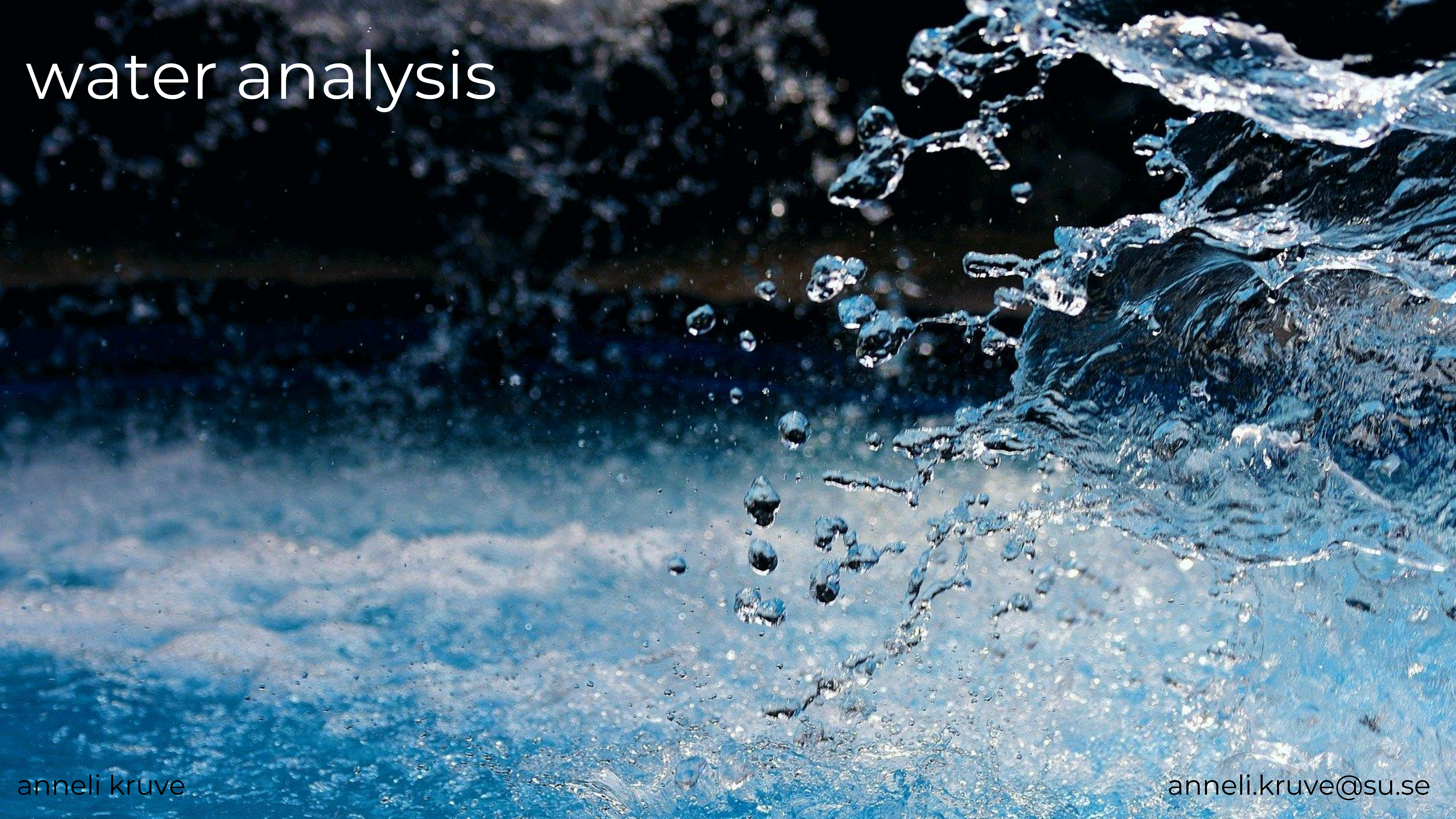
how can machine learning help us to evaluate the risk possessed by emerging contaminants?

