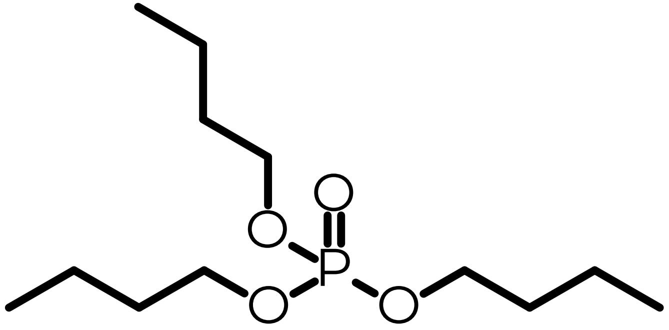
anneli kruve

anneli.kruve@su.se

toxicity assessment



LC_{50} = 9.3 mg/L



LC_{50} = ? mg/L

?

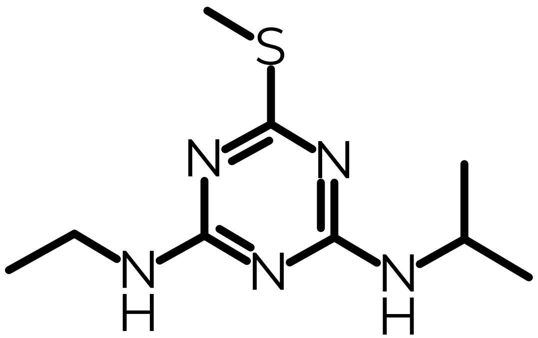
LC_{50} = ? mg/L

anneli kruve

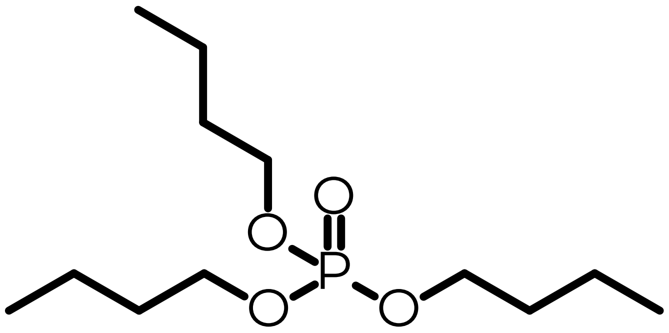
anneli.kruve@su.se

toxicity assessment

<1%



LC_{50} = 9.3 mg/L



LC_{50} = ? mg/L

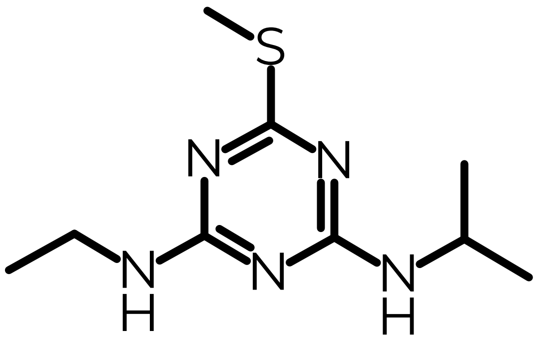
?

LC_{50} = ? mg/L

anneli.kruve@su.se

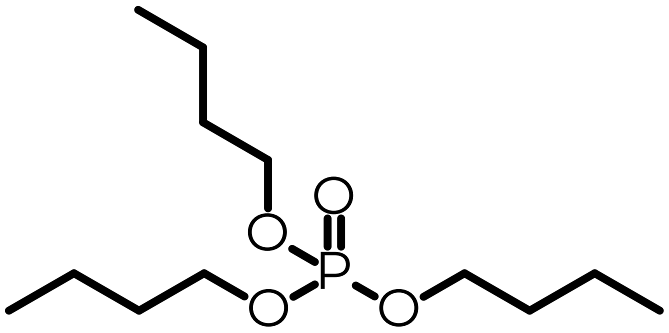
toxicity assessment

<1%



LC$_{50}$ = 9.3 mg/L

<2%



LC$_{50}$ = ? mg/L

?

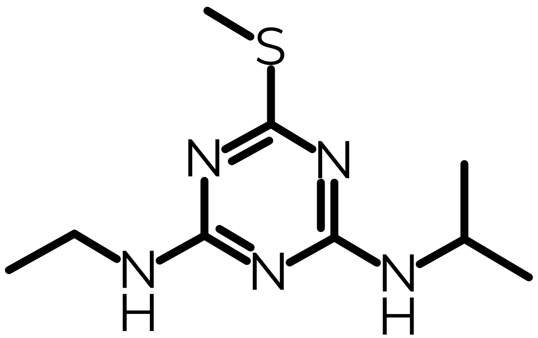
LC$_{50}$ = ? mg/L

anneli kruve

anneli.kruve@su.se

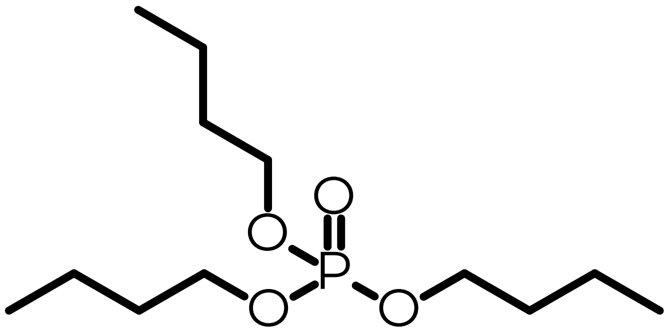
toxicity assessment

<1%



LC_{50} = 9.3 mg/L

<2%



LC_{50} = ? mg/L

~98%

?