anneli kruve
anneli.kruve@su.se
kruvelab.com

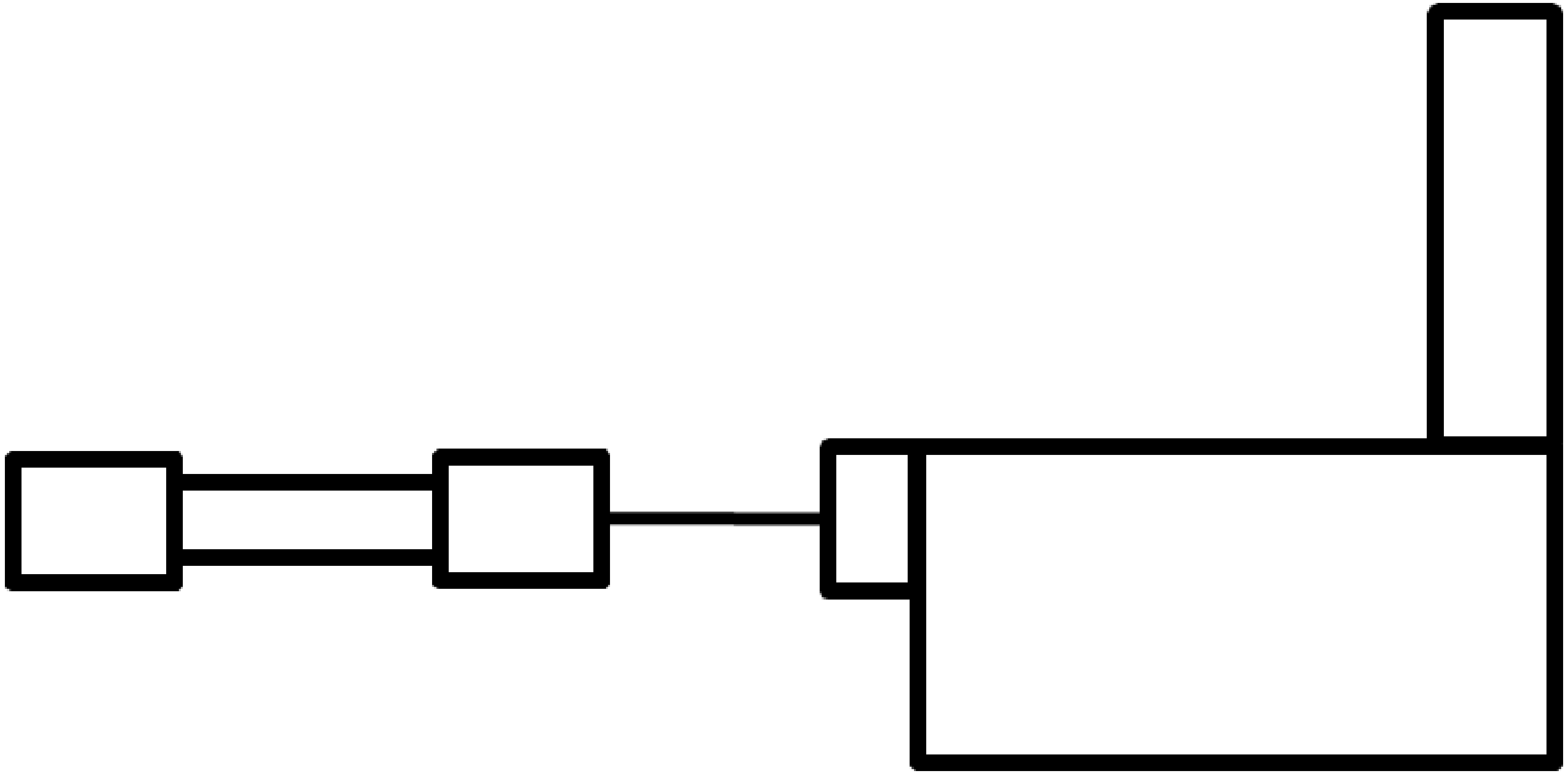


water analysis

anneli kruve

anneli.kruve@su.se

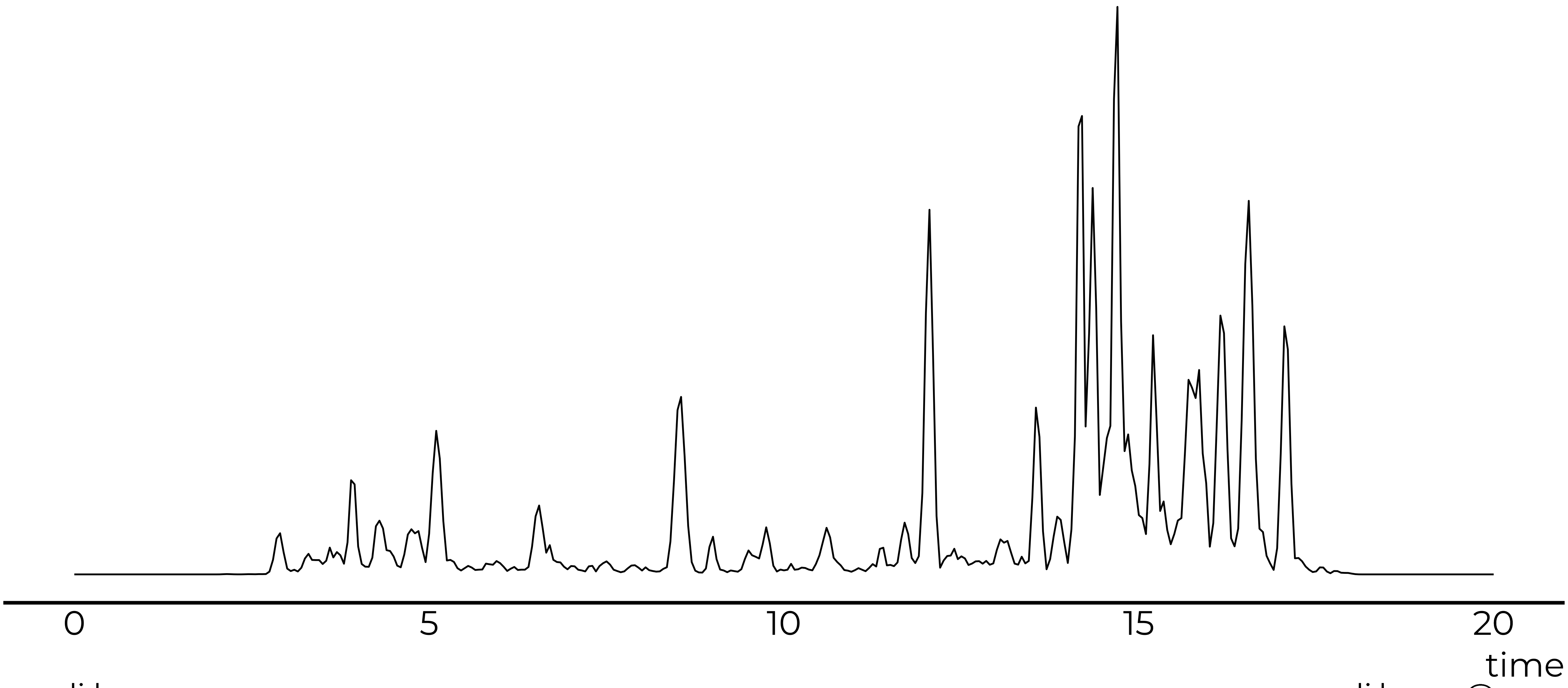
nontarget screening with LC/HRMS



anneli kruve

anneli.kruve@su.se

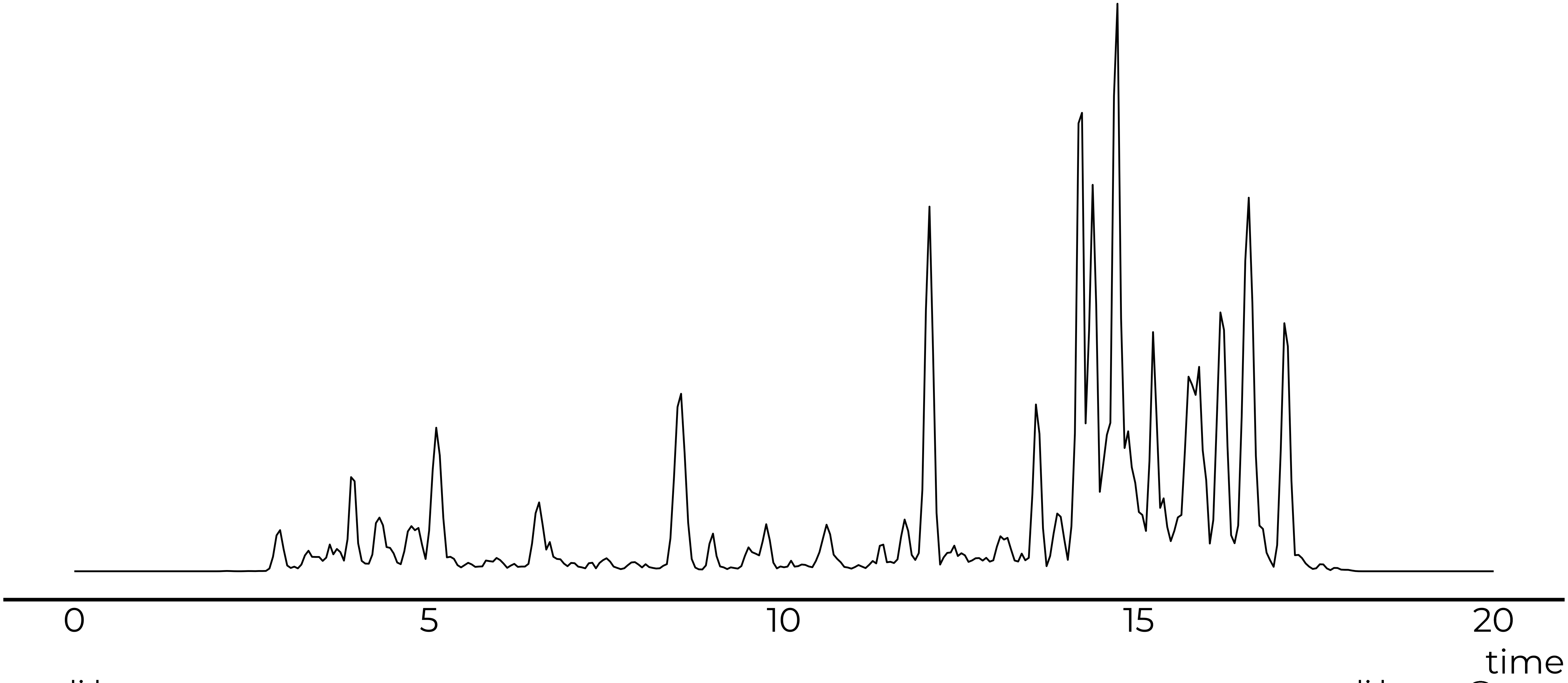
nontarget screening with LC/HRMS



anneli kruve

anneli.kruve@su.se

what next?