LC_{50} = ? mg/L

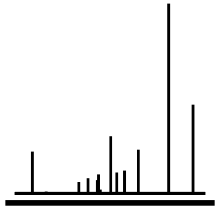
anneli kruve

anneli.kruve@su.se

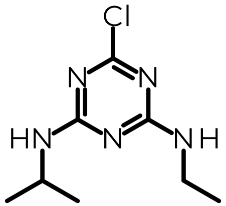
predicting toxicity

for detected chemicals

workflow

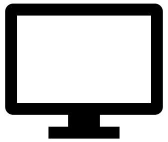


MS^2 spectra



structure as SMILES

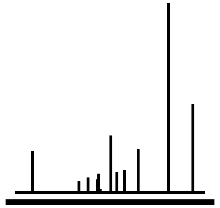
molecular descriptors



predict toxicity

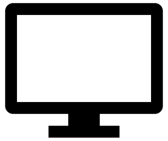
anneli kruve
anneli.kruve@su.se

workflow



MS^2 spectra

molecular descriptors



predict toxicity

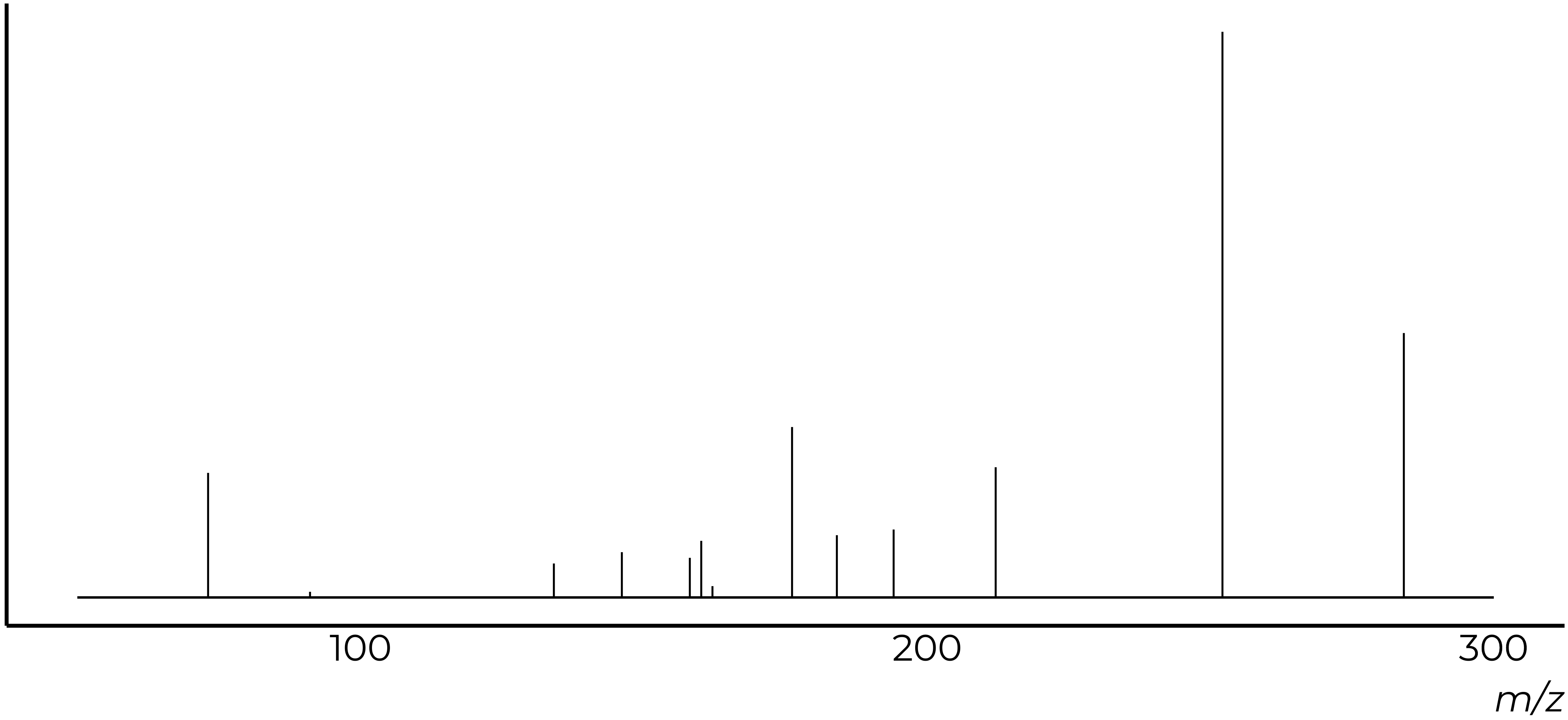
anneli kruve

anneli.kruve@su.se

information available

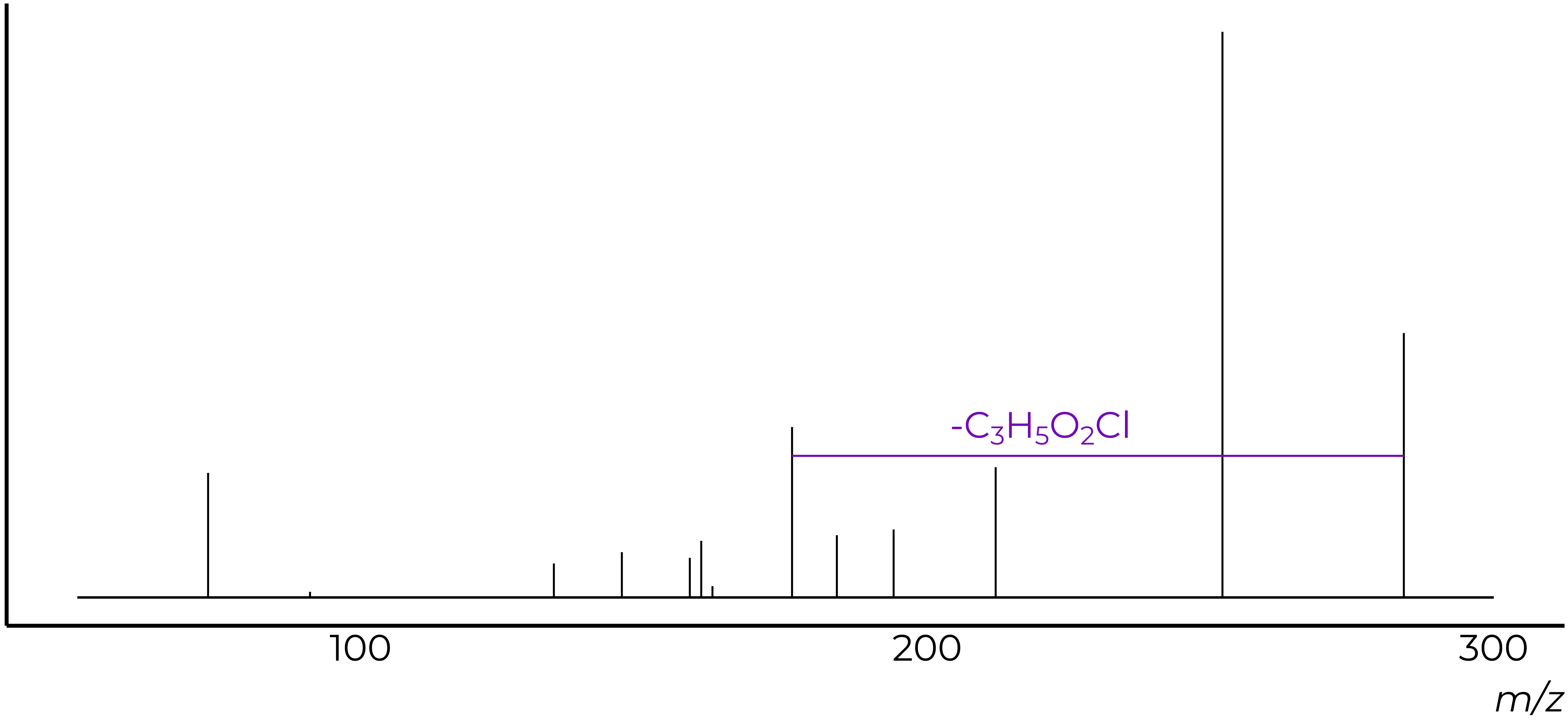
in MS^2 spectra

MS² spectra



anneli.kruve@su.se

MS^2 spectra



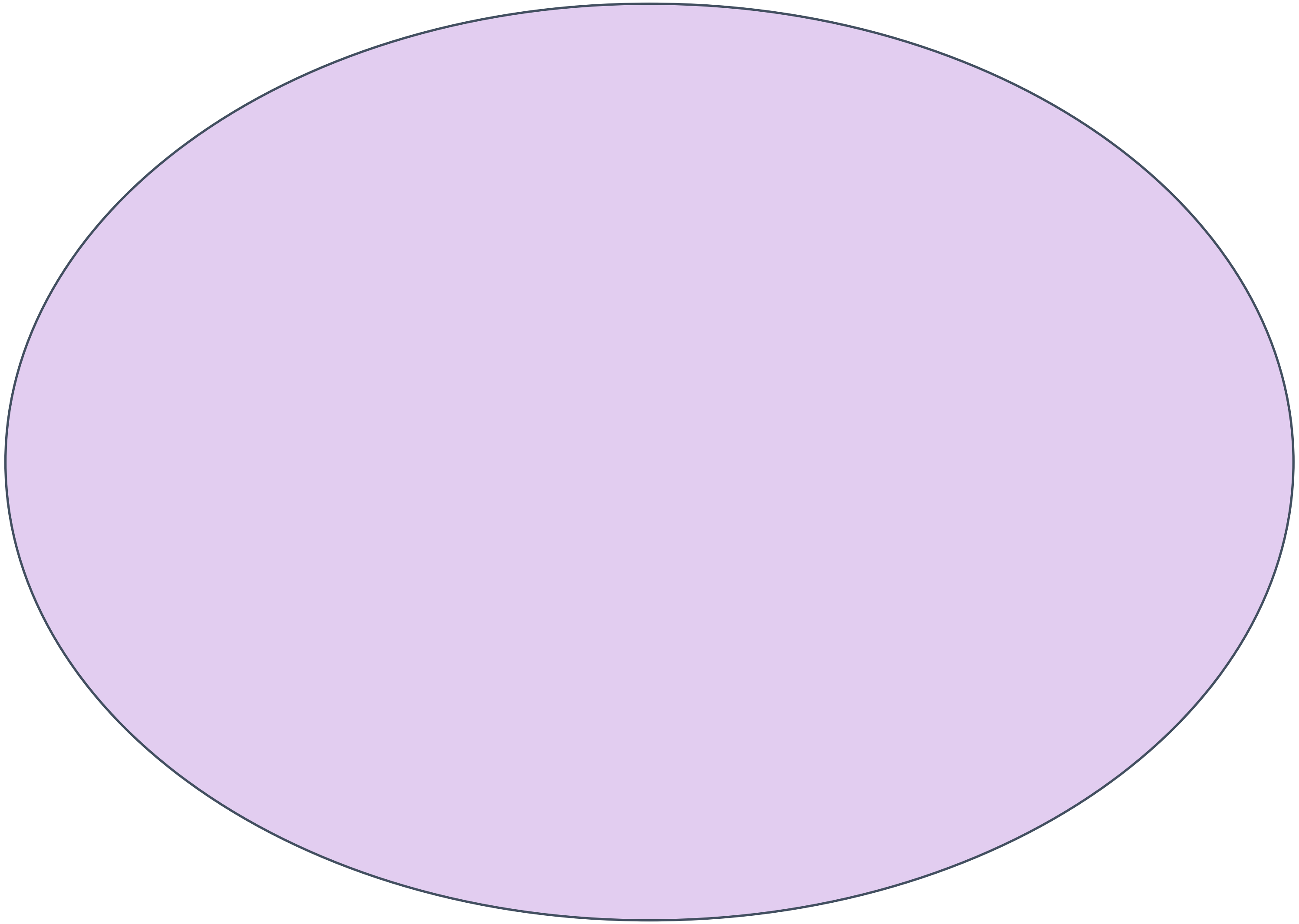
anneli.kruve@su.se

data for machine learning models

anneli.kruve@su.se

data for machine learning models

CompTox



all toxicity values

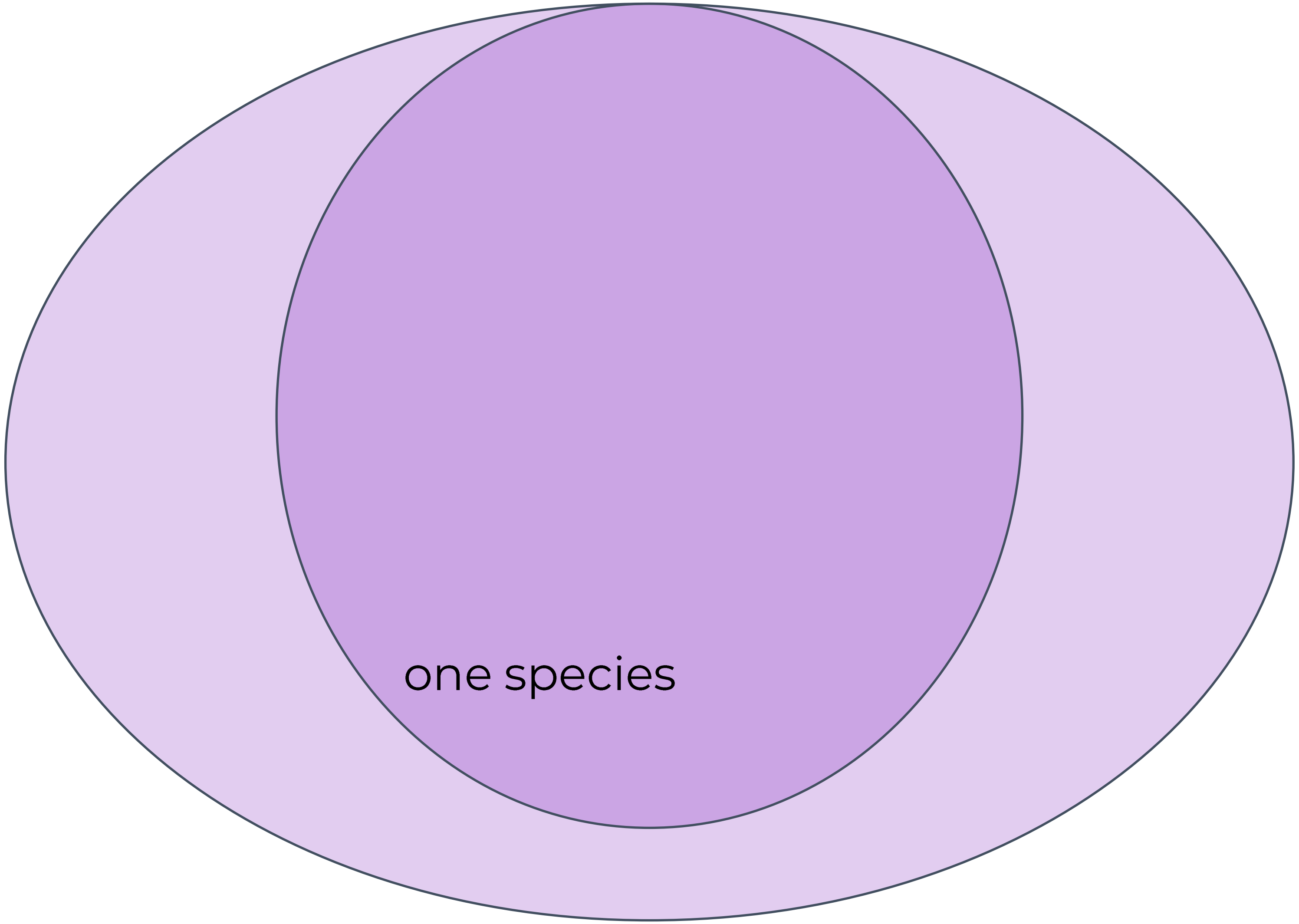
anneli kruve

anneli.kruve@su.se

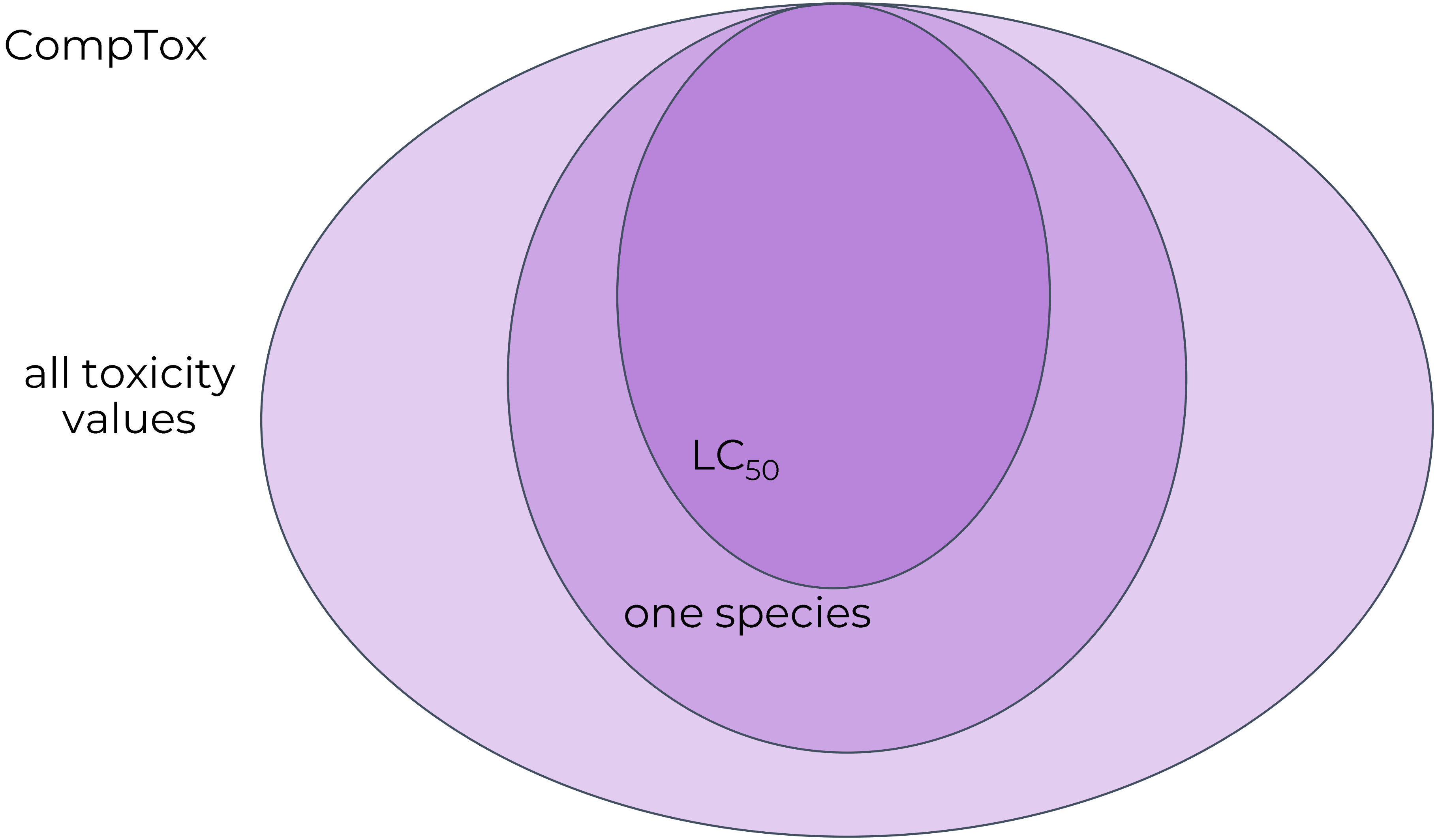
data for machine learning models

CompTox

all toxicity values



data for machine learning models

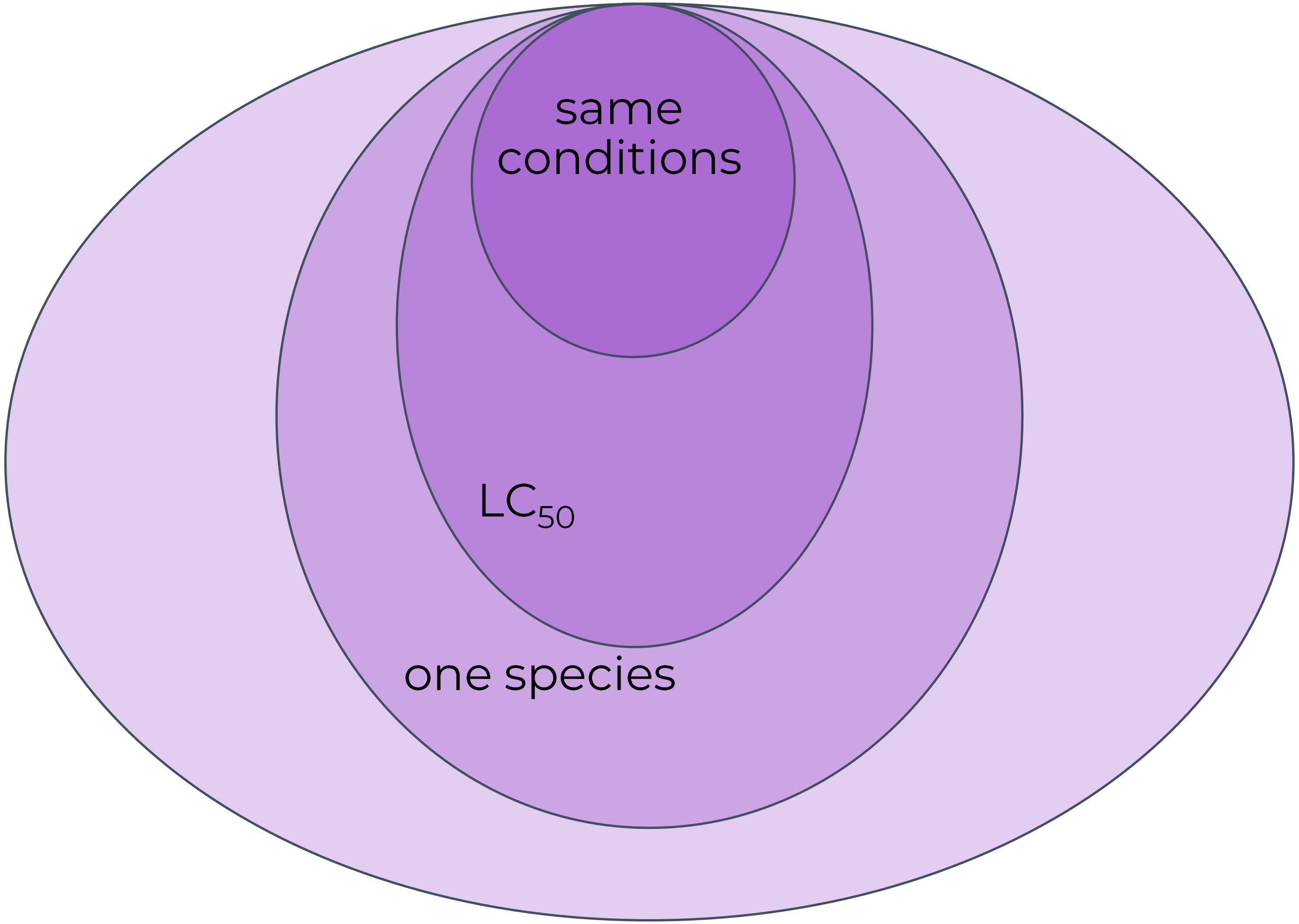


anneli kruve
anneli.kruve@su.se

data for machine learning models

CompTox

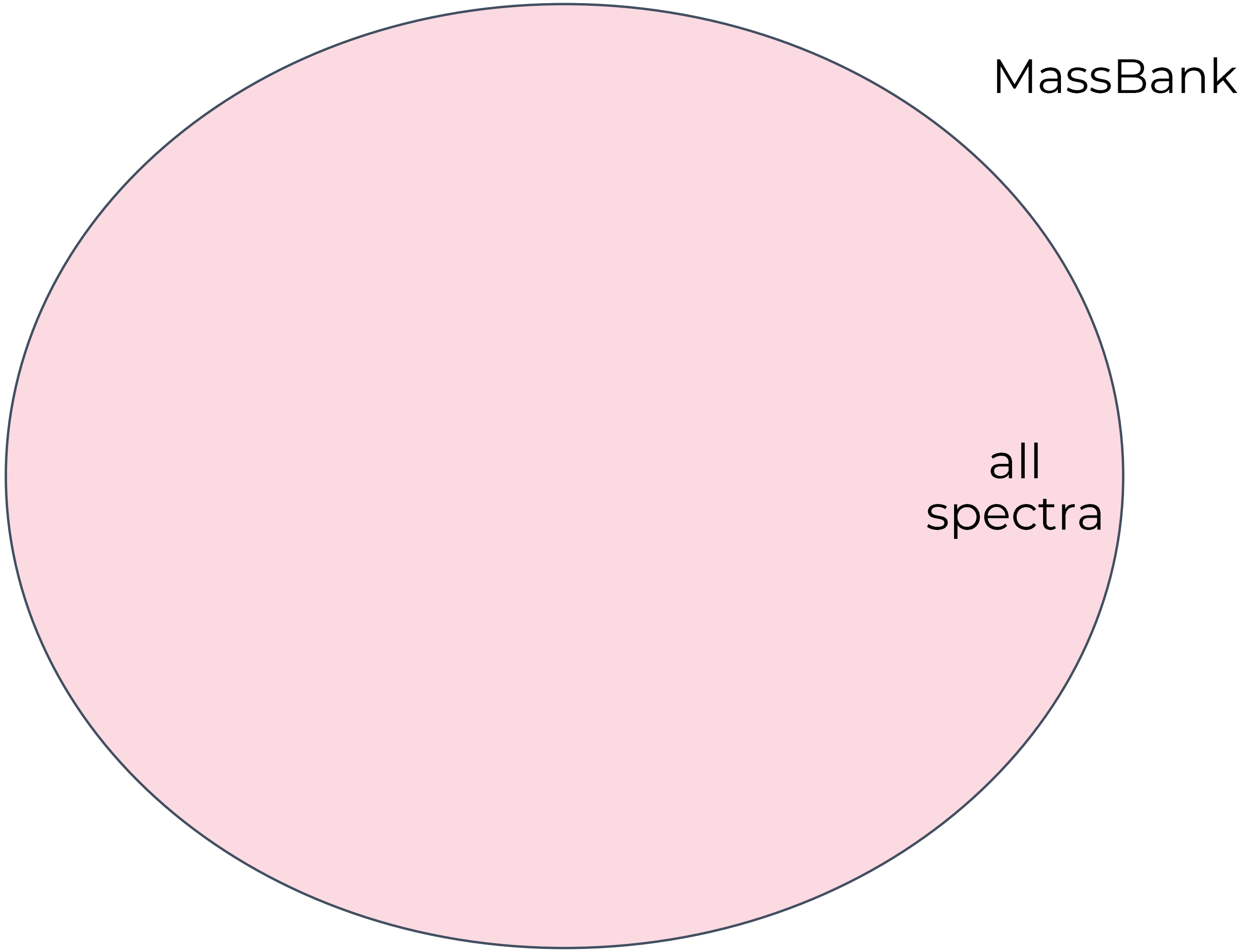
all toxicity values



anneli kruve

anneli.kruve@su.se

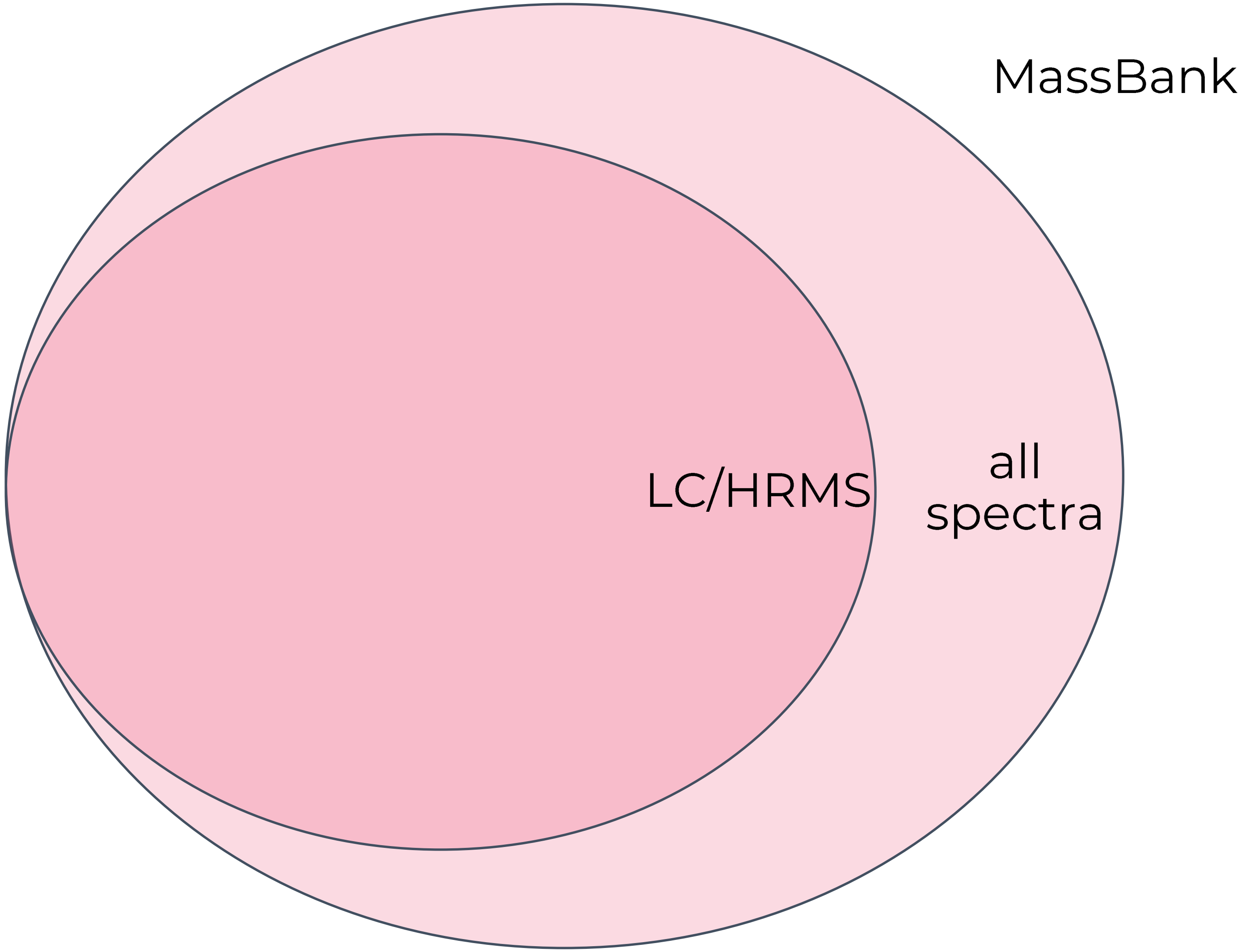
data for machine learning models



anneli kruve

anneli.kruve@su.se

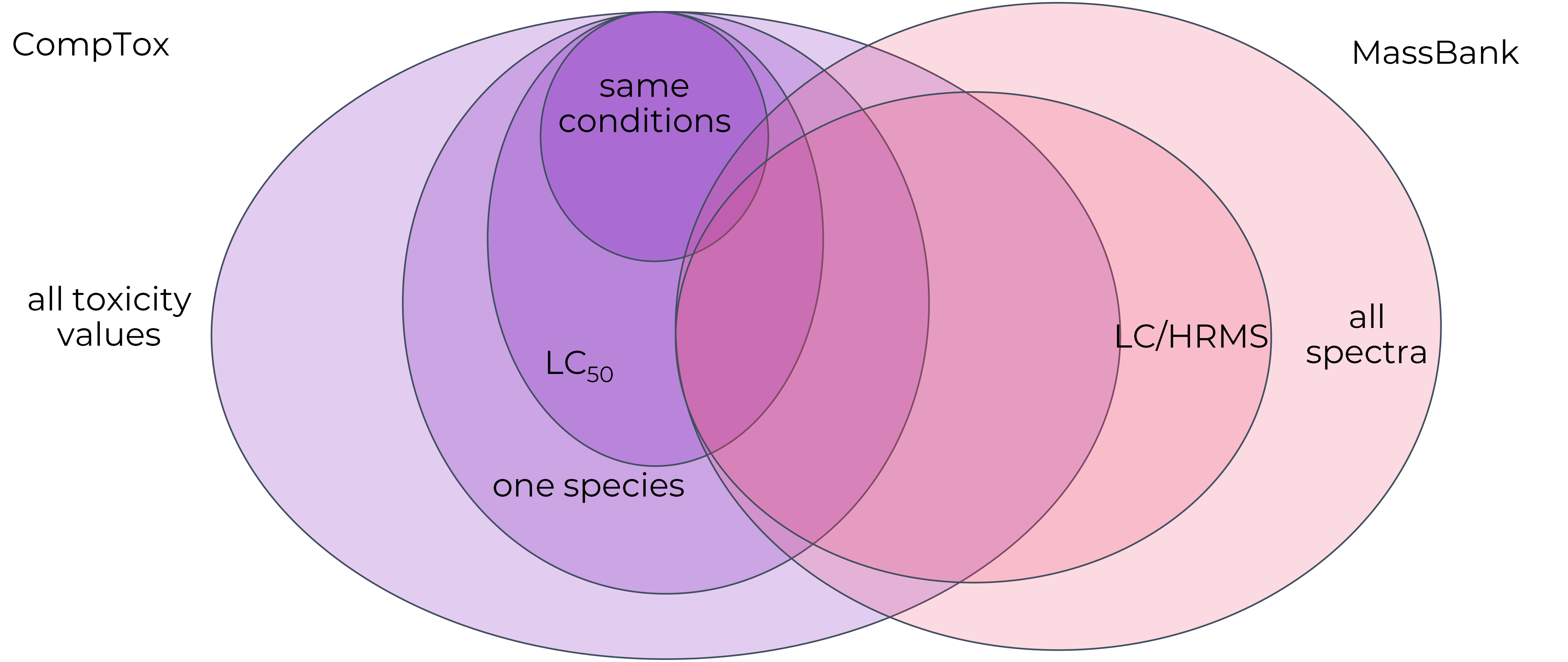
data for machine learning models



anneli kruve

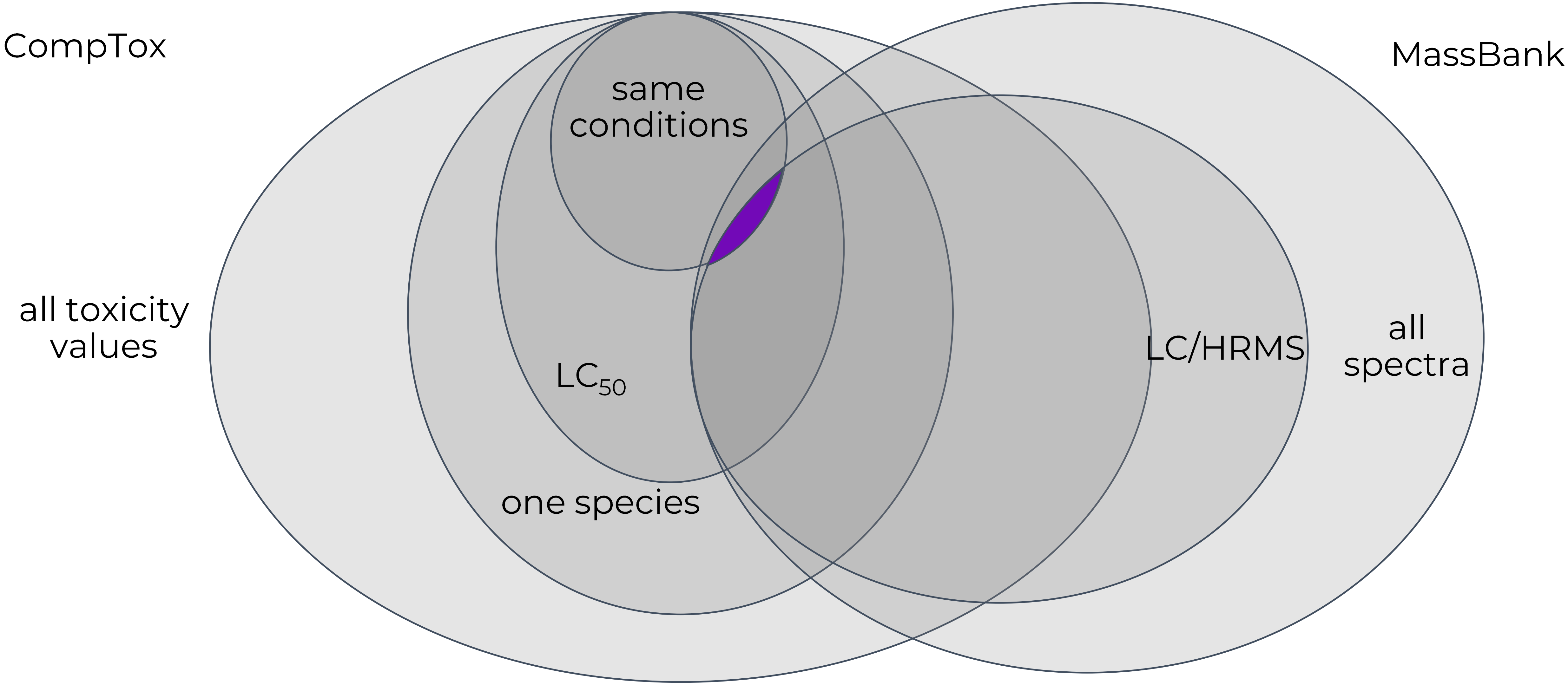
anneli.kruve@su.se

data for machine learning models



anneli.kruve@su.se

data for machine learning models



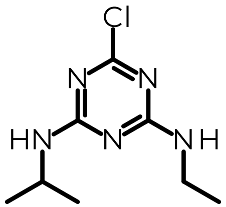
anneli kruve

anneli.kruve@su.se

predicting toxicity

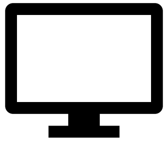
from the structure

workflow



structure as SMILES

molecular fingerprints



machine learning for predicting toxicity

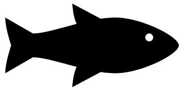
anneli kruve

anneli.kruve@su.se

selected endpoint

anneli.kruve@su.se

selected endpoint

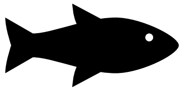


fathead minnow, bluegill, and rainbow trout

anneli kruve

anneli.kruve@su.se

selected endpoint



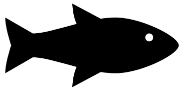
fathead minnow, bluegill, and rainbow trout

water flea

anneli kruve

anneli.kruve@su.se

selected endpoint

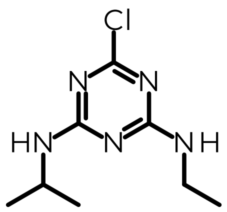


fathead minnow, bluegill, and rainbow trout

water flea

algae

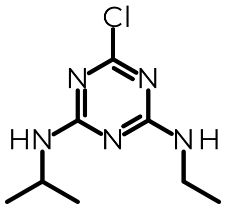
workflow



structure as SMILES

anneli.kruve@su.se

workflow

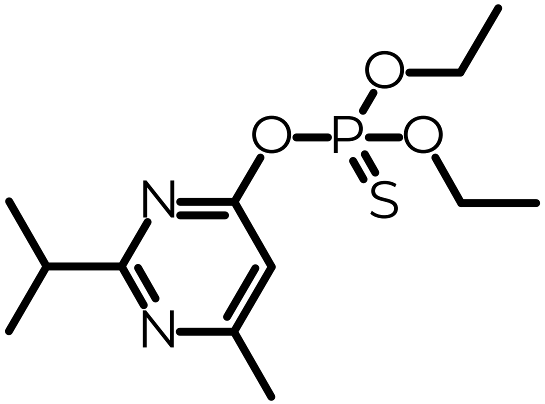


structure as SMILES

molecular fingerprints

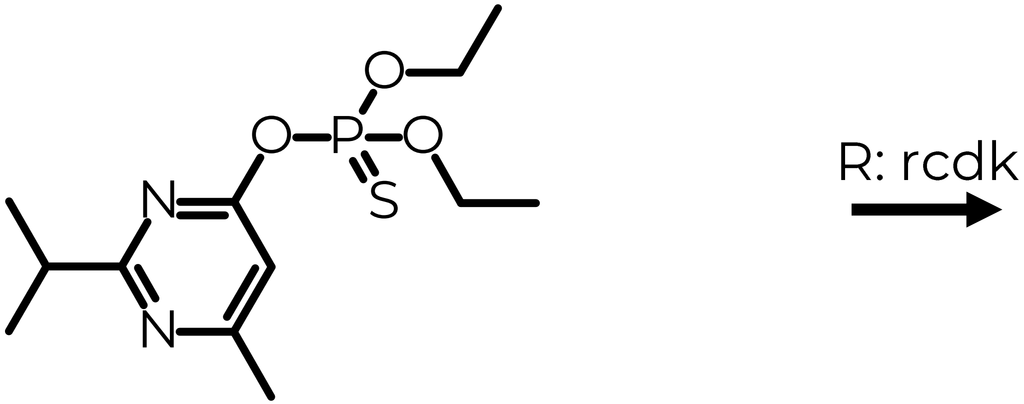
anneli.kruve@su.se

structural fingerprints

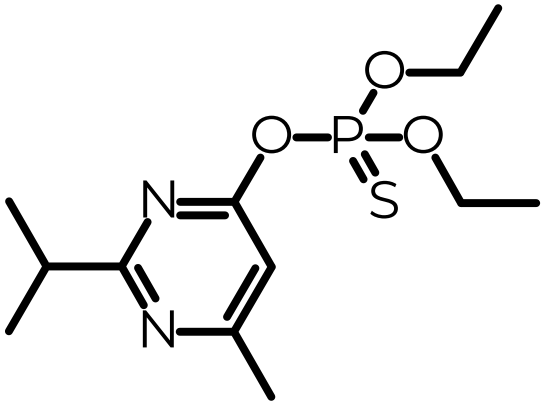


anneli.kruve@su.se

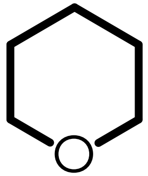
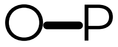
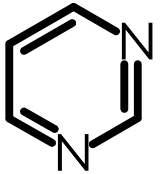
structural fingerprints



structural fingerprints

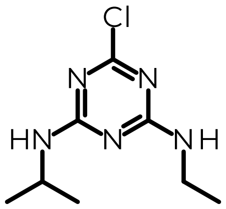


R: rcdk

0	
1	
1	
0	
1	

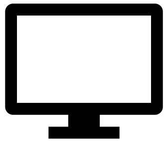
anneli.kruve@su.se

workflow



structure as SMILES

molecular fingerprints



machine learning for predicting LC_{50}

anneli kruve

anneli.kruve@su.se

model training

mass (Da)	fp1	...	fp243
317.32000	0	...	0
208.26100	1	...	0
240.21499	1	...	0
300.57998	0	...	0
201.22500	0	...	0

model training

mass (Da)	fp1	…	fp243
317.32000	0	…	0
208.26100	1	…	0
240.21499	1	…	0
300.57998	0	…	0
201.22500	0	…	0

training set

517
chemicals

test set

130
chemicals

anneli kruve

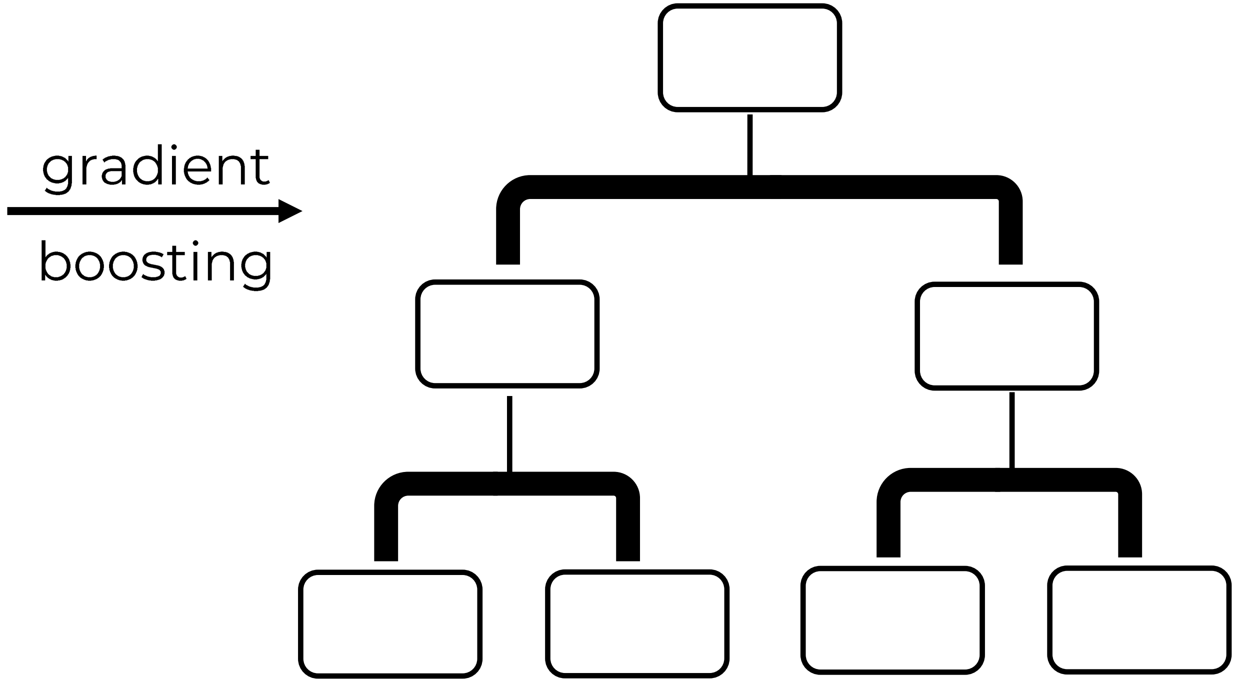
anneli.kruve@su.se

model training

mass (Da)	fp1	...	fp243
317.32000	0	...	0
208.26100	1	...	0
240.21499	1	...	0
300.57998	0	...	0
201.22500	0	...	0

training set
517 chemicals

test set
130 chemicals



anneli kruve

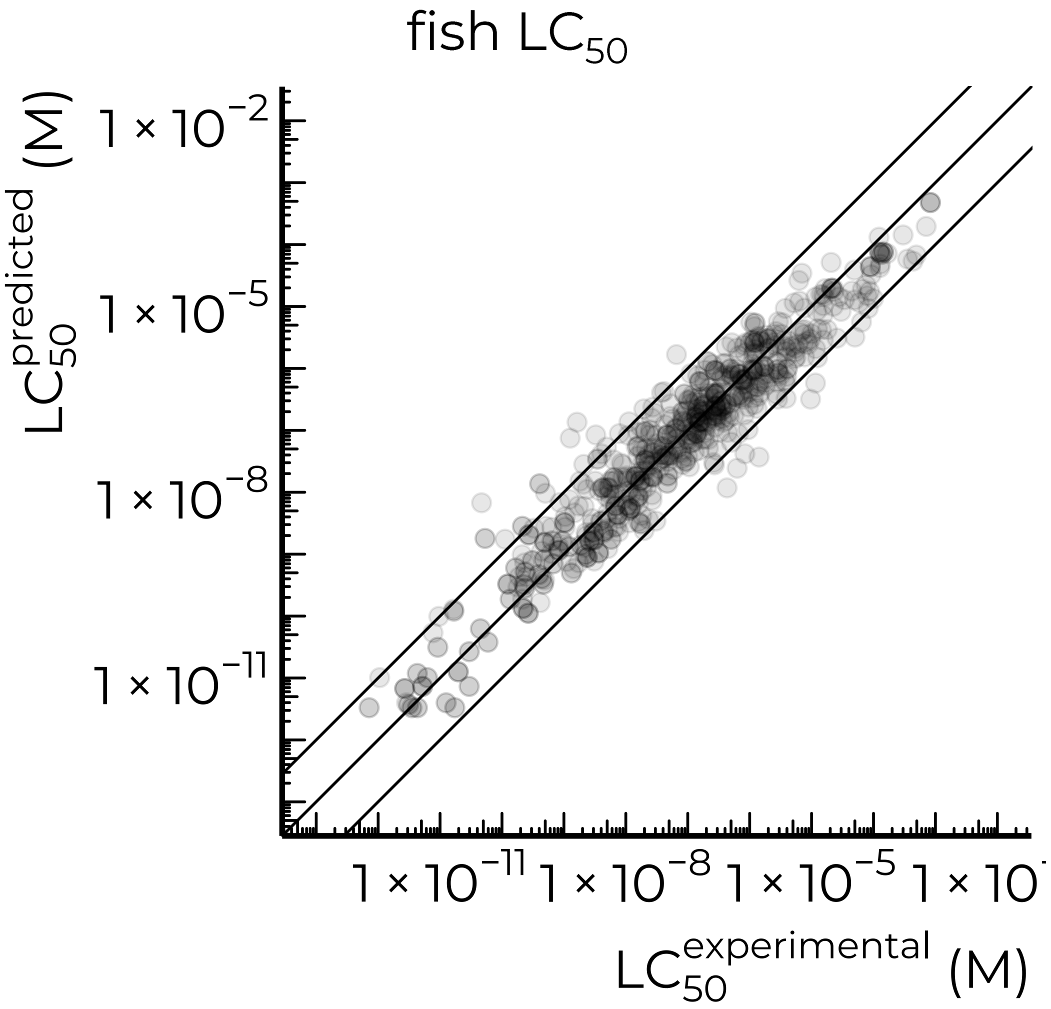
anneli.kruve@su.se

performance

of LC_{50} predictions with molecular fingerprints

LC_{50} predictions

Peets et al. ES&T 2022



training set

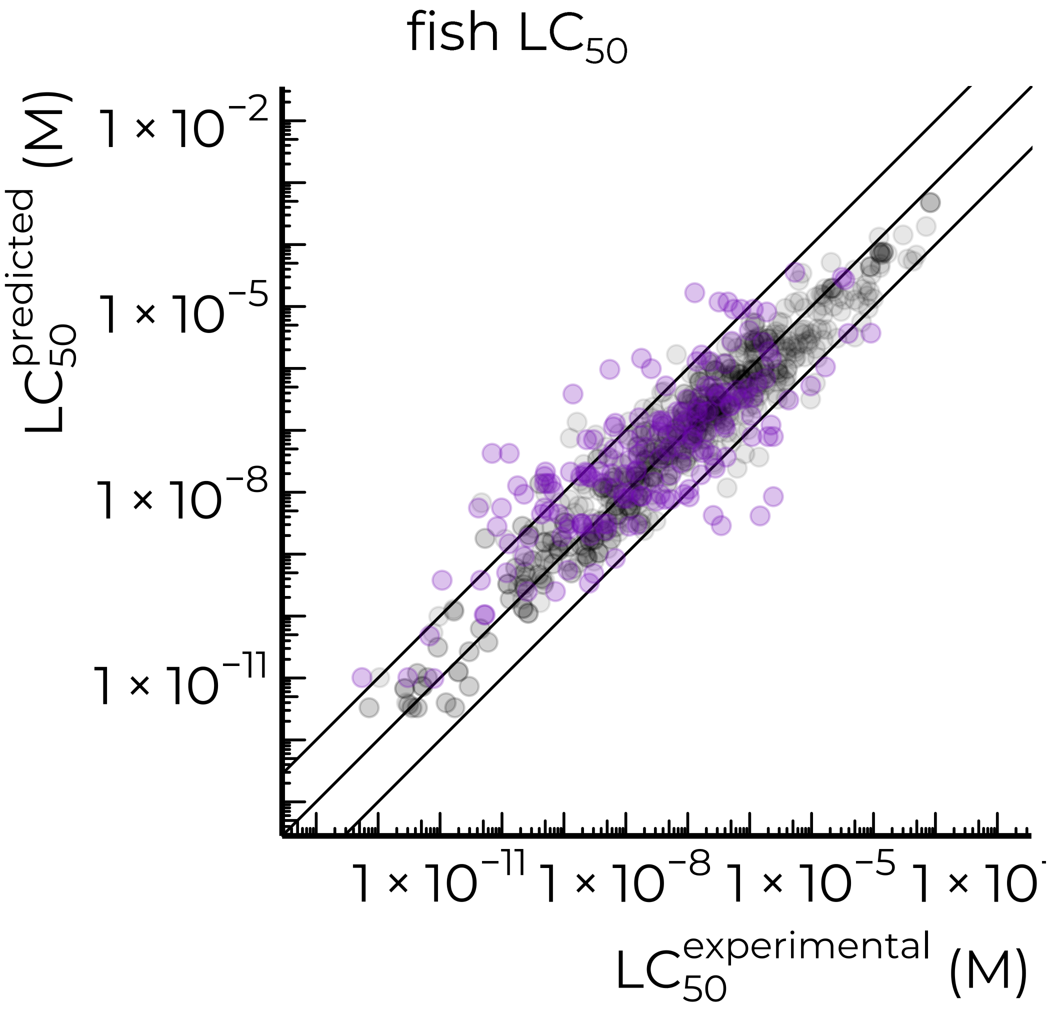
RMSE 0.52 log(M)

anneli kruve

anneli.kruve@su.se

LC_{50} predictions

Peets et al. ES&T 2022



training set

RMSE 0.52 log(M)

test set

RMSE 0.78 log(M)

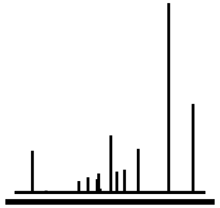
anneli kruve

anneli.kruve@su.se

unidentified chemicals

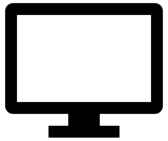
from MS^2 spectra

workflow



MS^2 spectra

molecular fingerprints with SIRIUS+CSI:FingerID

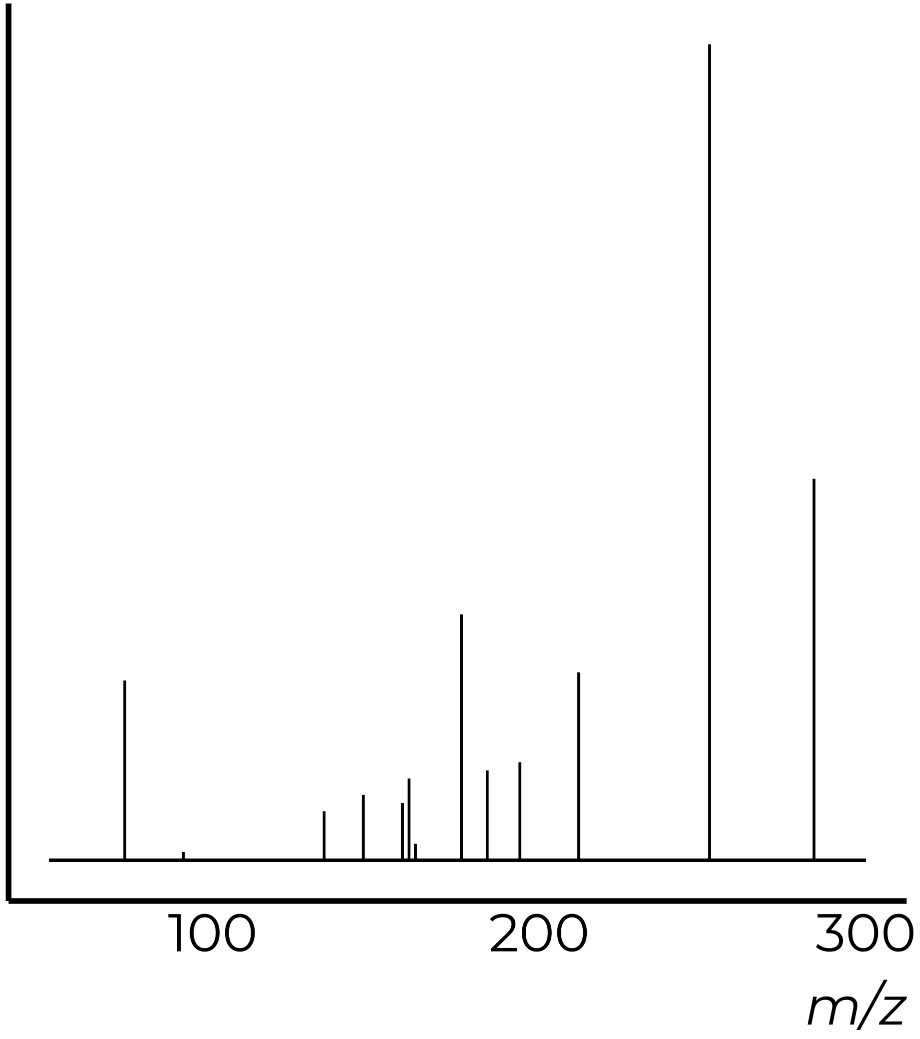


predict LC_{50} with pretrained gradient boosting

anneli kruve

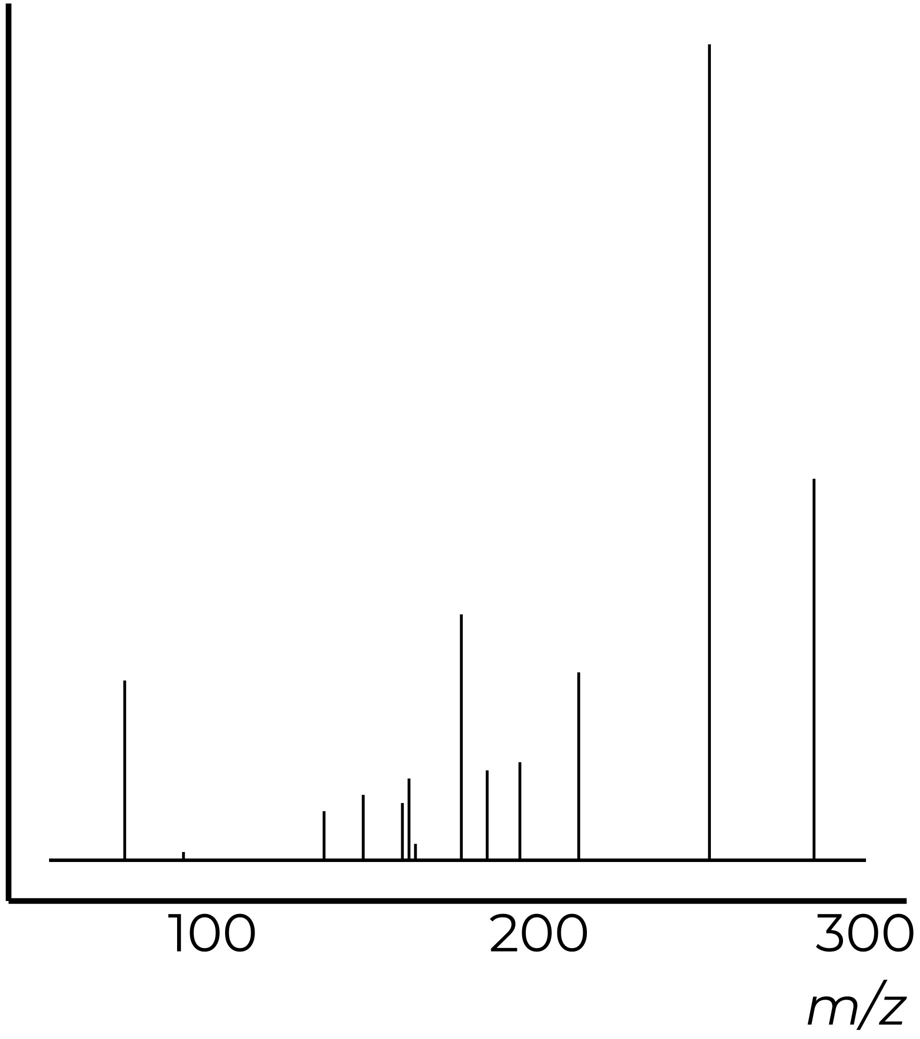
anneli.kruve@su.se

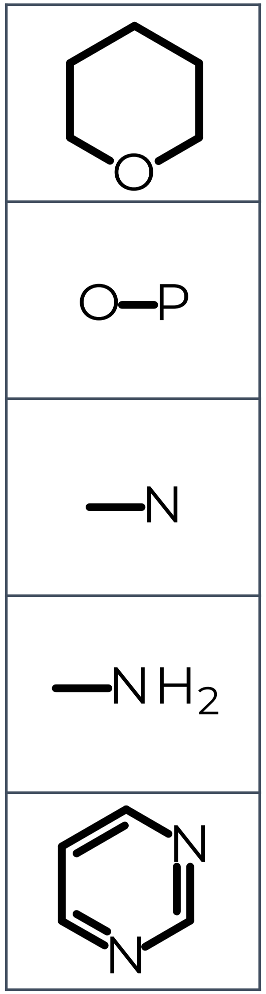
predict for unknown chemicals



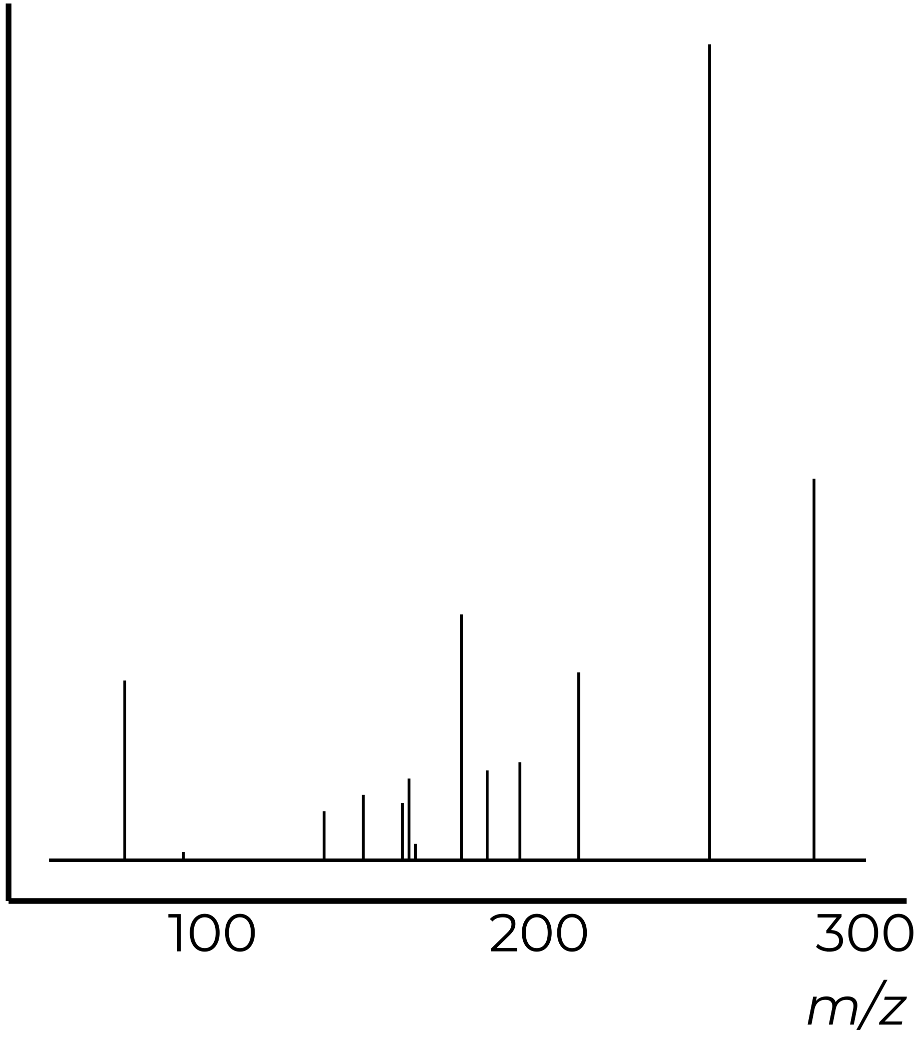
anneli.kruve@su.se

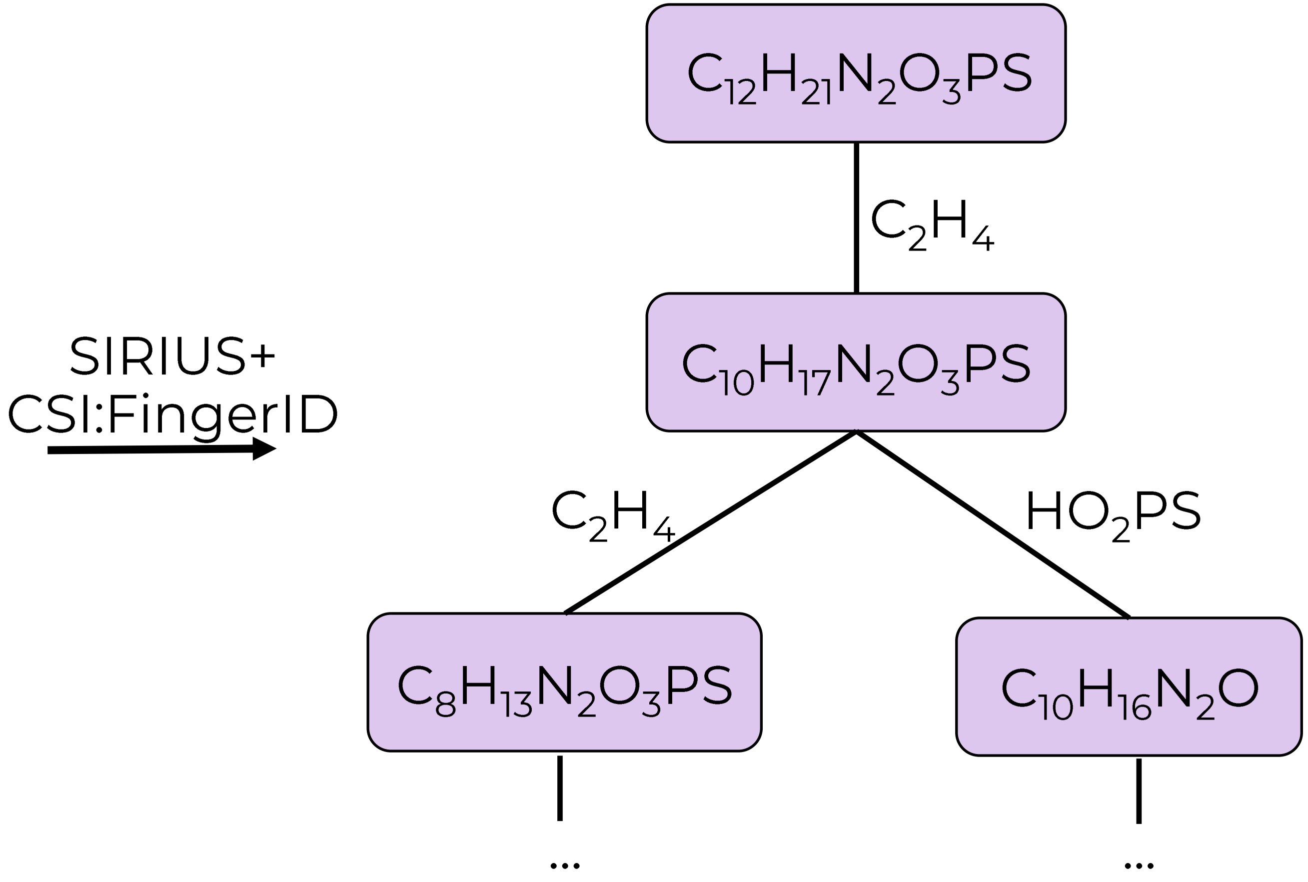
predict for unknown chemicals





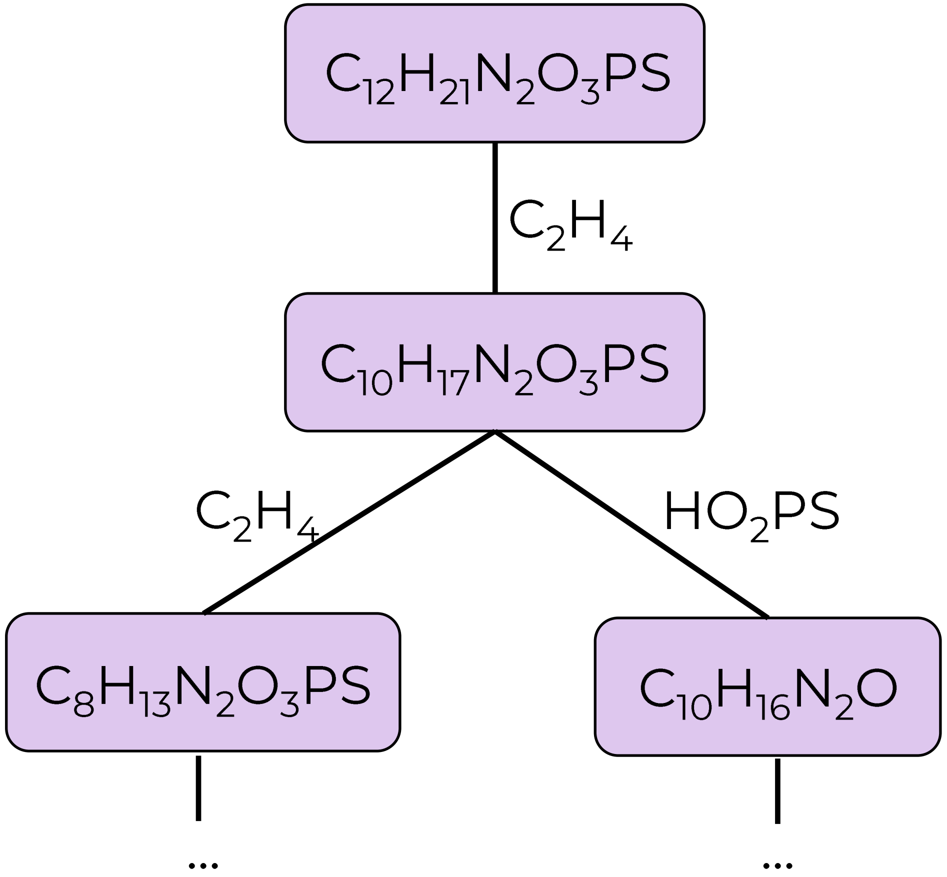
predict for unknown chemicals



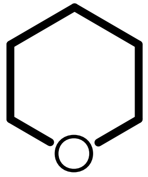
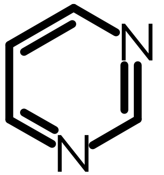


anneli.kruve@su.se

predict for unknown chemicals



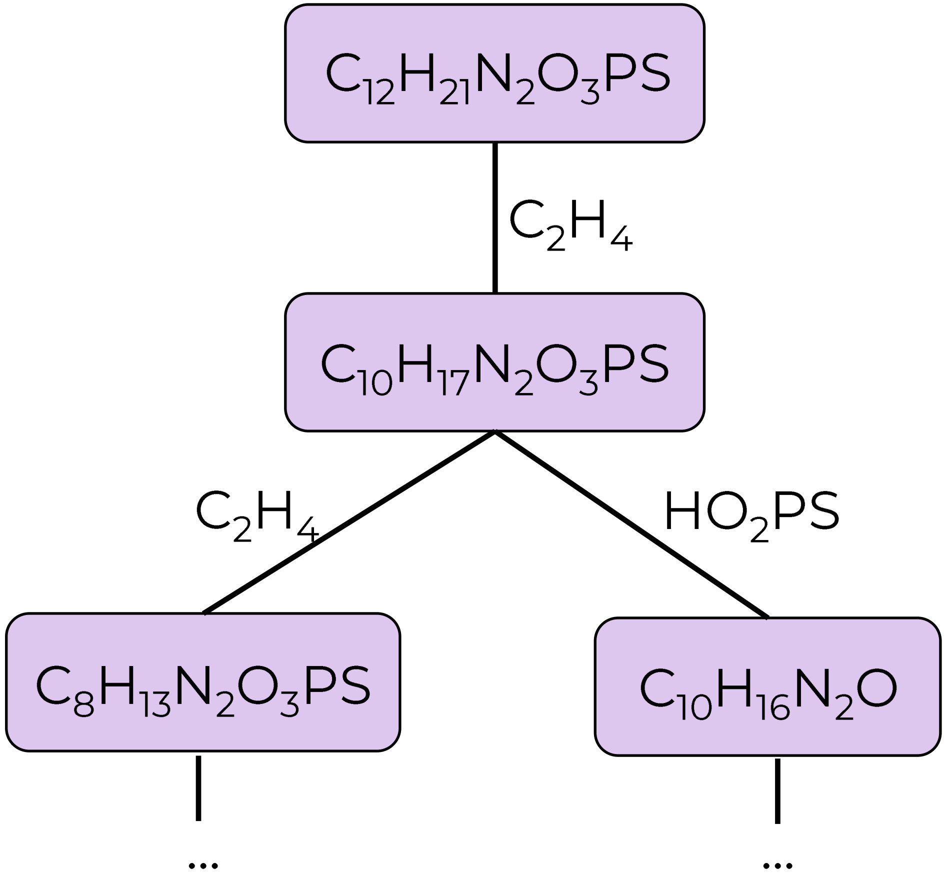
SIRIUS+
CSI:FingerID →

0.001	
0.999	O–P
0.999	–N
0.198	–NH_2
0.988	

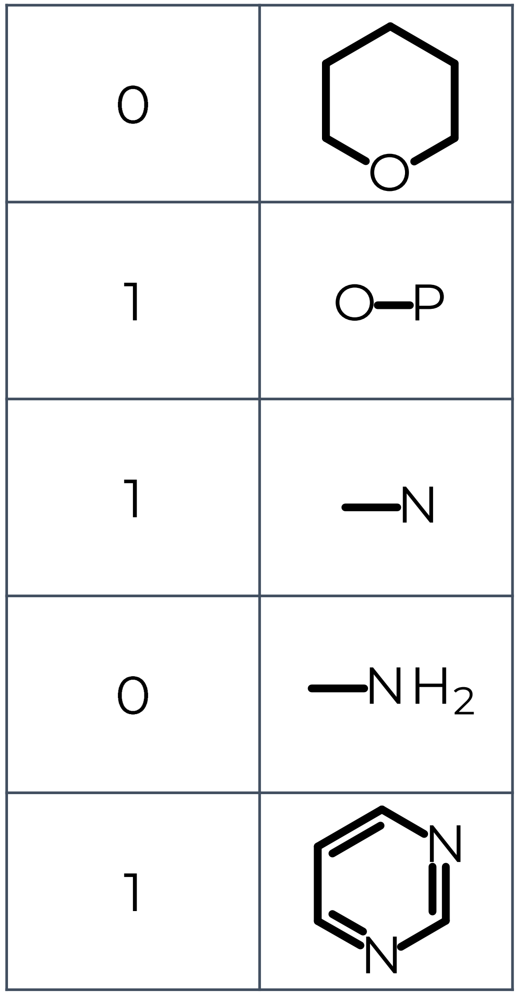
anneli kruve

anneli.kruve@su.se

predict for unknown chemicals



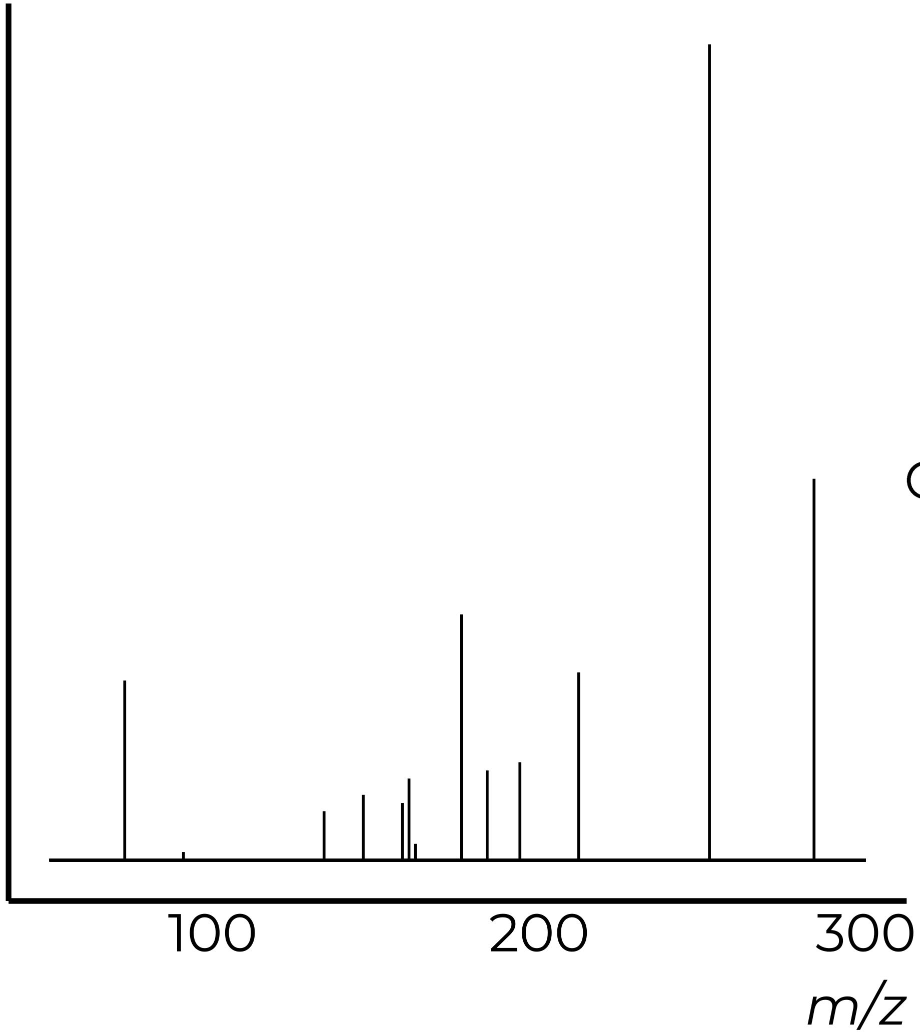
SIRIUS+
CSI:FingerID



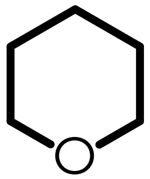
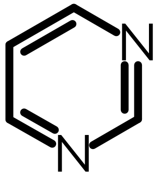
anneli kruve