anneli kruve

anneli.kruve@su.se

prioritization

🐟 toxicity

prioritization

toxicity

concentration

prioritization

toxicity

concentration

risk

prioritization

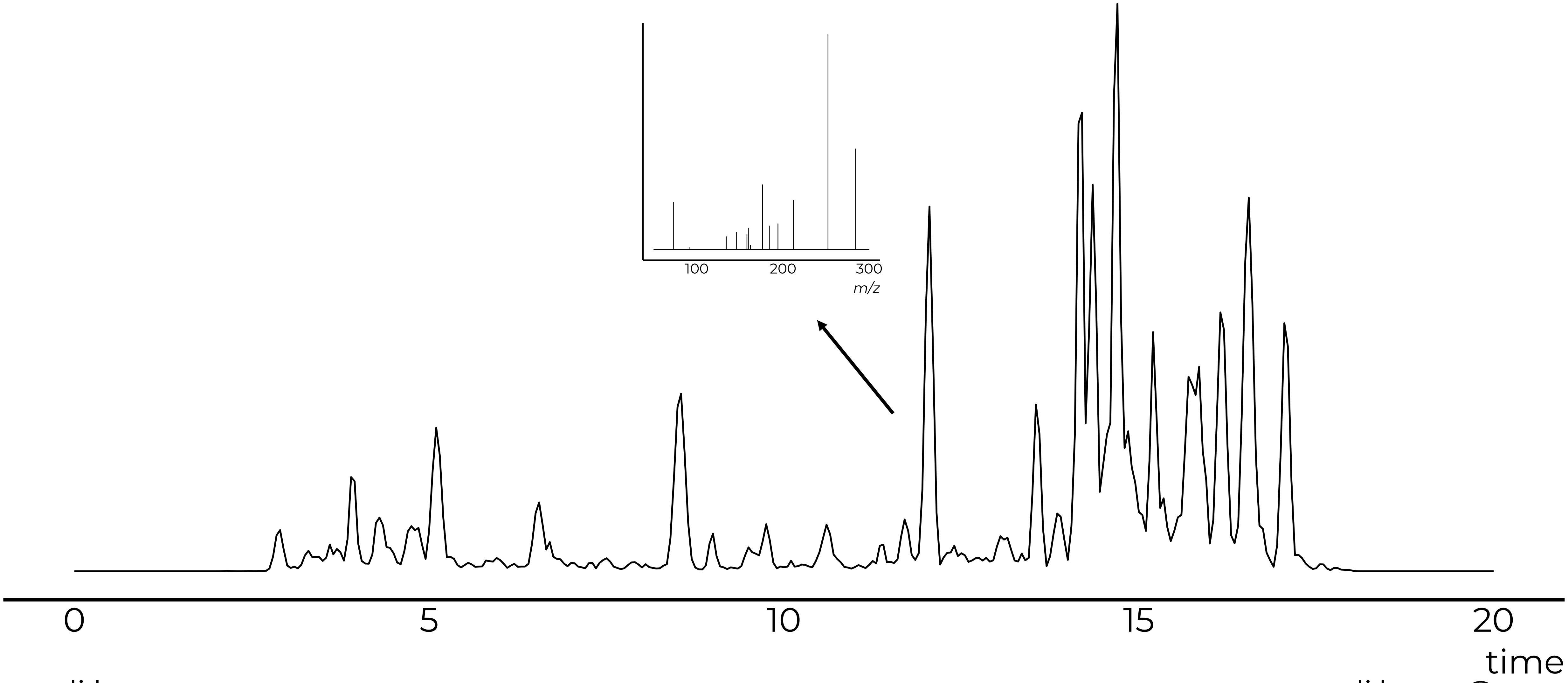