anneli.kruve@su.se

predict for unknown chemicals



SIRIUS+
CSI:FingerID →

0	
1	O—P
1	—N
0	—NH_2
1	

gradient
boosting →

LC_{50} =
-2.2 log(mM)

anneli kruve

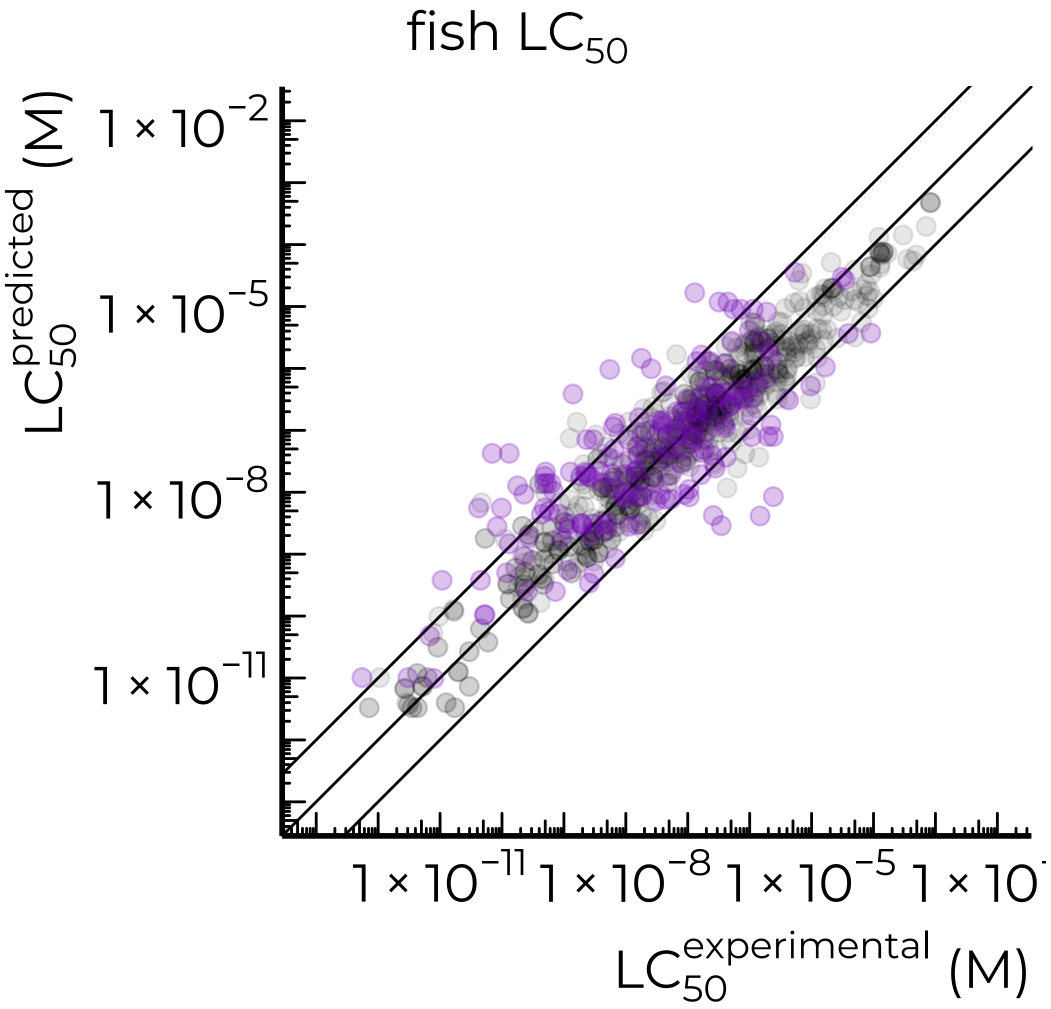
anneli.kruve@su.se

LC_{50} predictions

anneli.kruve@su.se

LC_{50} predictions

Peets et al. ES&T 2022



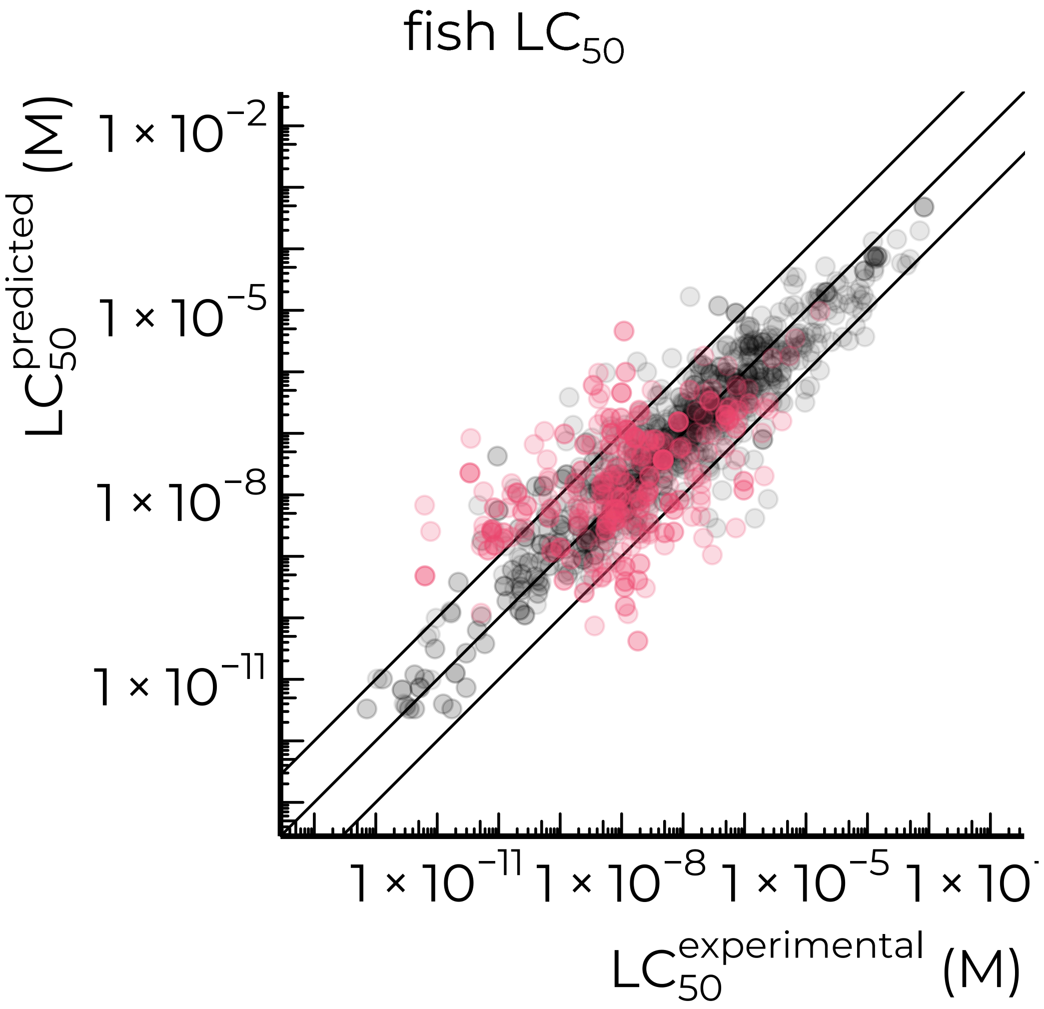
test set on structures

RMSE 0.78 log(M)

anneli.kruve@su.se

LC_{50} predictions

Peets et al. ES&T 2022



test set on structures

RMSE 0.78 log(M)

validation on MassBank

$RMSE_{model}$ 0.88 log(M)

$SD_{experimental}$ 0.44 log(mM)

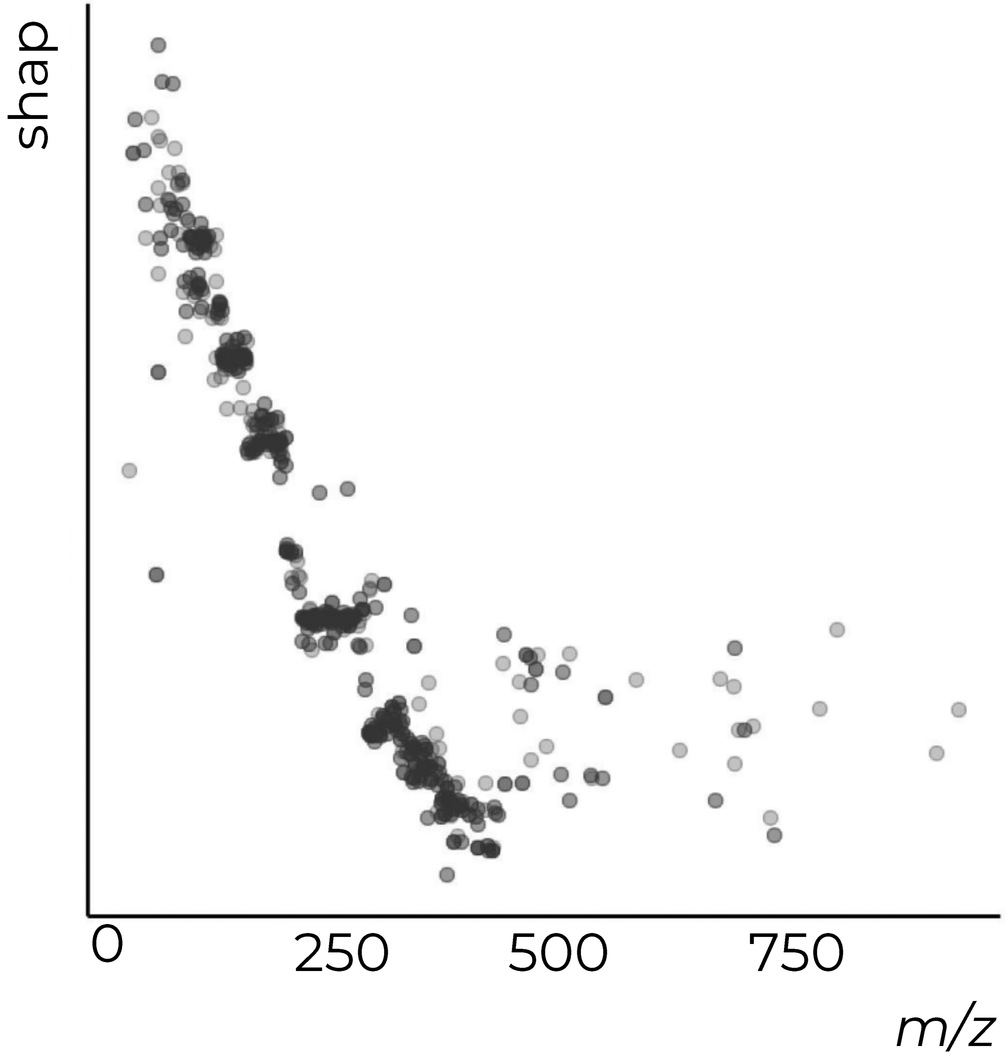
anneli.kruve@su.se

model interpretation

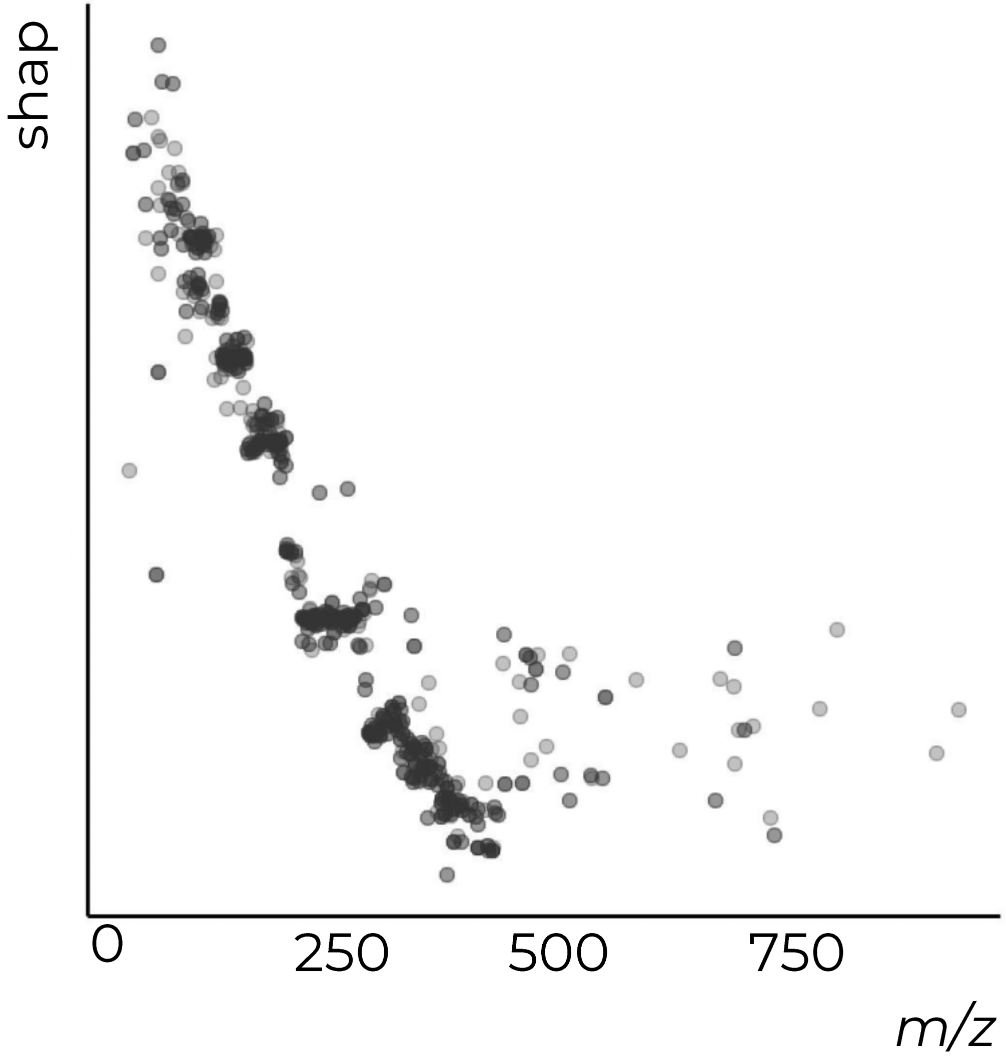
anneli kruve

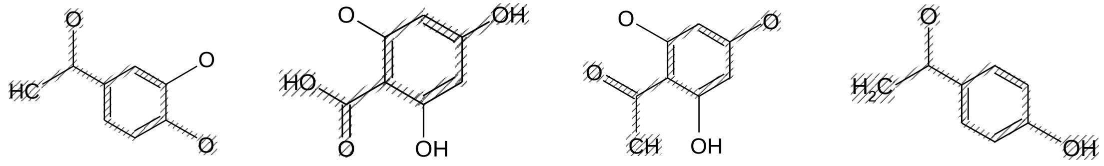
anneli.kruve@su.se

model interpretation



model interpretation

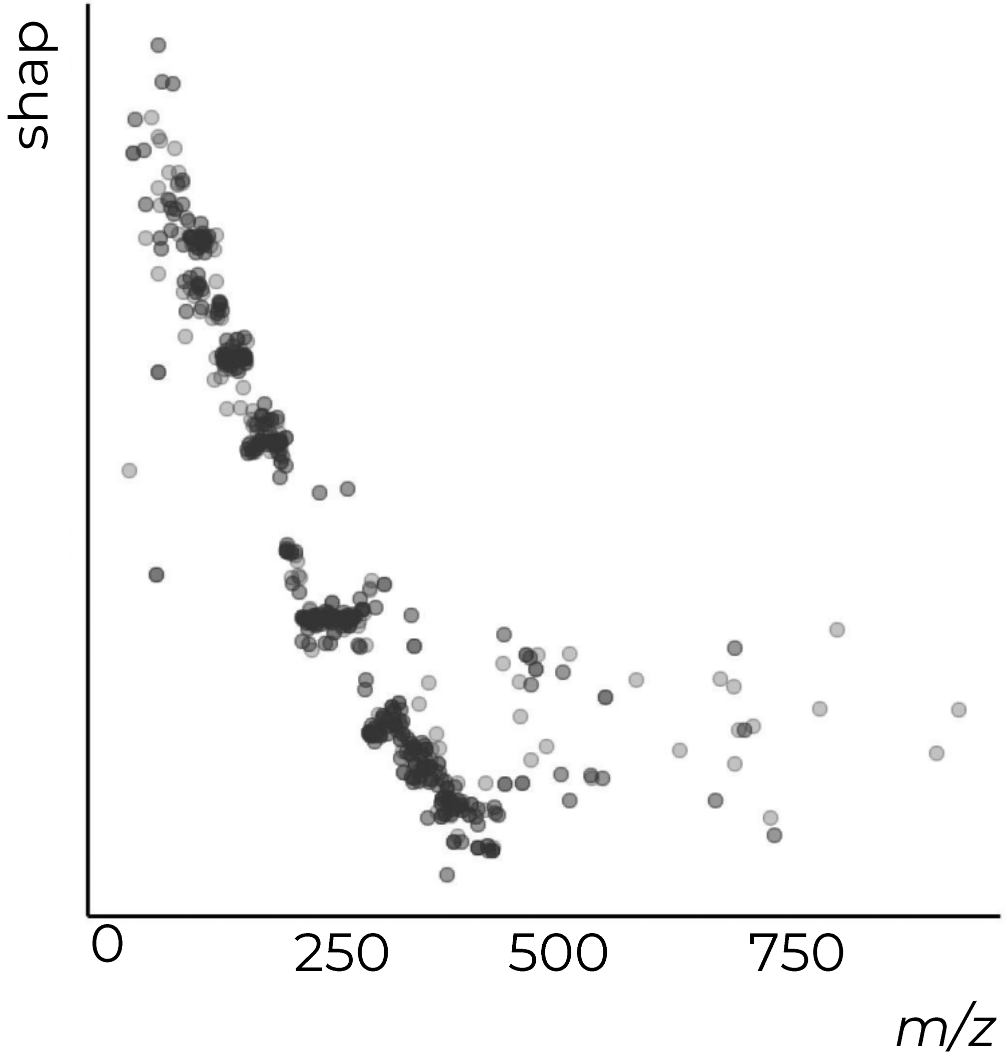


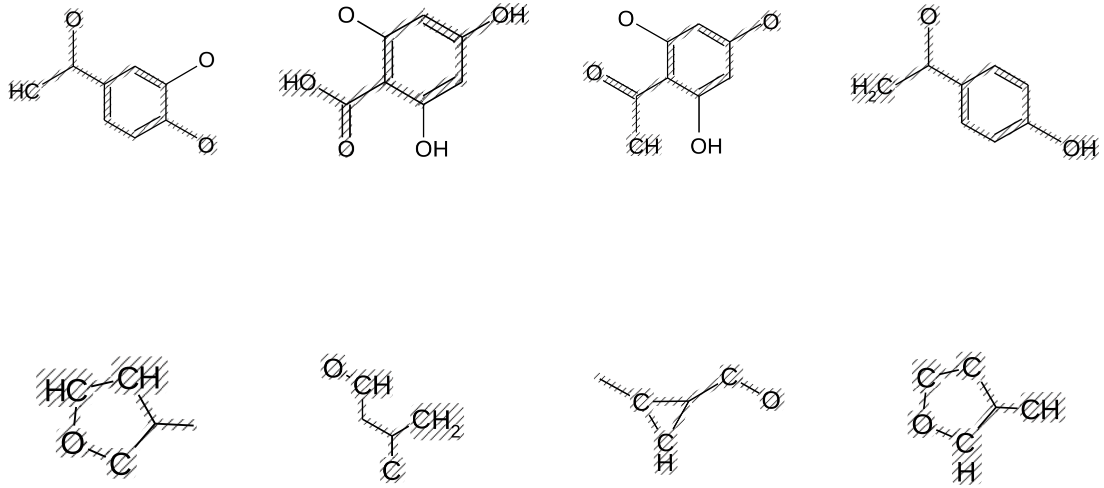


anneli kruve

anneli.kruve@su.se

model interpretation

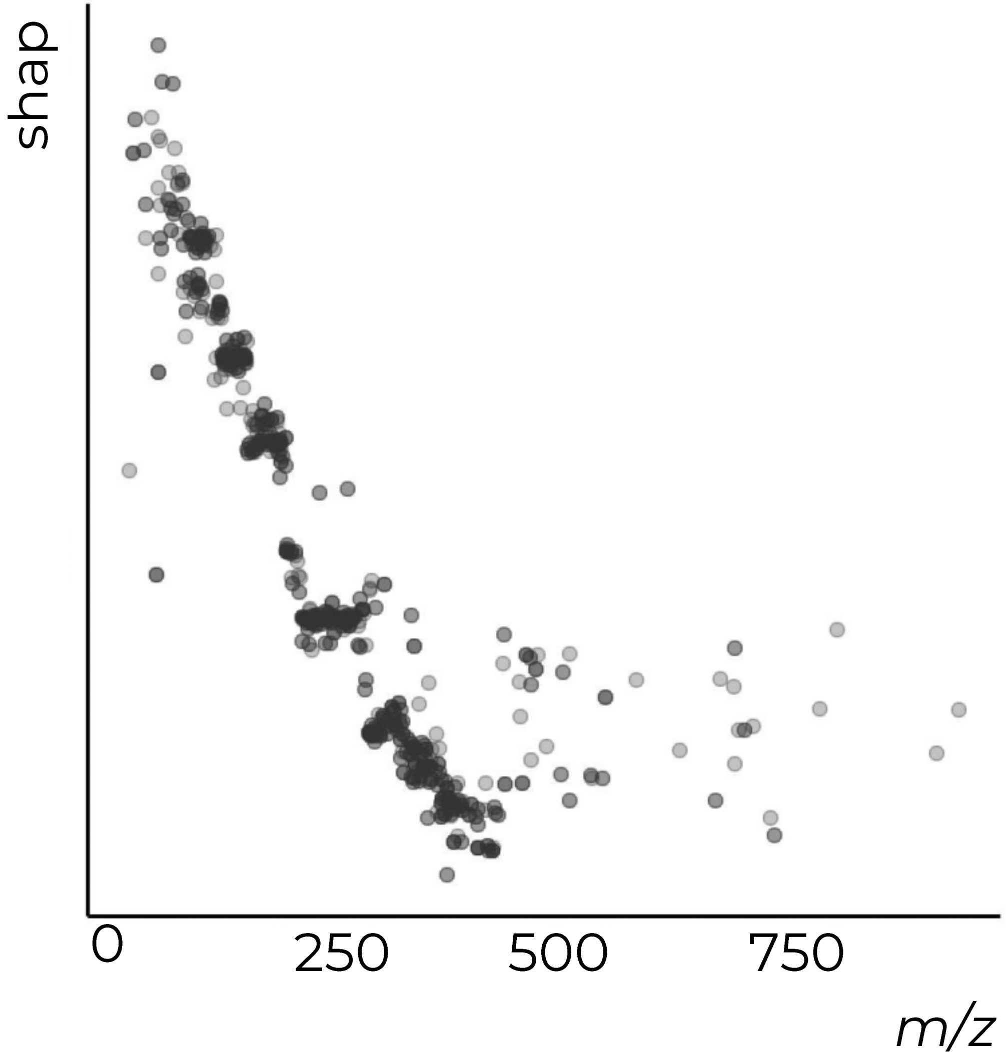


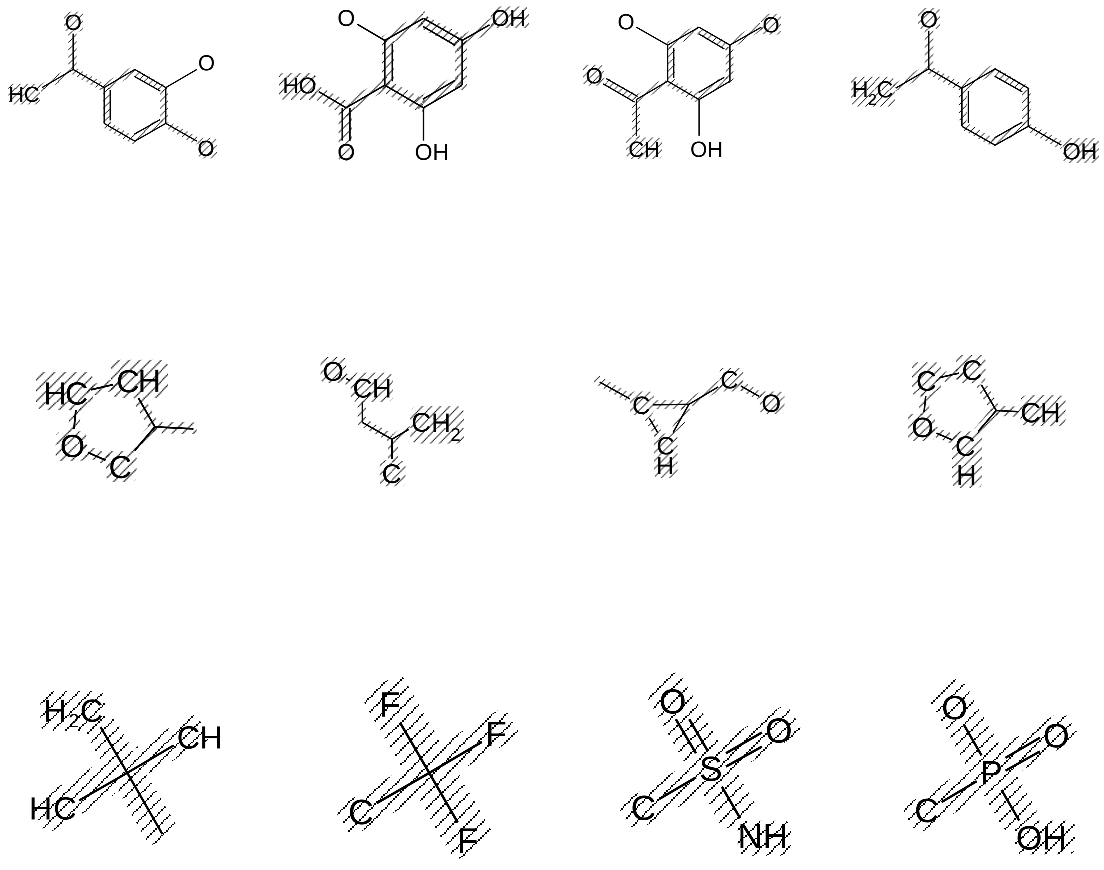


anneli kruve

anneli.kruve@su.se

model interpretation





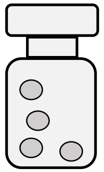
anneli kruve

anneli.kruve@su.se

toxic chemicals

in wastewater

case study on wastewater

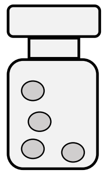


wastewater samples

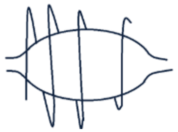
anneli kruve

anneli.kruve@su.se

case study on wastewater



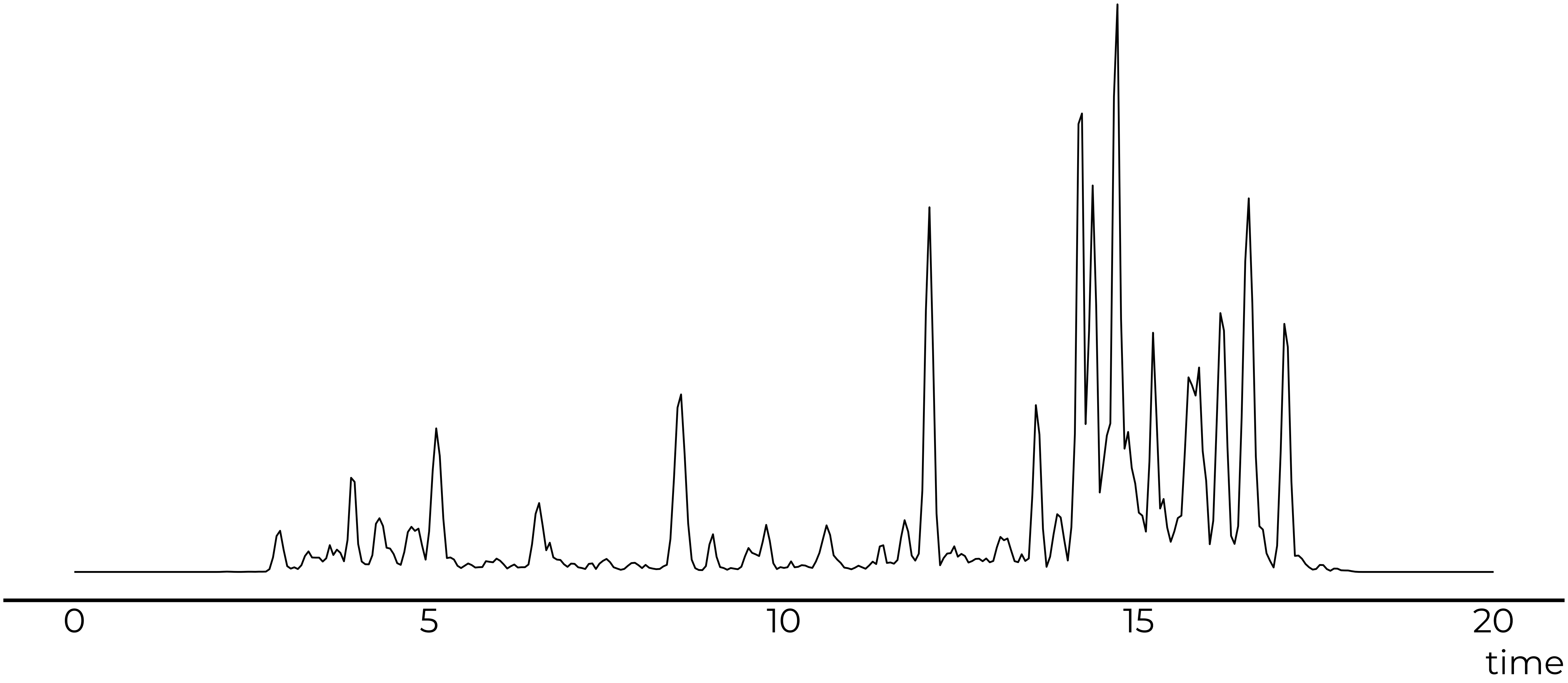
wastewater samples



LC/HRMS analysis

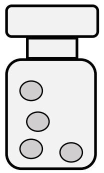
anneli.kruve@su.se

case study on wastewater

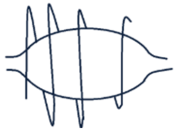


anneli kruve
anneli.kruve@su.se

case study on wastewater



wastewater samples



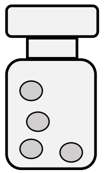
LC/HRMS analysis

molecular fingerprints with SIRIUS+CSI:FingerID

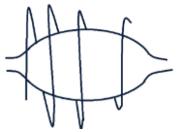
anneli kruve

anneli.kruve@su.se

case study on wastewater

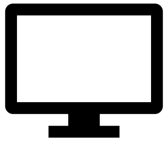


wastewater samples



LC/HRMS analysis

molecular fingerprints with SIRIUS+CSI:FingerID



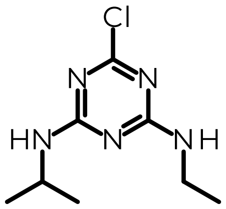
predict LC_{50} with pretrained gradient boosting