toxicity

concentration

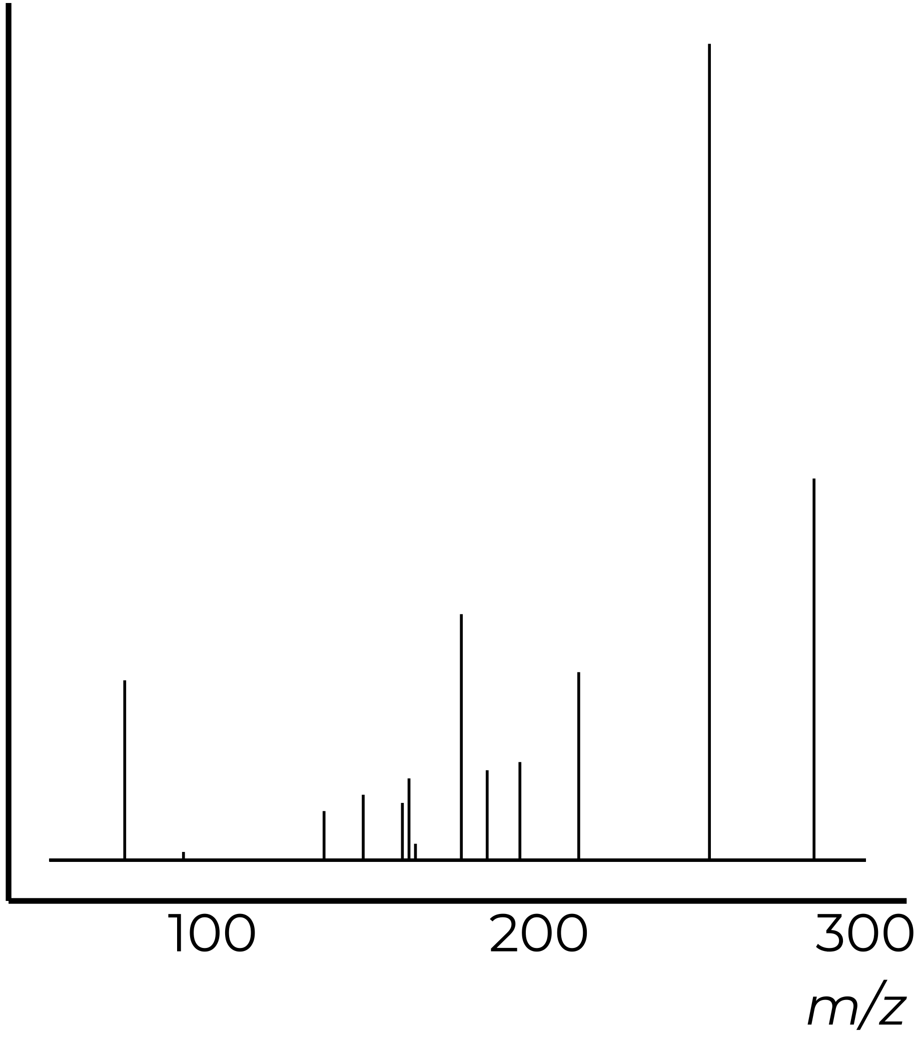
risk

$$\mathrm{PriorityScore} = \frac{c_{\mathrm{predicted}}}{\mathrm{AC}_{50}^{\mathrm{5th\ percentile}}}$$