anneli kruve

anneli.kruve@su.se

quality control

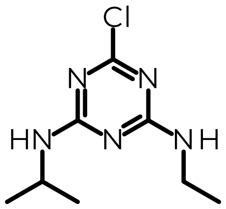
quality control



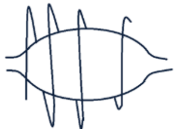
216 analytical standard

anneli.kruve@su.se

quality control

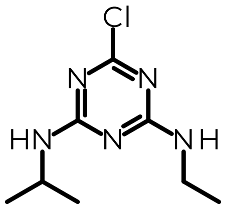


216 analytical standard

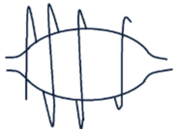


DIA and DDA MS^2 data

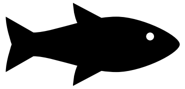
quality control



216 analytical standard



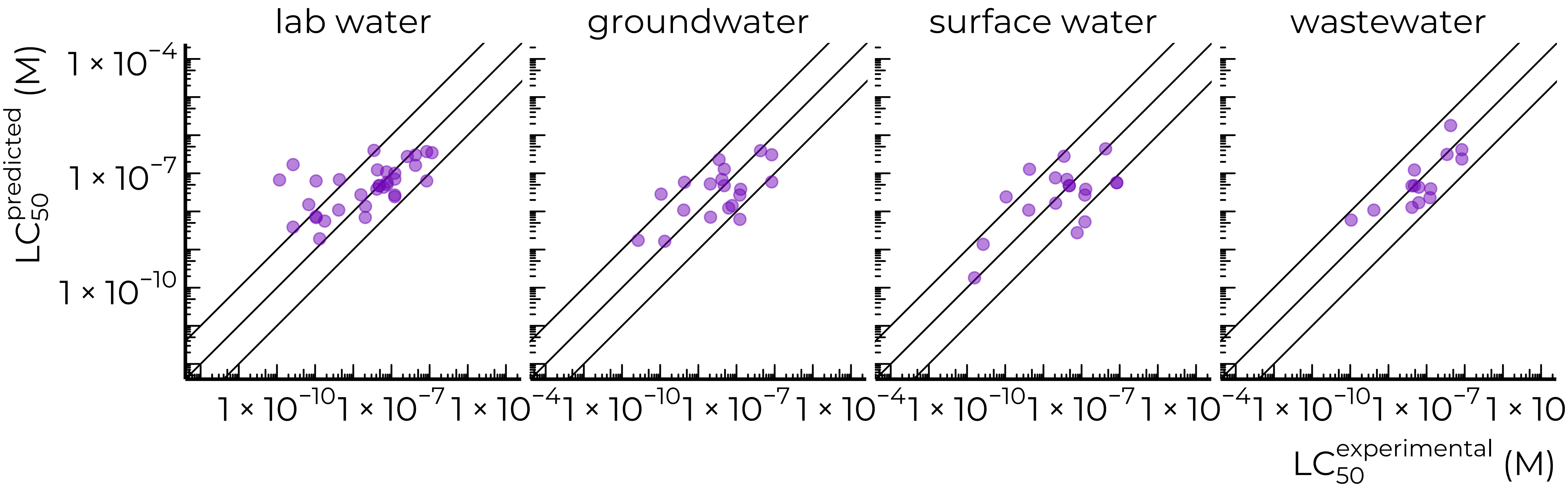
DIA and DDA MS^2 data



comparison with experimental LC_{50}

anneli.kruve@su.se

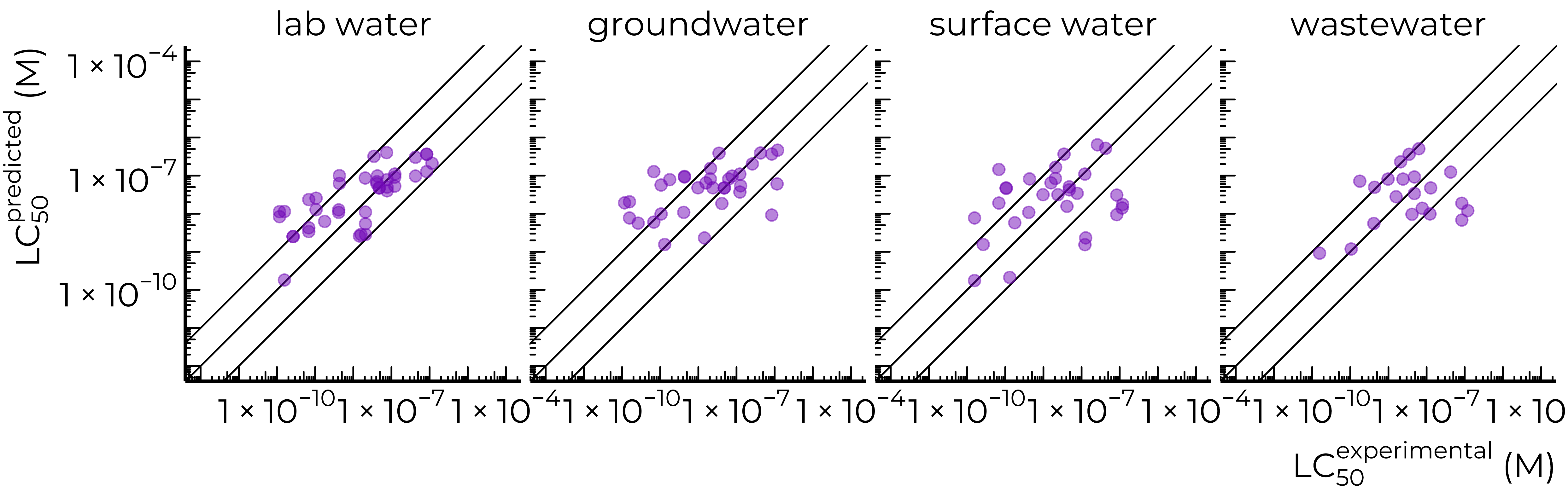
DDA



RMSE = 0.95 log-mM 0.74 log-mM 0.86 log-mM 0.47 log-mM

anneli kruve anneli.kruve@su.se

DIA

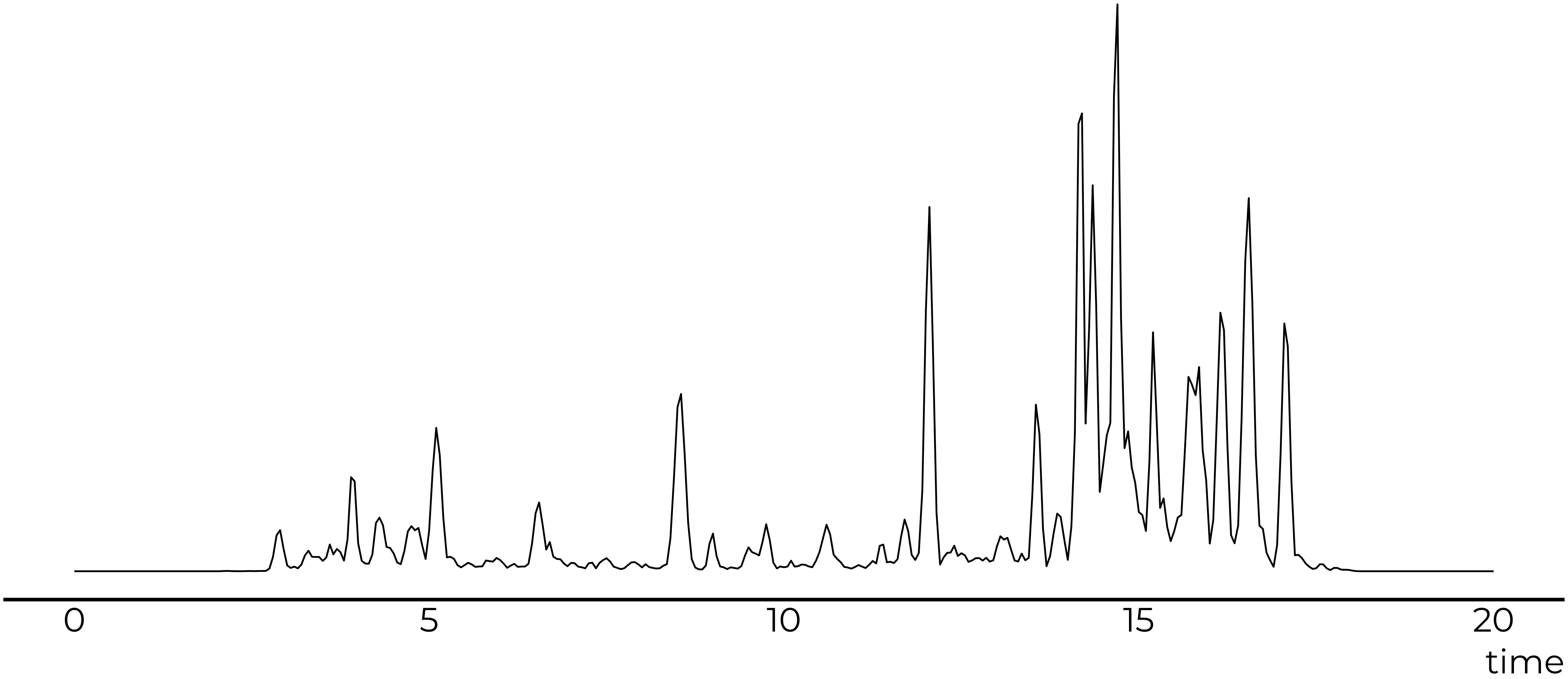


RMSE = 0.85 log-mM 1.09 log-mM 1.18 log-mM 1.03 log-mM

anneli kruve anneli.kruve@su.se

pinpointing toxic chemicals

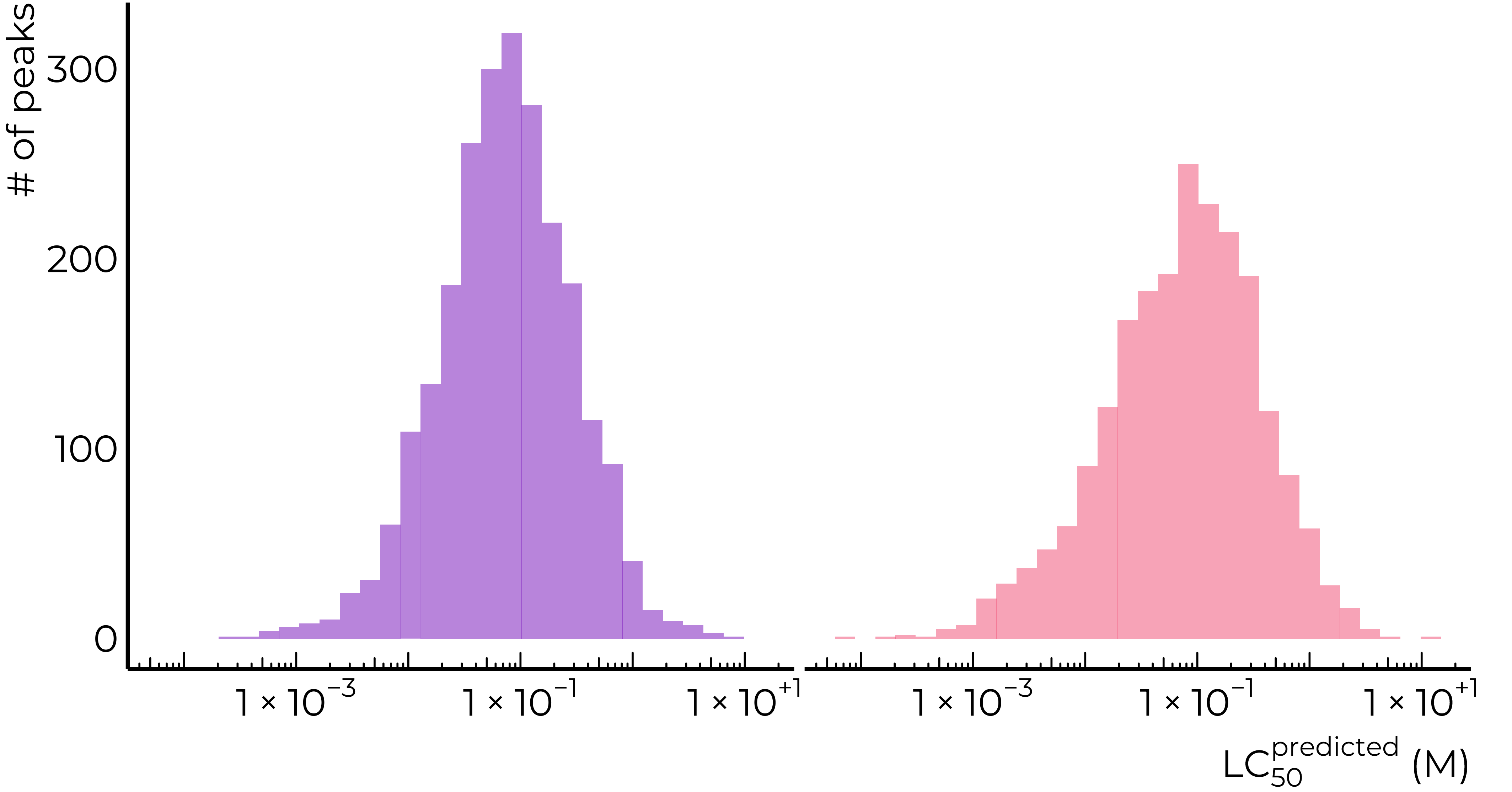
case study on wastewater



anneli kruve

anneli.kruve@su.se

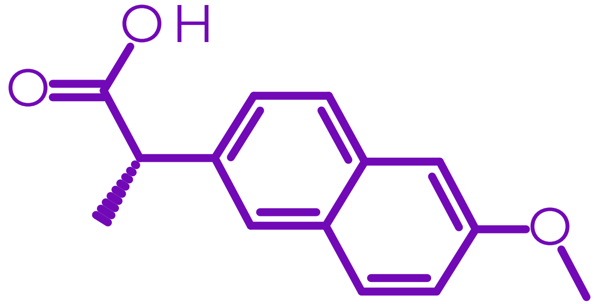
LC_{50} distribution

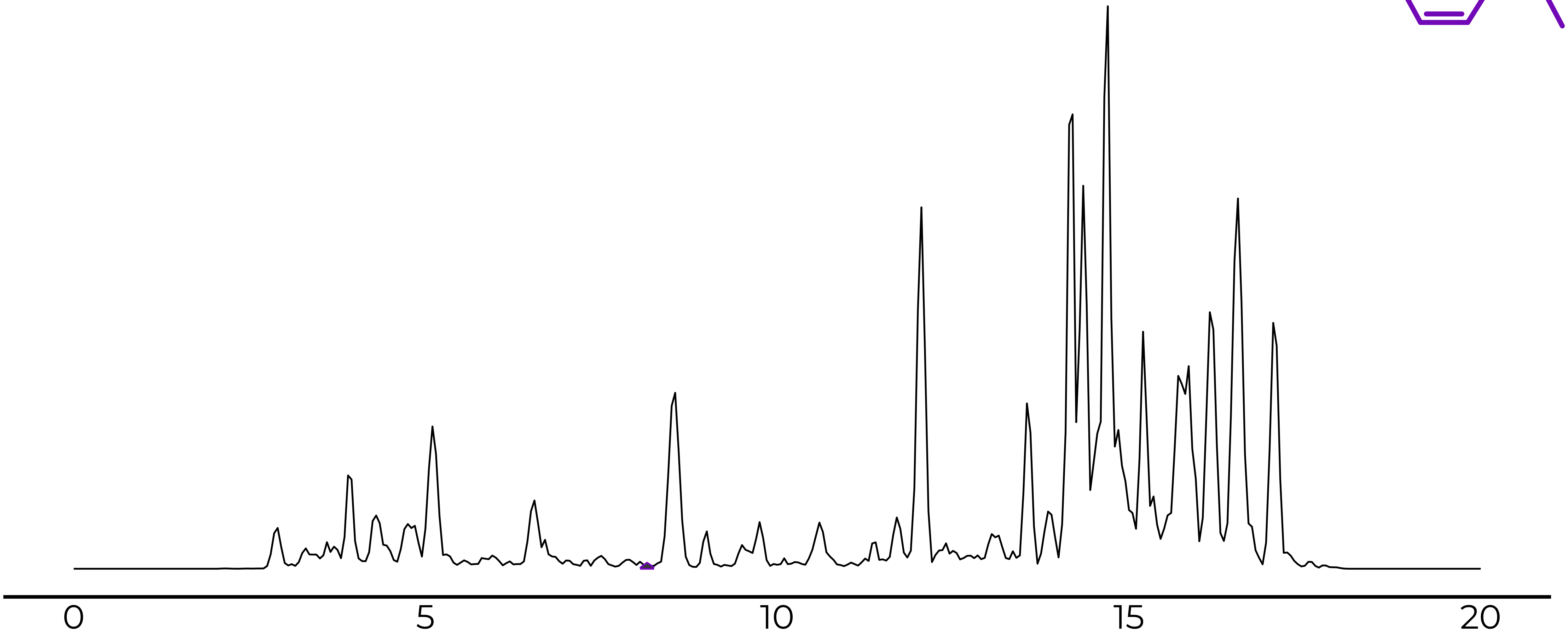


anneli kruve

anneli.kruve@su.se

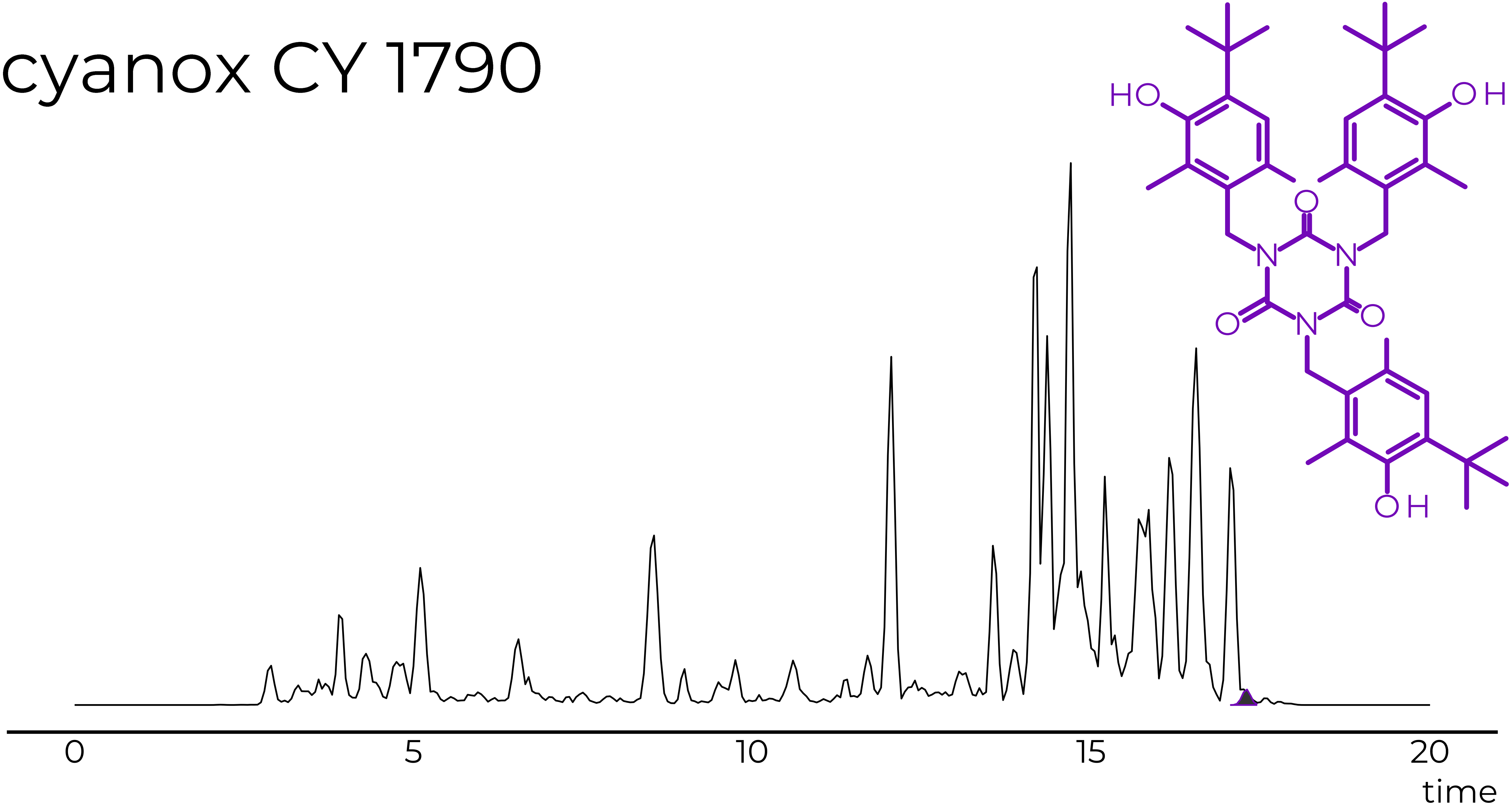
naproxen





anneli kruve
anneli.kruve@su.se

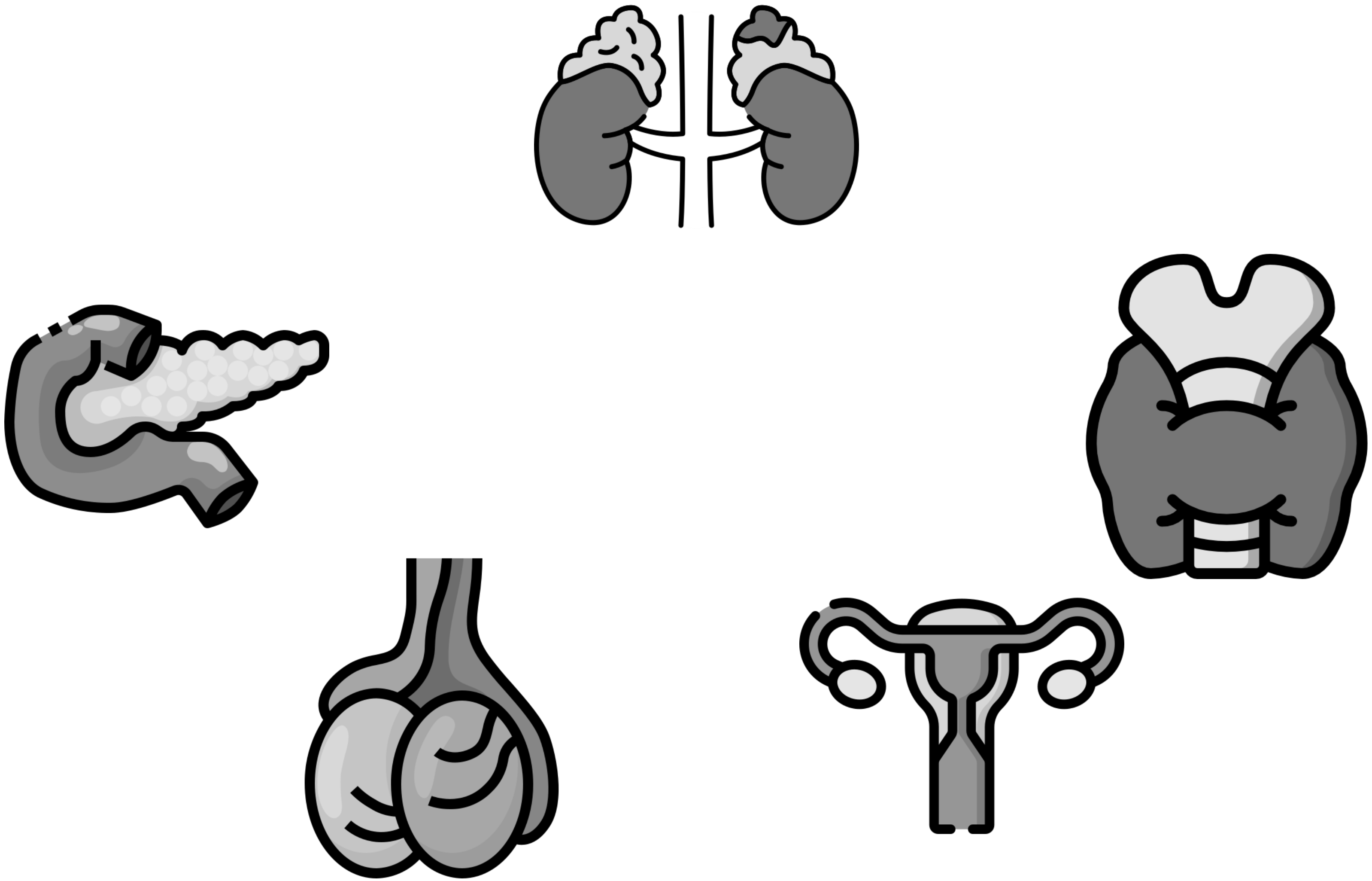
cyanox CY 1790



anneli kruve
anneli.kruve@su.se

endocrine disrupting chemicals

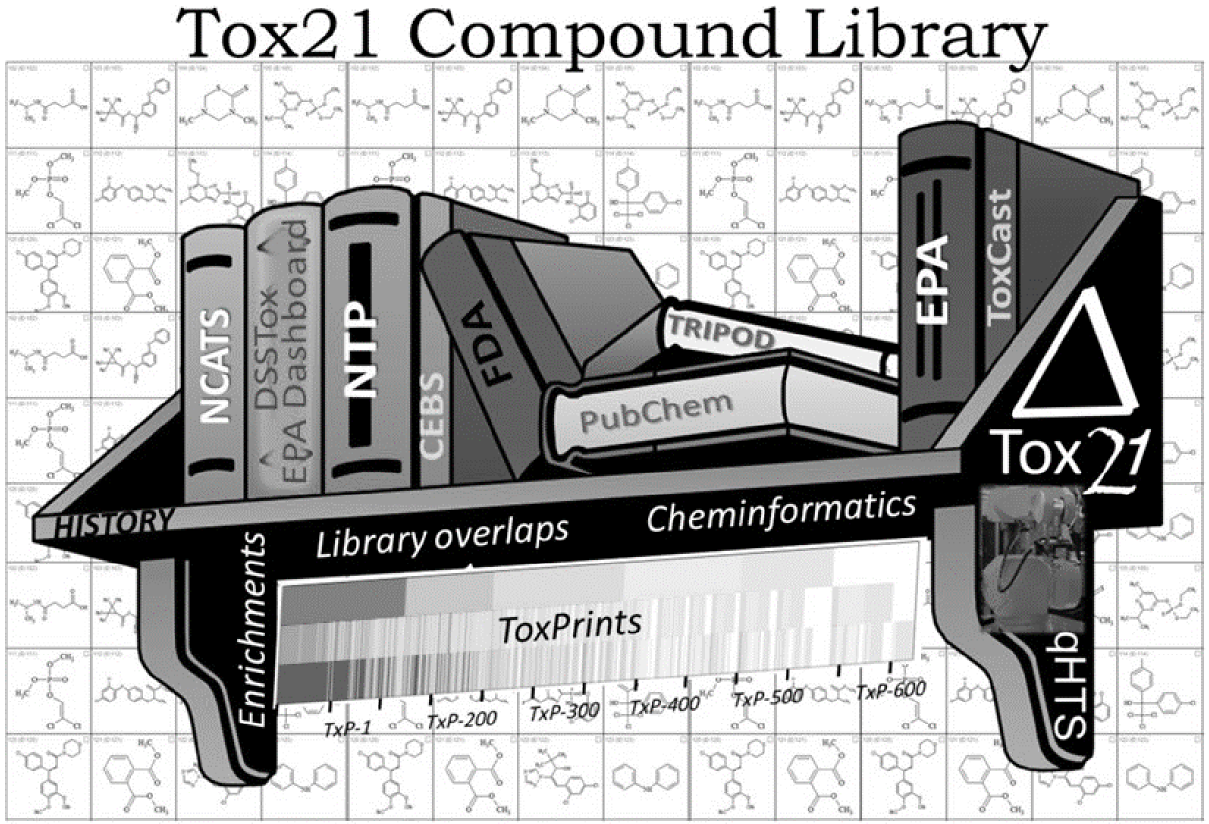
endocrine disruption



anneli kruve

anneli.kruve@su.se

data



nuclear receptor panel

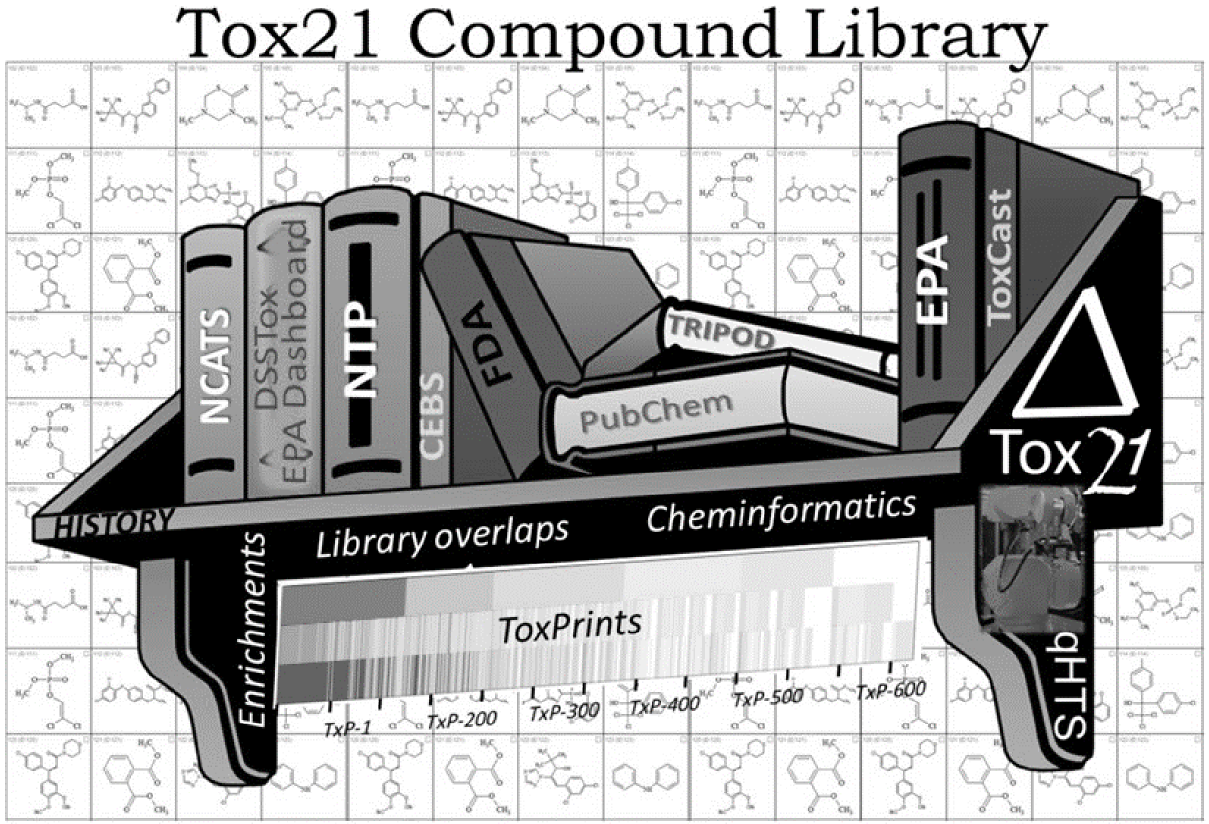
- nr.ahr
- nr.ar.lbd
- nr.ar
- nr.aromatase
- nr.er.lbd
- nr.er
- nr.ppar.gamma

stress response panel

- sr.are
- sr.atad5
- sr.hse
- sr.mmp
- sr.p53

 anneli.kruve@su.se

data



8,043 chemicals

5090 no replica

2953 with replica

replica often inconsistent

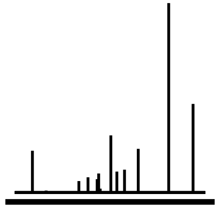
precautionary principle

active chemicals

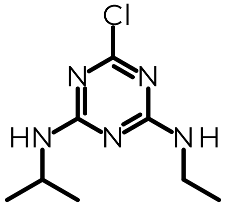
4% to 16%

anneli.kruve@su.se

workflow: training

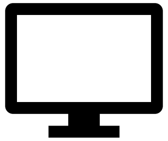


MS^2 spectra



structure as SMILES

molecular descriptors



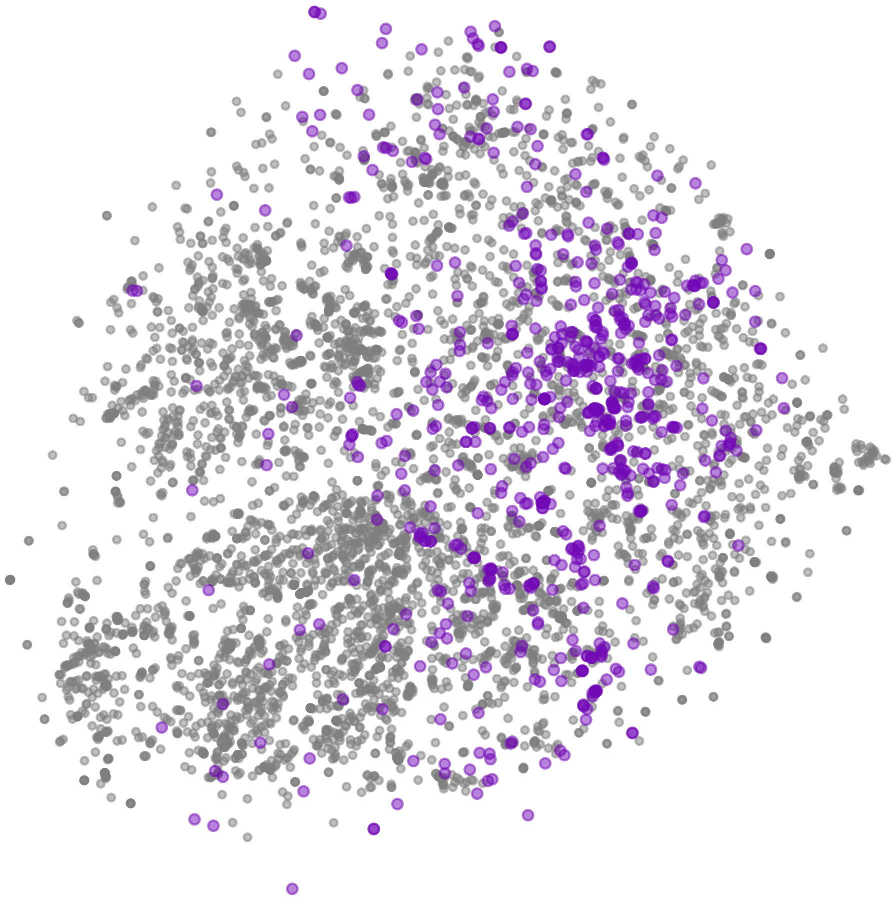
predict toxicity

anneli kruve

anneli.kruve@su.se

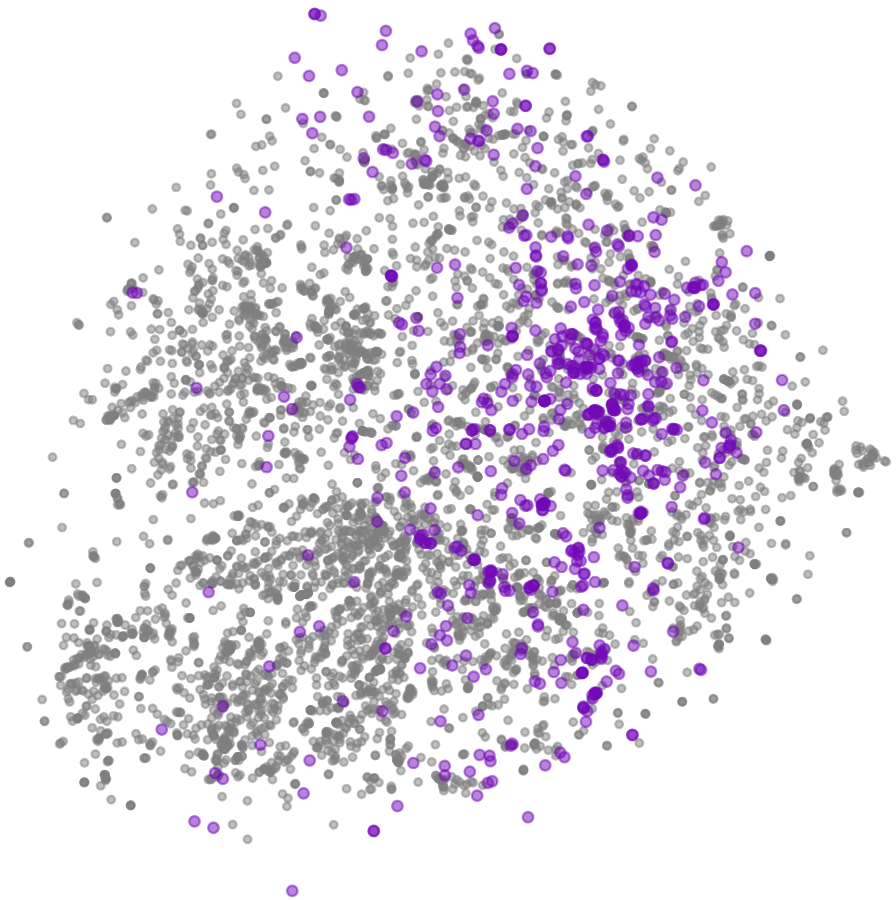
results: t-SNE

aryl hydrocarbon receptor

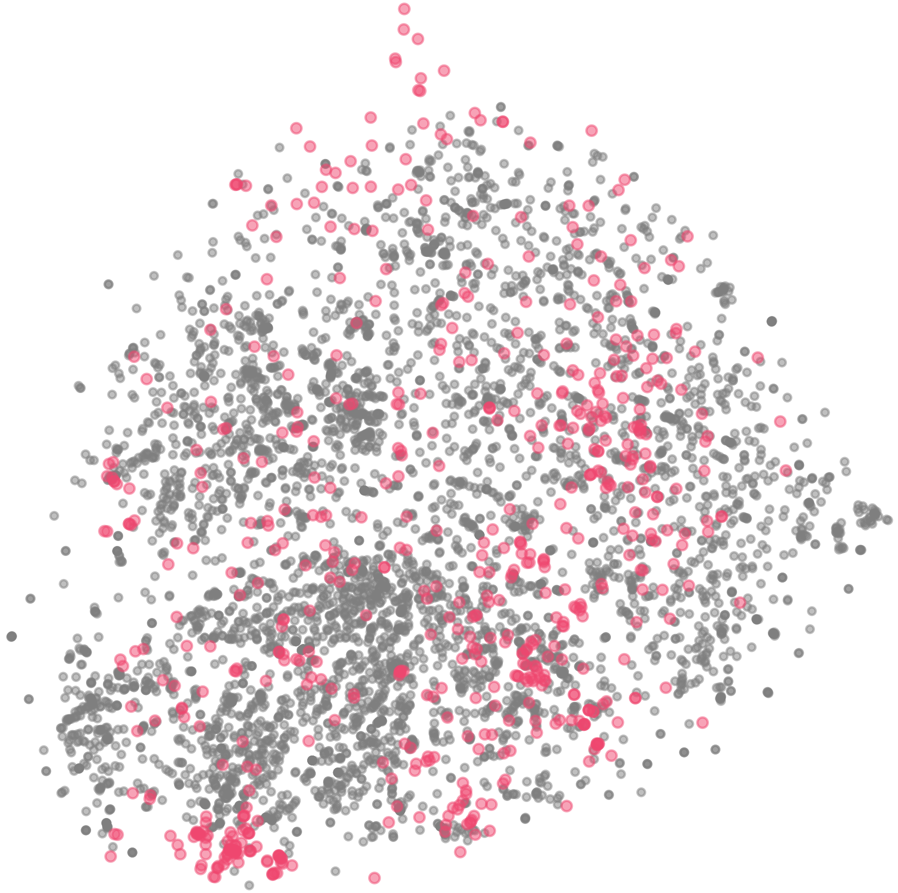


results: t-SNE

aryl hydrocarbon receptor activation



estrogen receptor activation



anneli kruve

anneli.kruve@su.se

metrics

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

which is more dramatic:

type I error

OR

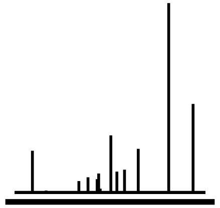
type II error?

FPR @ TPR = 0.9

anneli kruve

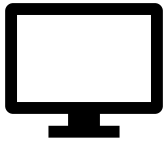
anneli.kruve@su.se

workflow: validation



MS^2 spectra

molecular fingerprints with SIRIUS+CSI:FingerID



predict LC_{50} with pretrained gradient boosting

anneli kruve

anneli.kruve@su.se

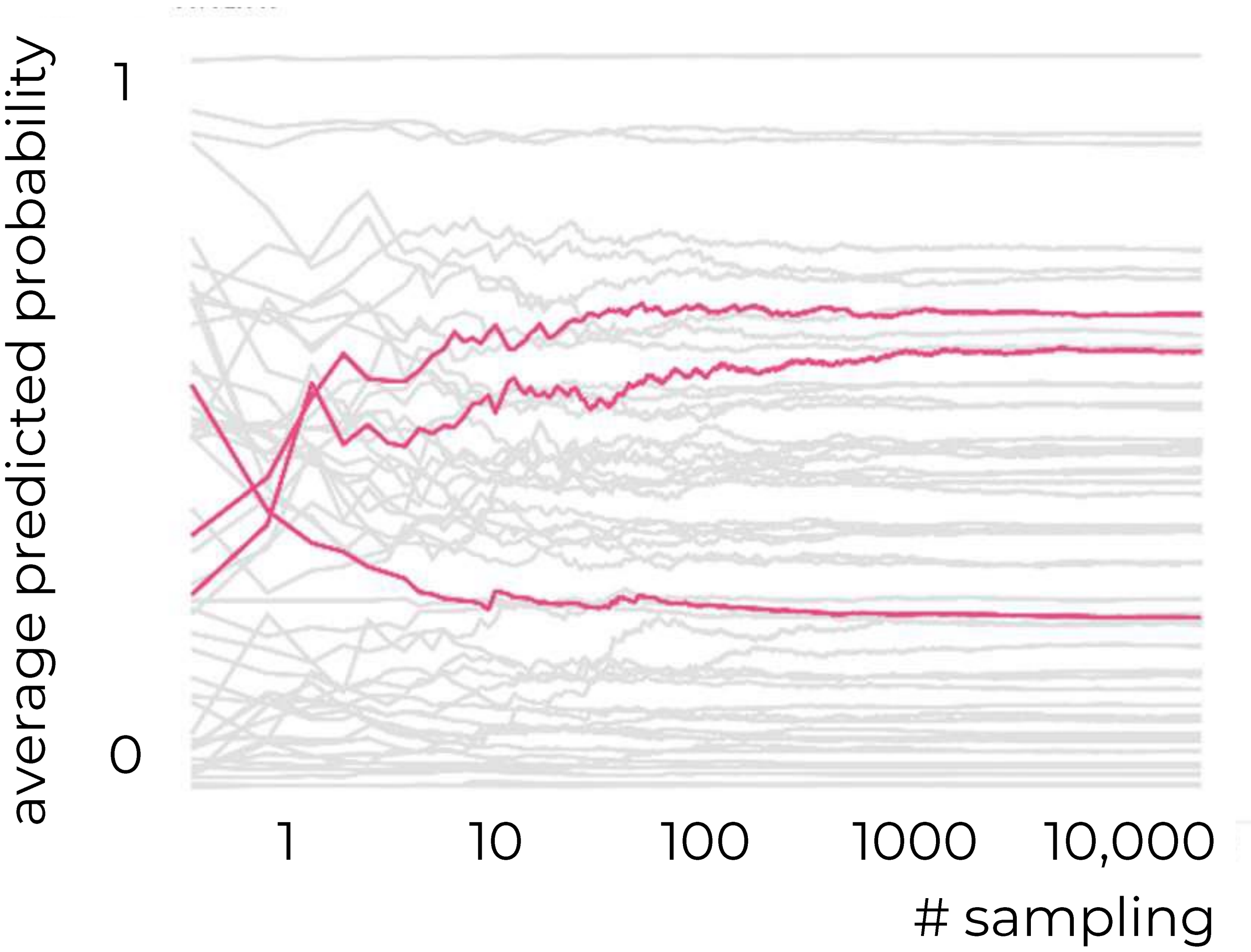
prediction accuracy

bioassay	FPR
sr.mmp	25.1%
sr.p53	25.4%
nr.ahr	41.8%
...	...
nr.ar	82.4%
nr.er	85.0%

MassBank & MoNA

748 compounds with MS^2 & tox

handling probabilistic fingerprints



Monte Carlo sampling

- sampling each fingerprint

prediction changed with # of samples

- 2%...25%
- more prone for multi-output

anneli.kruve@su.se

MS2Quant

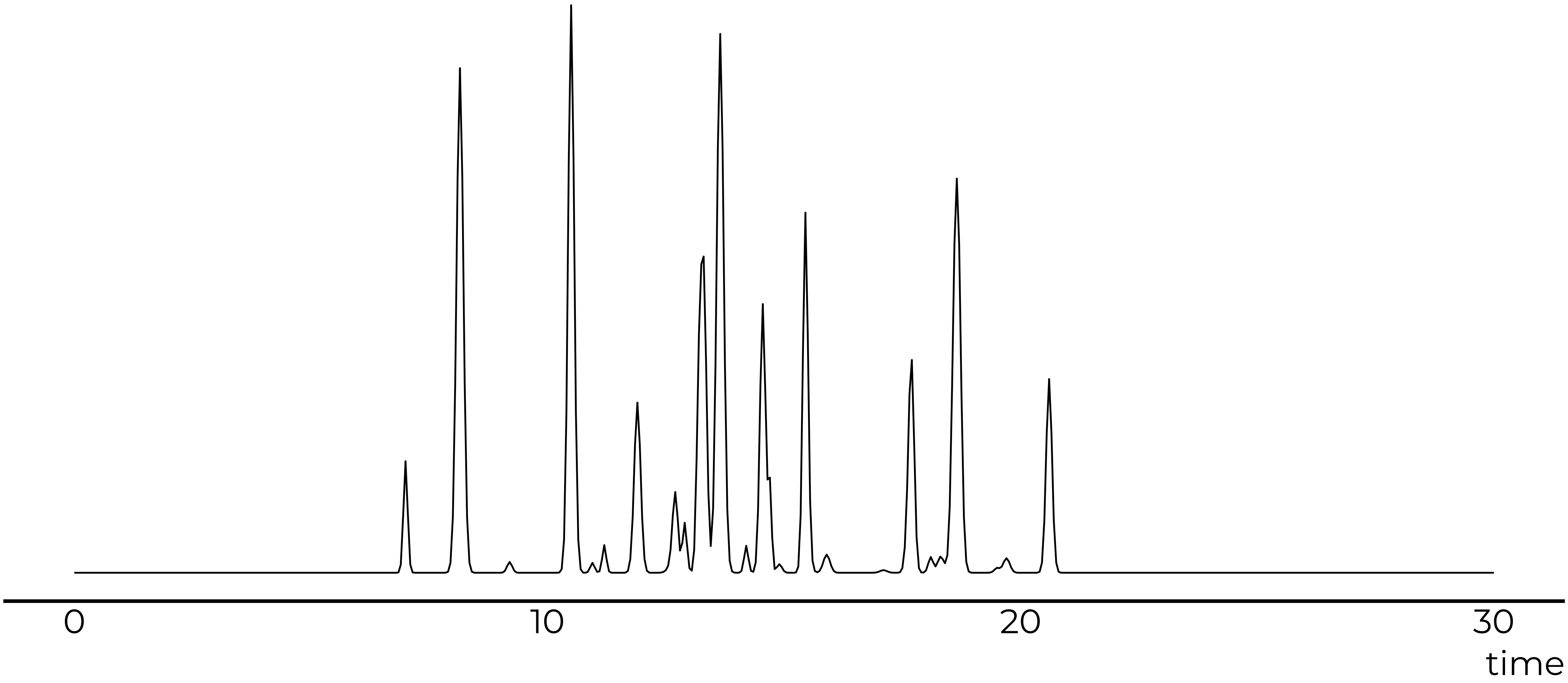
quantification in ESI/HRMS

anneli kruve

anneli.kruve@su.se

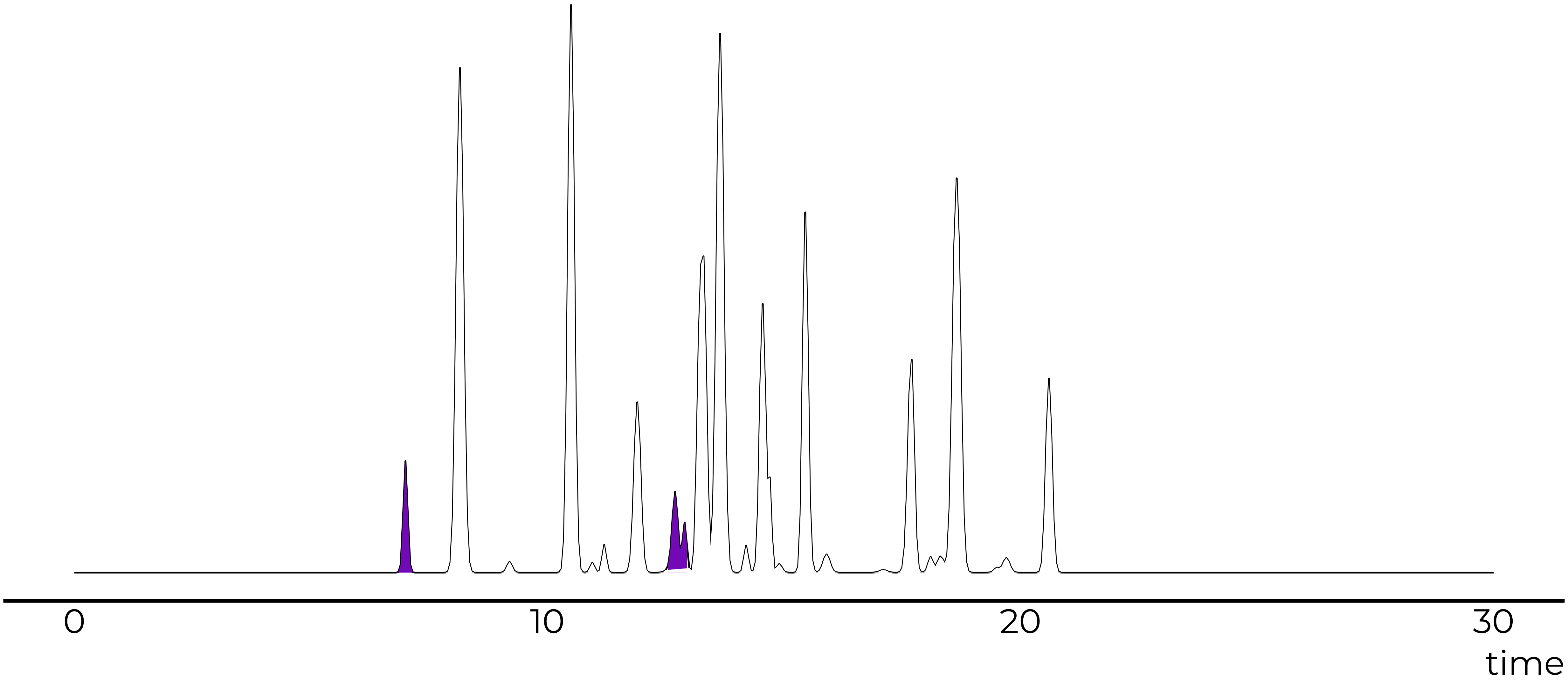
quantification in ESI/HRMS

Malm et al. Molecules 2021



quantification in ESI/HRMS

Malm et al. Molecules 2021

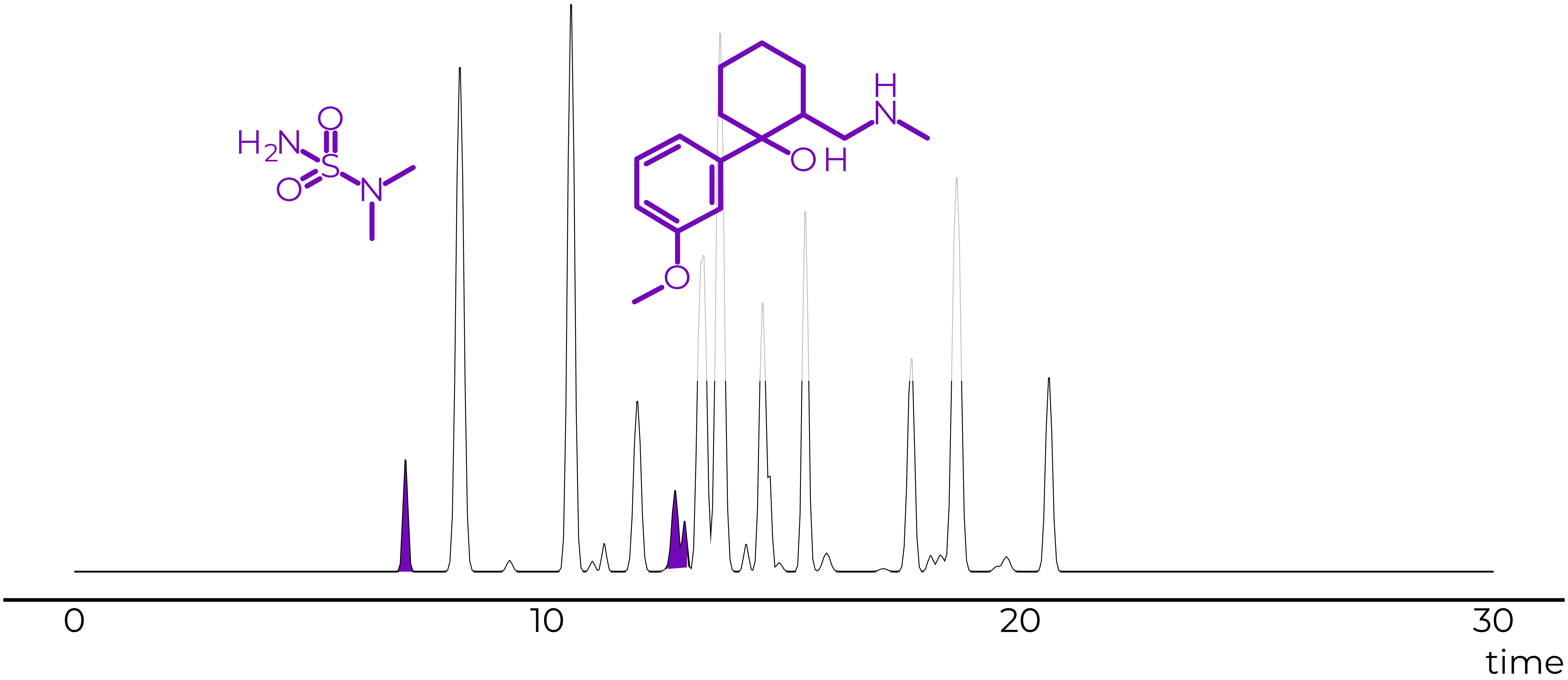


anneli kruve

anneli.kruve@su.se

quantification in ESI/HRMS

Malm et al. Molecules 2021

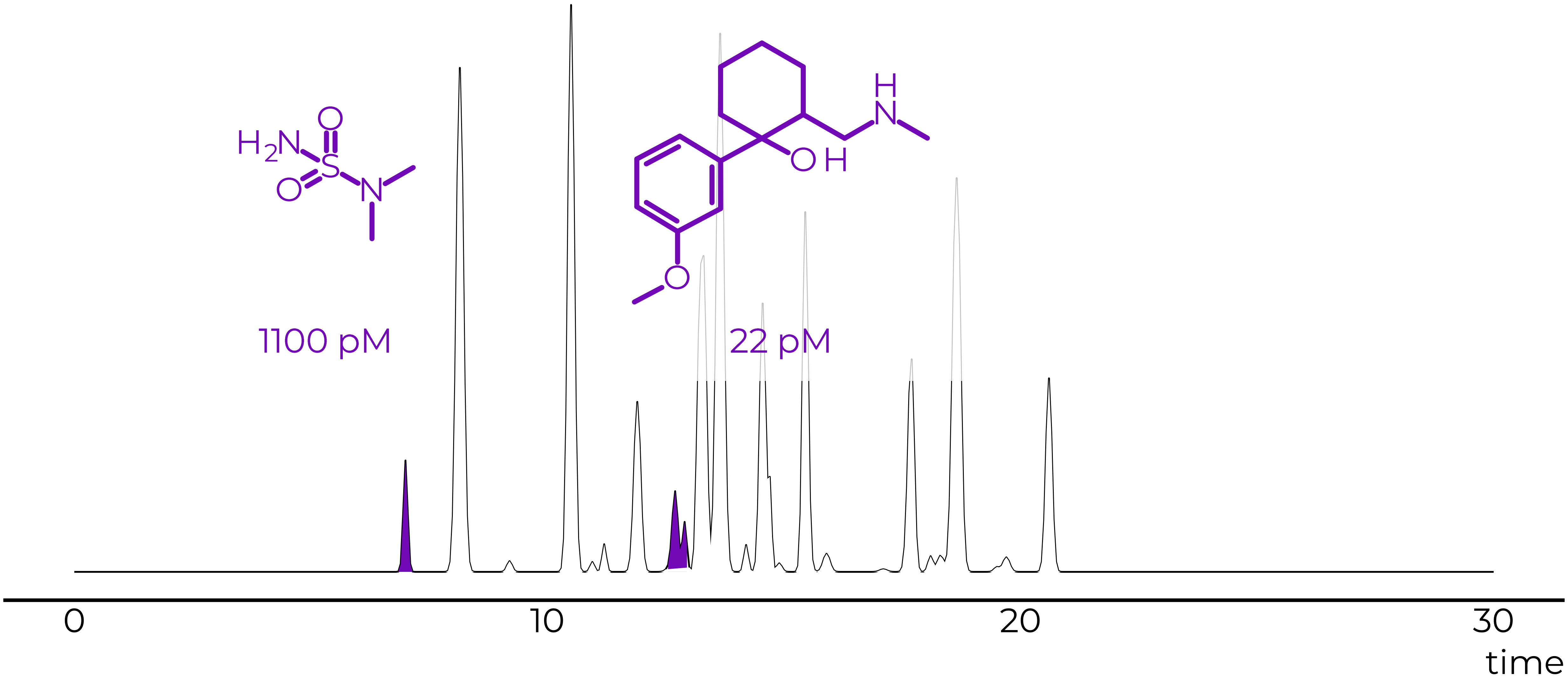


anneli kruve

anneli.kruve@su.se

quantification in ESI/HRMS

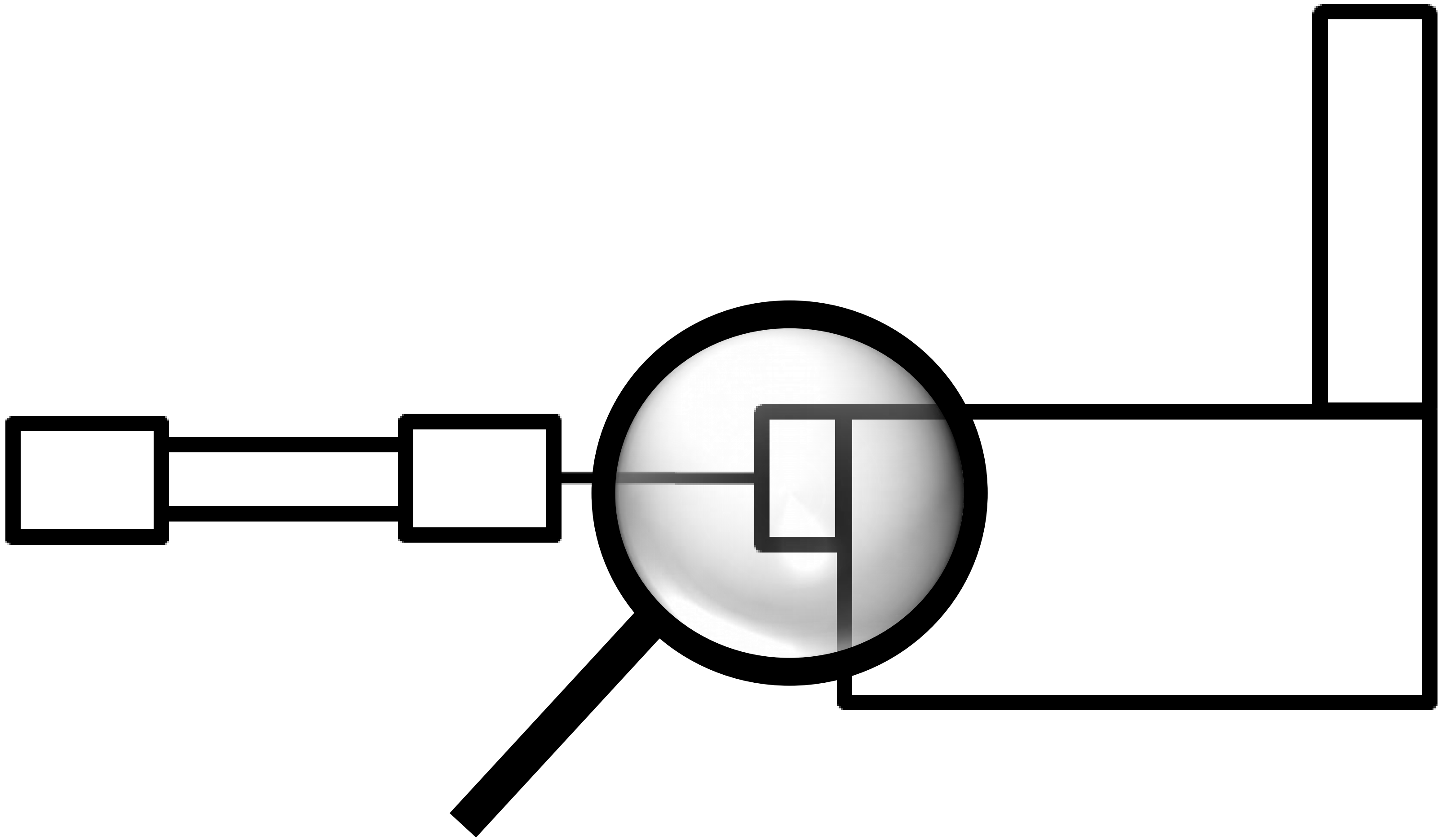
Malm et al. Molecules 2021



anneli kruve

anneli.kruve@su.se

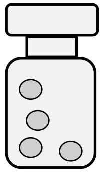
electrospray



anneli kruve

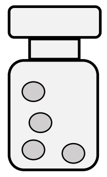
anneli.kruve@su.se

workflow

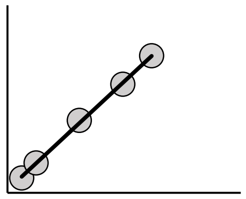


flow injections

workflow

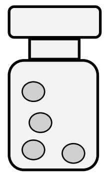


flow injections

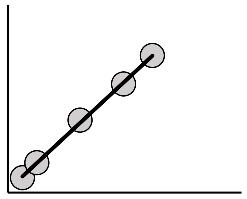


calibration graph

workflow



flow injections



calibration graph

$$\frac{slope_1}{slope_2} \rightarrow IE$$

relative measurements

anneli.kruve@su.se

structure

anneli.kruve@su.se

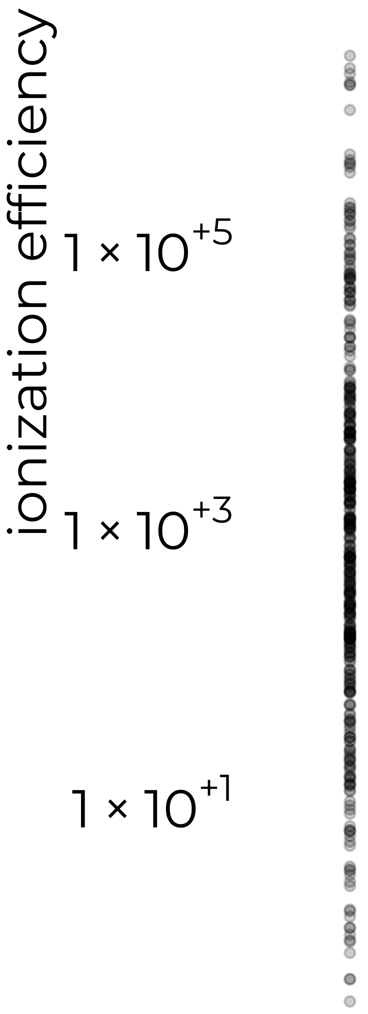
structure

one solvent, purely analyte properties

377 chemicals

anneli.kruve@su.se

structure

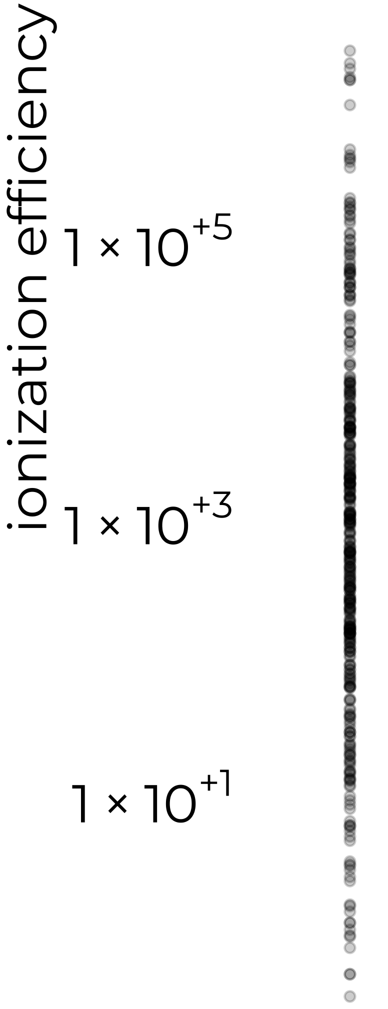


one solvent, purely analyte properties

377 chemicals

anneli.kruve@su.se

structure



one solvent, purely analyte properties

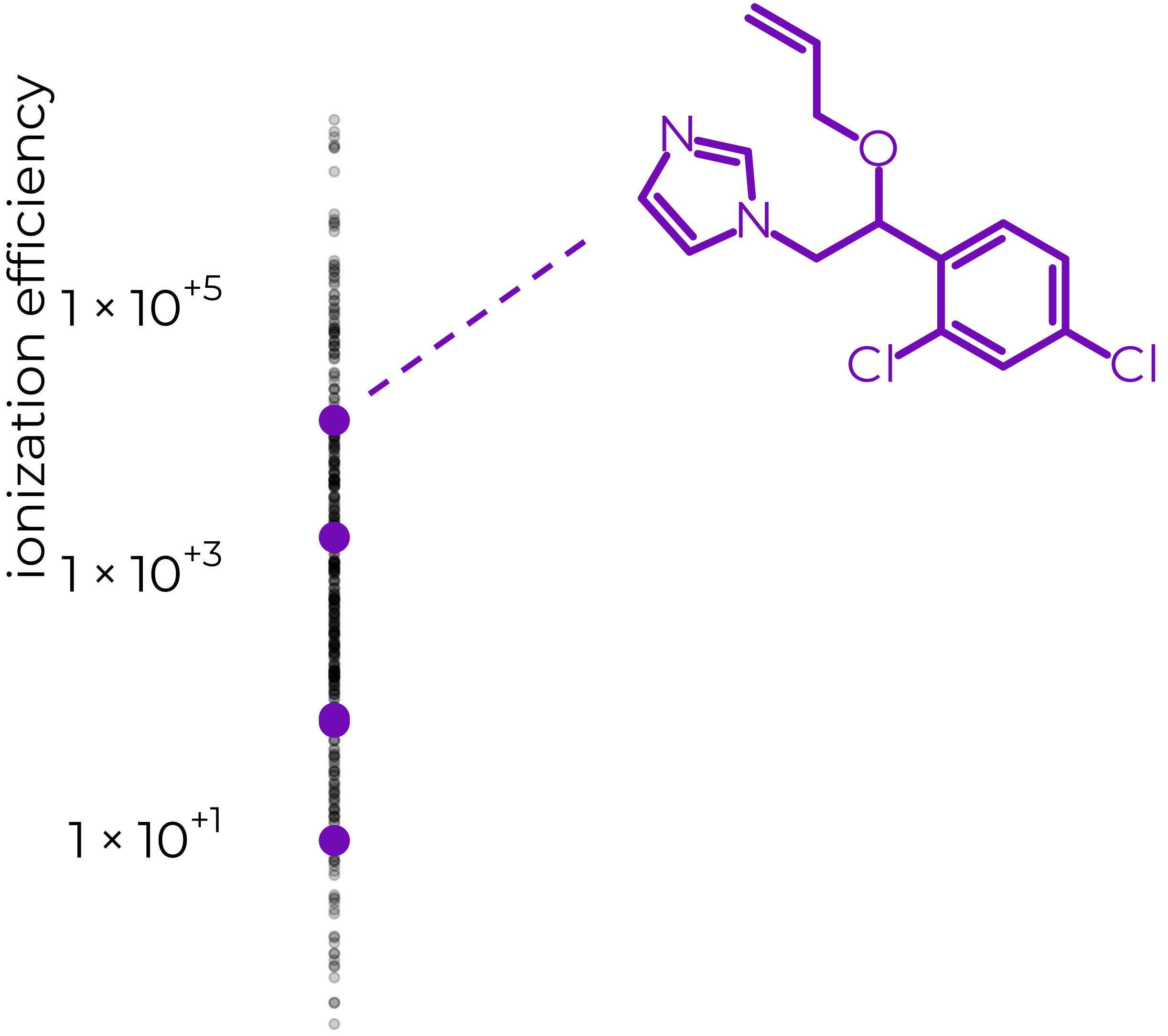
377 chemicals

10,000,000x difference in ionization efficiency

anneli kruve

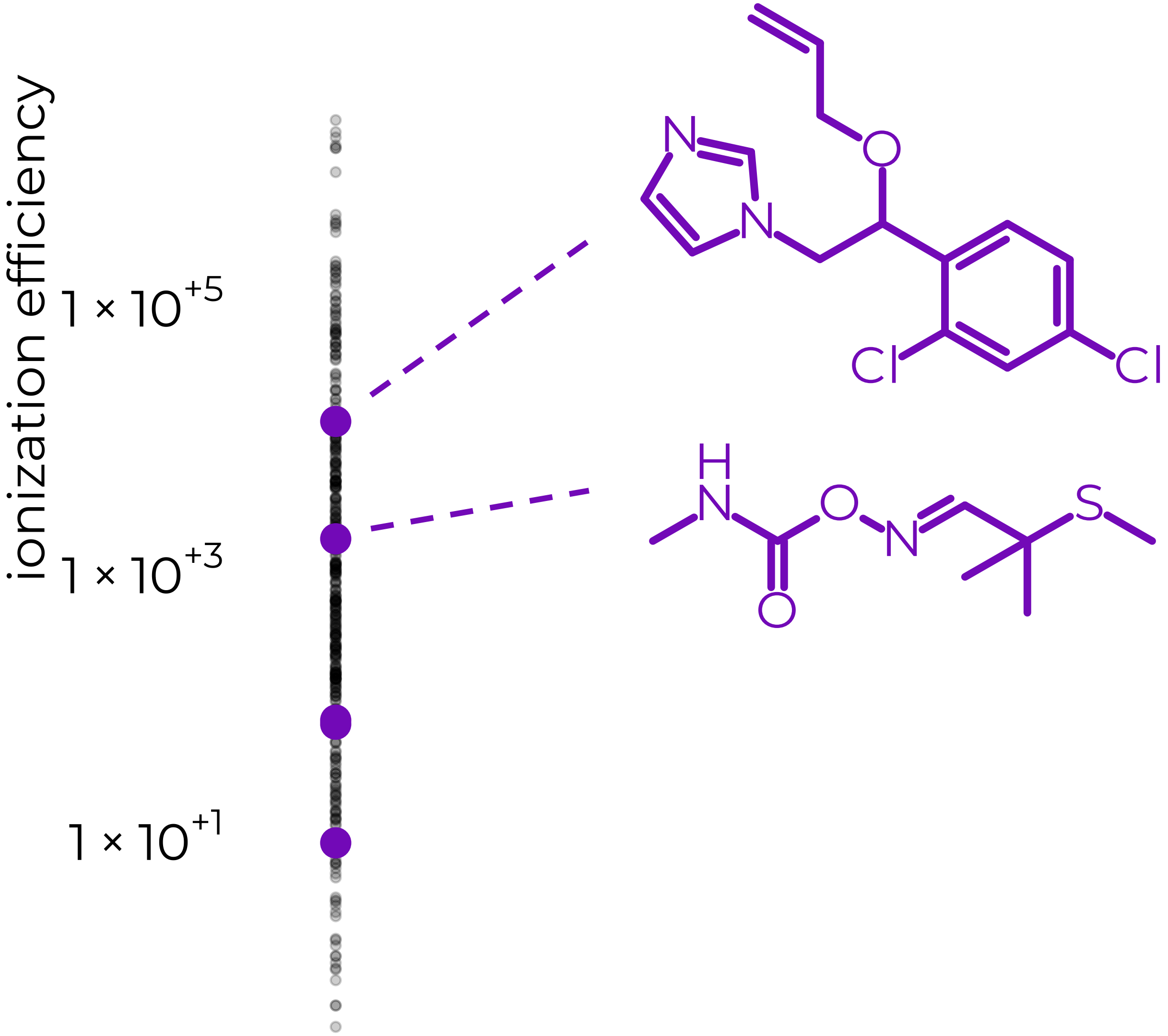
anneli.kruve@su.se

structure



anneli.kruve@su.se

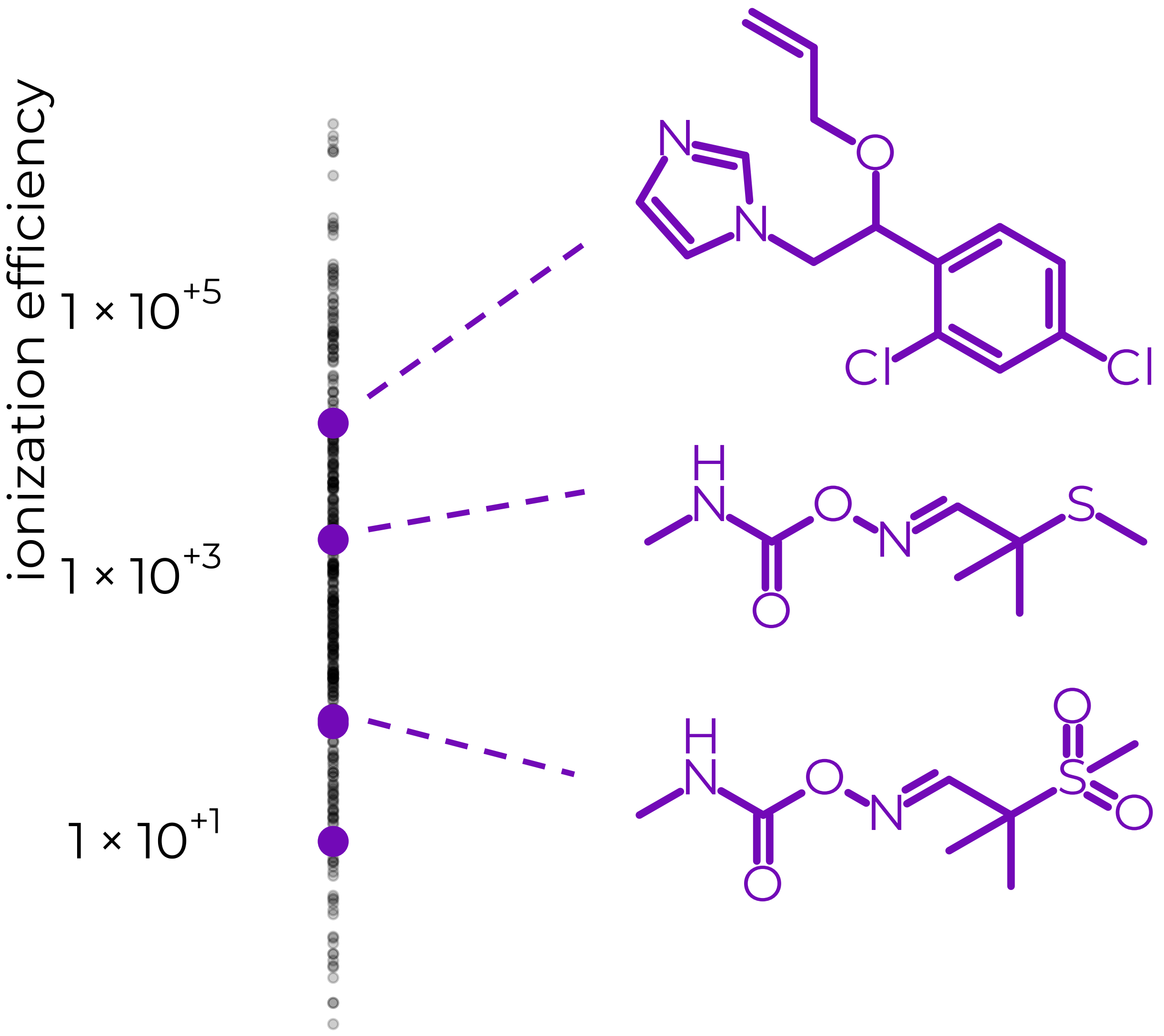
structure



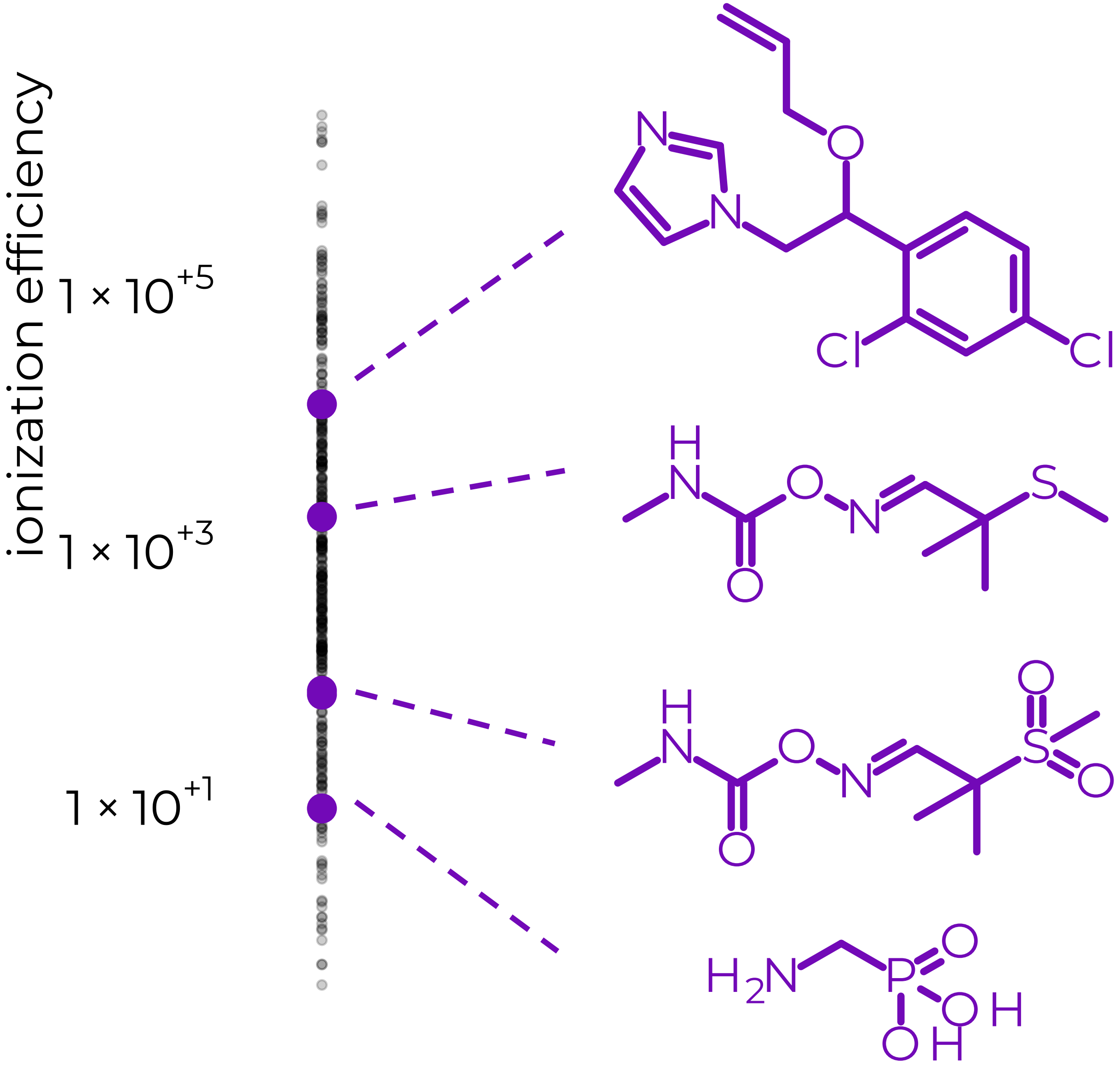
anneli kruve

anneli.kruve@su.se

structure


ionization efficiency
1 × 10+5
1 × 10+3
1 × 10+1
N
N
O
Cl
Cl
H
N
O
N
S
O
H
N
O
N
O
S
O
O

structure



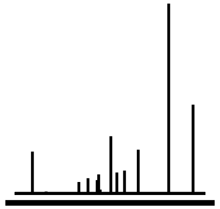
anneli kruve

anneli.kruve@su.se

quantification

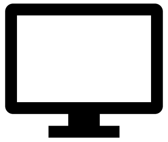
with machine learning

workflow



MS^2 spectra

molecular fingerprints with SIRIUS



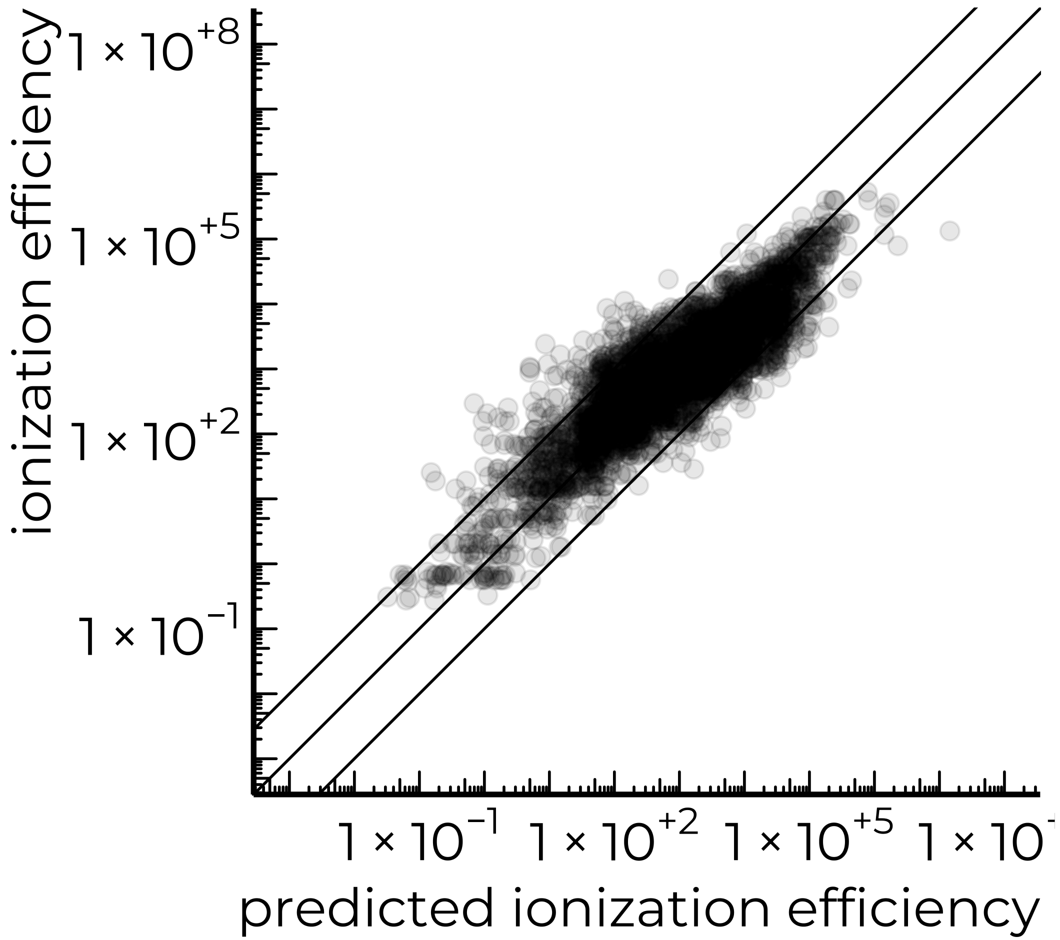
predict toxicity and ionization efficiency

anneli kruve

anneli.kruve@su.se

performance

Sepman et al. Anal Chem 2023



IE range

100,000,000

training set

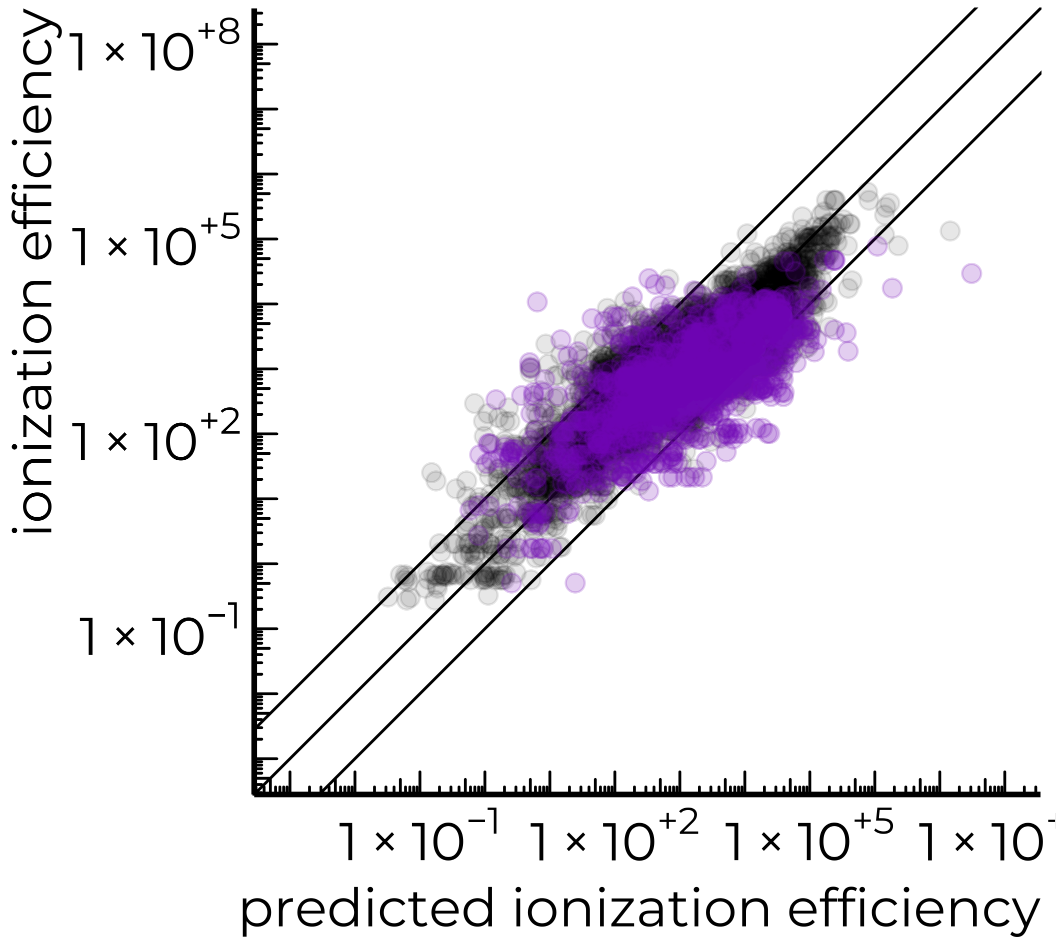
RMSE 3.6x

anneli kruve

anneli.kruve@su.se

performance

Sepman et al. Anal Chem 2023



IE range

100,000,000

training set

RMSE 3.6x

test set

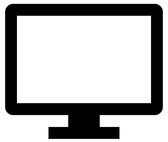
RMSE 5.6x

application

compound	peak area
methiocarb sulfoxide	5,300
pyridaben	5,400
aldicarb-sulfone	70,800

anneli.kruve@su.se

application



predict ionization efficiency

anneli kruve

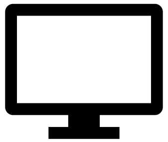
anneli.kruve@su.se

application

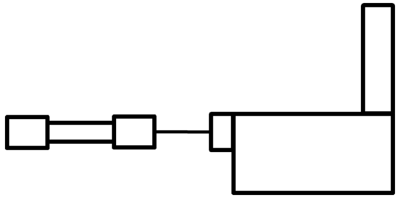
compound	peak area	log*IE*$_{pred}$
methiocarb sulfoxide	5,300	2.57
pyridaben	5,400	3.78
aldicarb-sulfone	70,800	1.99

anneli.kruve@su.se

application



predict ionization efficiency



convert to instrument specific values

anneli kruve

anneli.kruve@su.se

application

compound	**peak area**	**log*IE*$_{pred}$**	**c (nM)**
methiocarb sulfoxide	5,300	2.57	
pyridaben	5,400	3.78	
aldicarb-sulfone	70,800	1.99	
atrazine-D5			4.5
gabapentin-lactam			0.35
sitagliptin			0.23