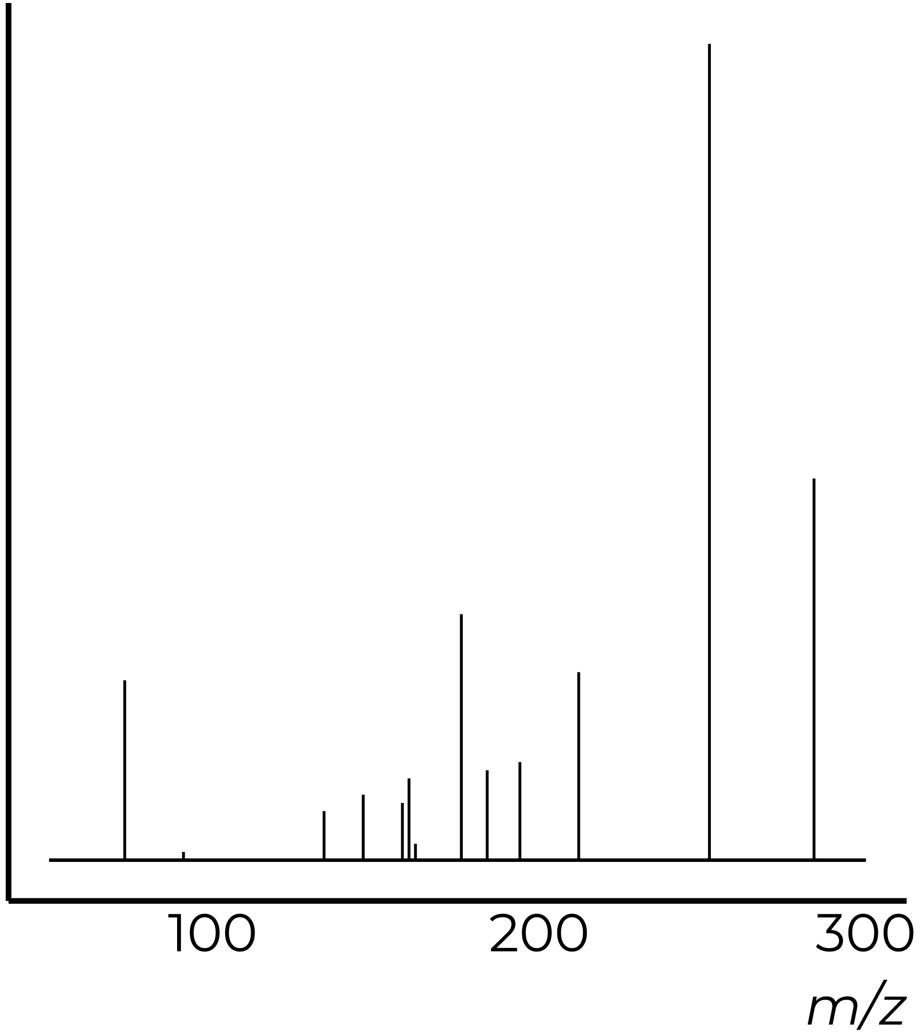
nontarget screening with LC/HRMS

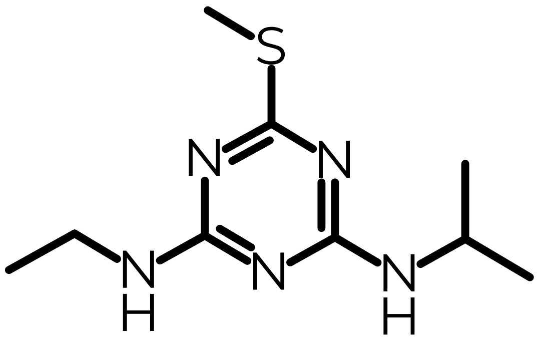


toxicity assessment

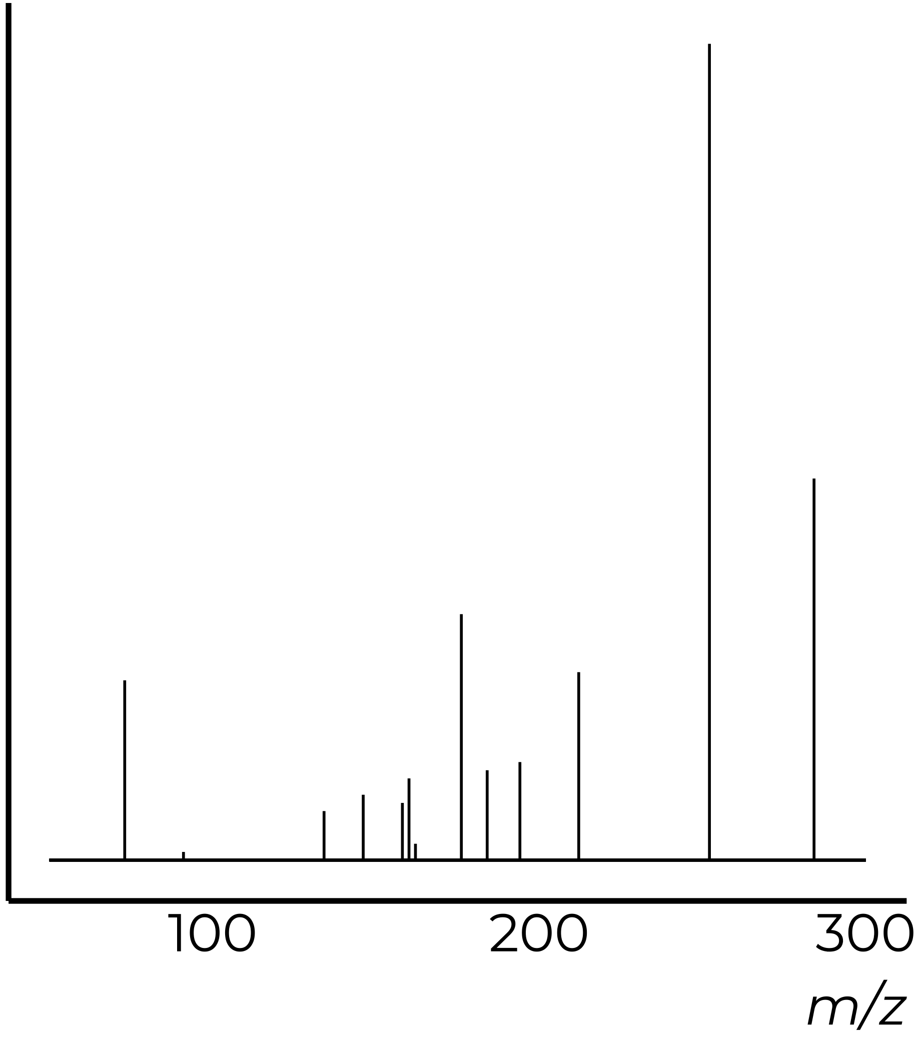


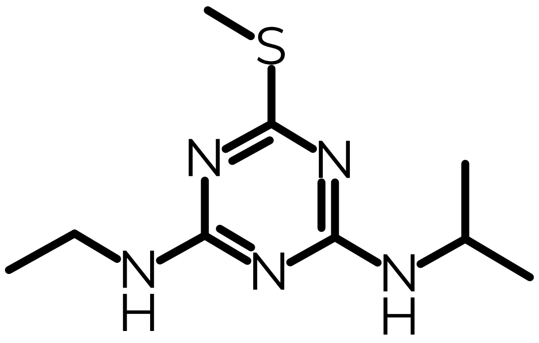
toxicity assessment





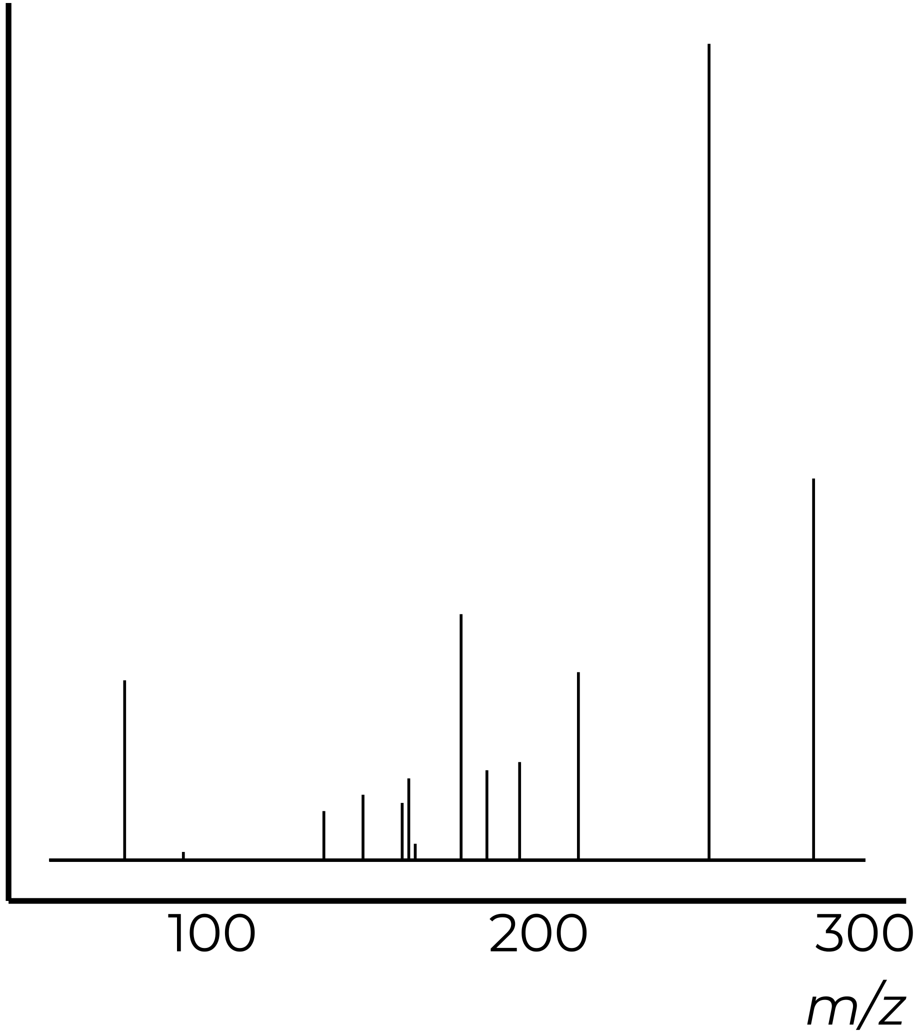
toxicity assessment

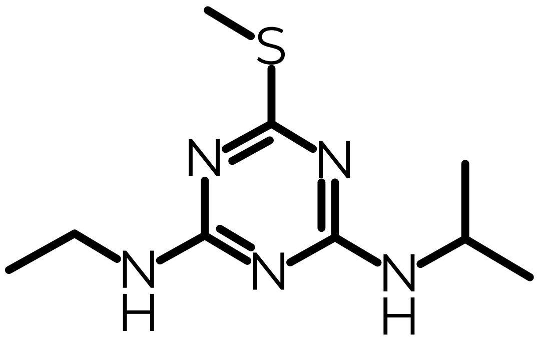