5-methyl-1H-benzotriazole			0.94
neburon			3.4
caffeine			0.50

application

compound	peak area	log*IE*$_{pred}$	c (nM)
methiocarb sulfoxide	5,300	2.57	
pyridaben	5,400	3.78	
aldicarb-sulfone	70,800	1.99	
atrazine-D5	450,000		4.5
gabapentin-lactam	10,400		0.35
sitagliptin	8,100		0.23
5-methyl-1H-benzotriazole	27,000		0.94
neburon	243,000		3.4
caffeine	5,600		0.50

anneli.kruve@su.se

application

$RF_{measured}$ = peak area $/c$

compound	**peak area**	**log*IE*$_{pred}$**	**c (nM)**	***RF*$_{meas}$·10^{16}**
methiocarb sulfoxide	5,300	2.57		
pyridaben	5,400	3.78		
aldicarb-sulfone	70,800	1.99		
atrazine-D5	450,000		4.5	9.8
gabapentin-lactam	10,400		0.35	3.0
sitagliptin	8,100		0.23	3.5
5-methyl-1H-benzotriazole	27,000		0.94	2.9
neburon	243,000		3.4	7.2
caffeine	5,600		0.50	1.1

application

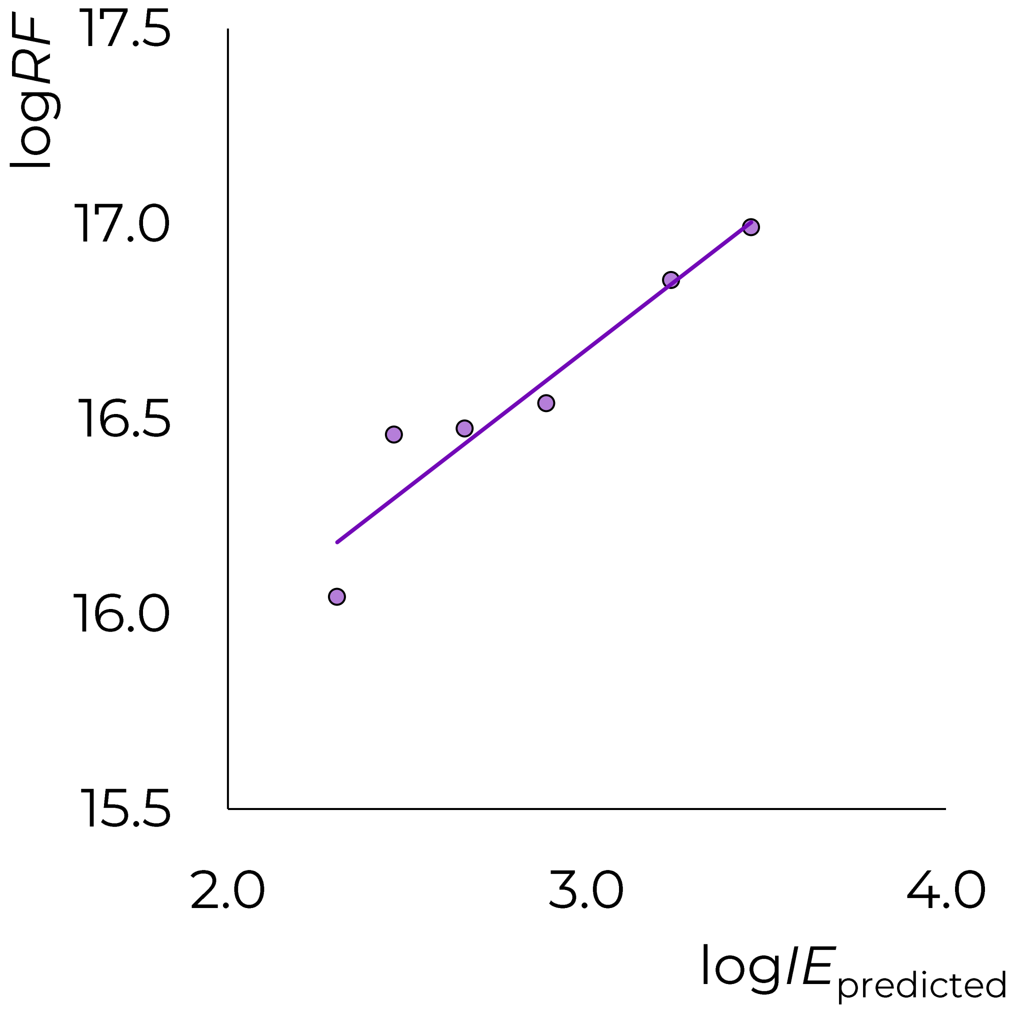
compound	peak area	$\log IE_{pred}$	c (nM)	$RF_{meas} \cdot 10^{16}$
methiocarb sulfoxide	5,300	2.57		
pyridaben	5,400	3.78		
aldicarb-sulfone	70,800	1.99		
atrazine-D5	450,000	3.46	4.5	9.8
gabapentin-lactam	10,400	2.66	0.35	3.0
sitagliptin	8,100	2.89	0.23	3.5
5-methyl-1H-benzotriazole	27,000	2.46	0.94	2.9
neburon	243,000	3.23	3.4	7.2
caffeine	5,600	2.30	0.50	1.1

anneli.kruve@su.se

application

compound	peak area	$\log IE_{pred}$
methiocarb sulfoxide	5,300	2.57
pyridaben	5,400	3.78
aldicarb-sulfone	70,800	1.99
atrazine-D5	450,000	3.46
gabapentin-lactam	10,400	2.66
sitagliptin	8,100	2.89
5-methyl-1H-benzotriazole	27,000	2.46
neburon	243,000	3.23
caffeine	5,600	2.30

$$\log RF = \text{slope} \cdot \log IE + \text{intercept}$$



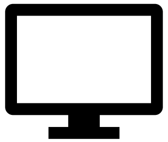
anneli kruve
anneli.kruve@su.se

application

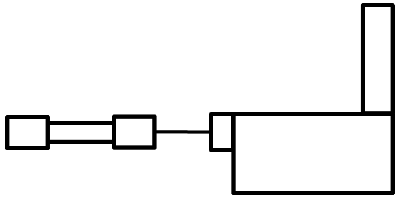
$\log RF_{\text{predicted}} = \text{slope} \cdot \log IE_{\text{predicted}} + \text{intercept}$

compound	**peak area**	**log*IE*$_{pred}$**	**c (nM)**	***RF*$_{meas}$$\cdot 10^{16}$**	***RF*$_{pred}$$\cdot 10^{16}$**
methiocarb sulfoxide	5,300	2.57			2.6
pyridaben	5,400	3.78			15.5
aldicarb-sulfone	70,800	1.99			1.1
atrazine-D5	450,000	3.46	4.5	9.8	
gabapentin-lactam	10,400	2.66	0.35	3.0	
sitagliptin	8,100	2.89	0.23	3.5	
5-methyl-1H-benzotriazole	27,000	2.46	0.94	2.9	
neburon	243,000	3.23	3.4	7.2	
caffeine	5,600	2.30	0.50	1.1	

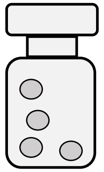
application



predict ionization efficiency



convert to instrument specific values



estimate concentration

anneli kruve

anneli.kruve@su.se

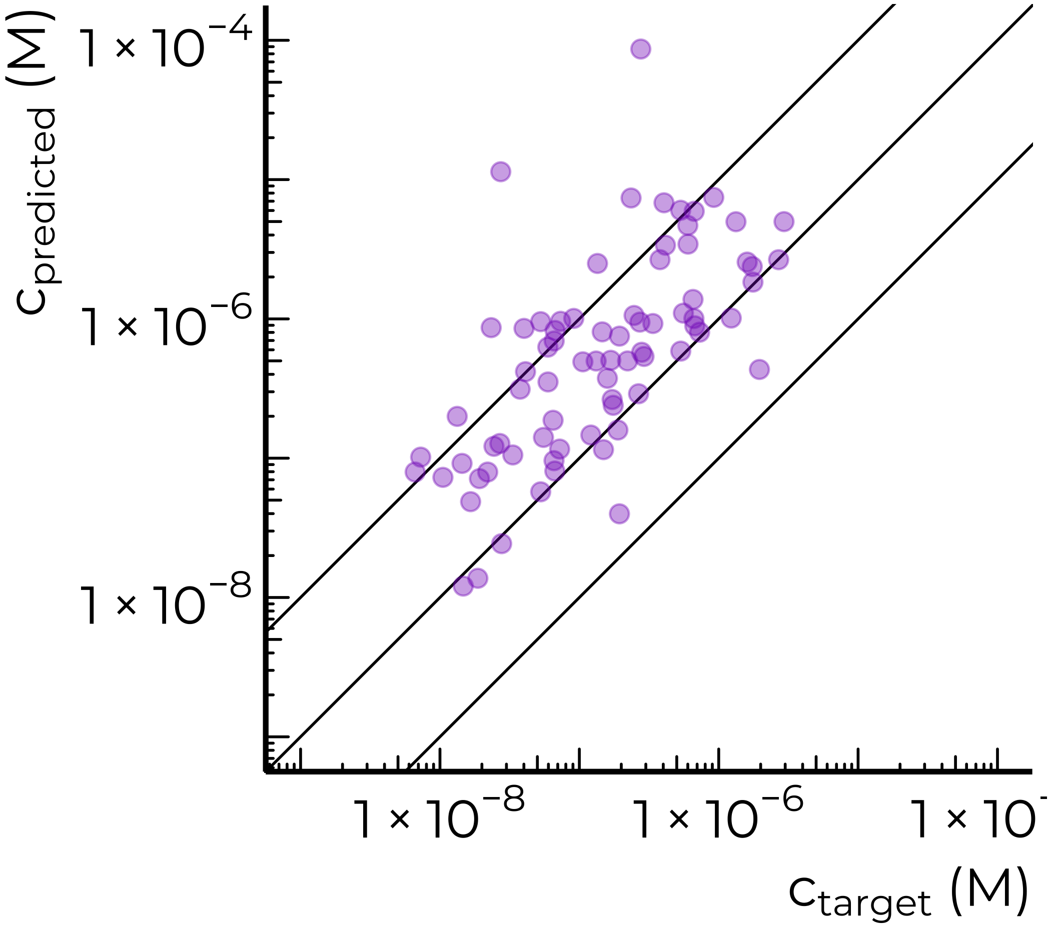
application

c = peak area / $RF_{predicted}$

compound	peak area	$\log IE_{pred}$	c (nM)	$RF_{meas} \cdot 10^{16}$	$RF_{pred} \cdot 10^{16}$	c_{pred} (nM)
methiocarb sulfoxide	5,300	2.57			2.6	0.20
pyridaben	5,400	3.78			15.5	0.035
aldicarb-sulfone	70,800	1.99			1.1	6.3
atrazine-D5	450,000	3.46	4.5	9.8		
gabapentin-lactam	10,400	2.66	0.35	3.0		
sitagliptin	8,100	2.89	0.23	3.5		
5-methyl-1H-benzotriazole	27,000	2.46	0.94	2.9		
neburon	243,000	3.23	3.4	7.2		
caffeine	5,600	2.30	0.50	1.1		

ionization efficiency

Sepman et al. Anal Chem 2023



mean prediction error

7.4x

geometric mean prediction error

4.5x

median prediction error

4.0x

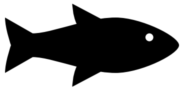
anneli kruve

anneli.kruve@su.se

summary

prioritization in NTS

toxicity



concentration

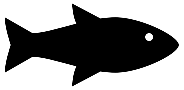
risk

anneli kruve

anneli.kruve@su.se

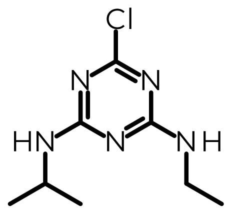
prioritization in NTS

toxicity

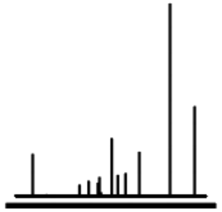


concentration

risk



structure



MS^2 spectrum

anneli kruve

anneli.kruve@su.se

kruvelab.com

anneli.kruve@su.se