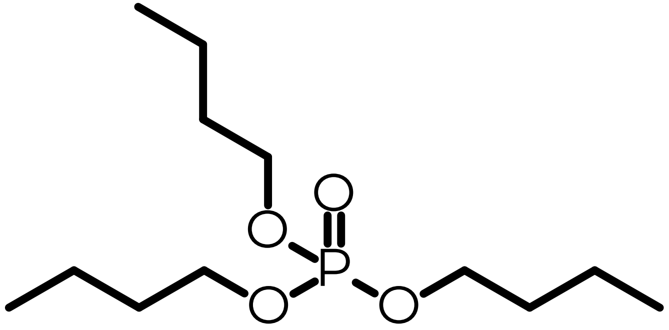
LC_{50} = 9.3 mg/L

toxicity assessment

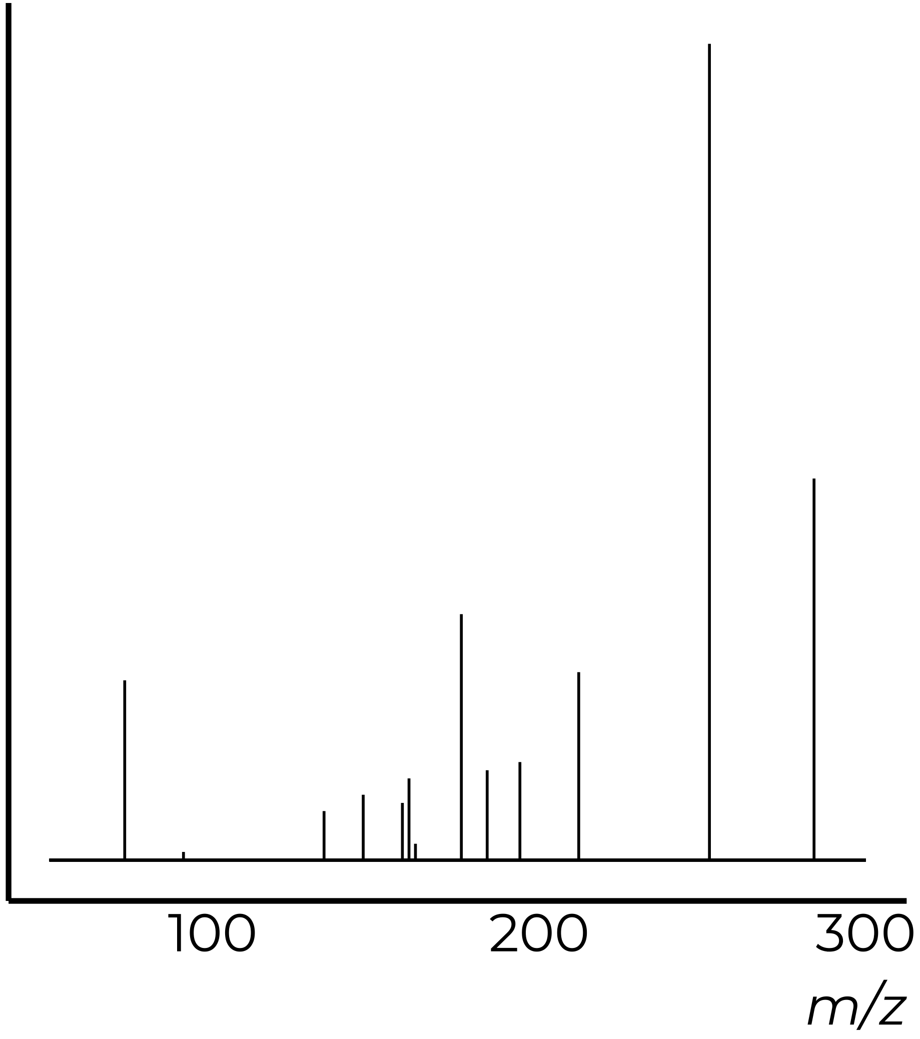


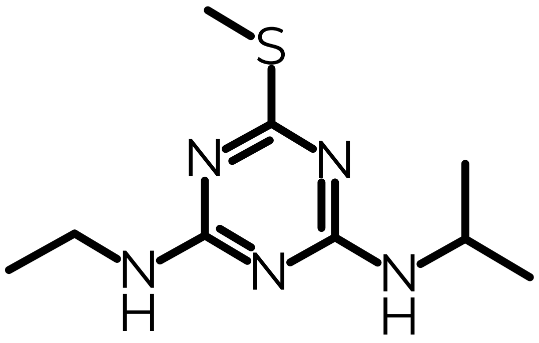


LC$_{50}$ = 9.3 mg/L

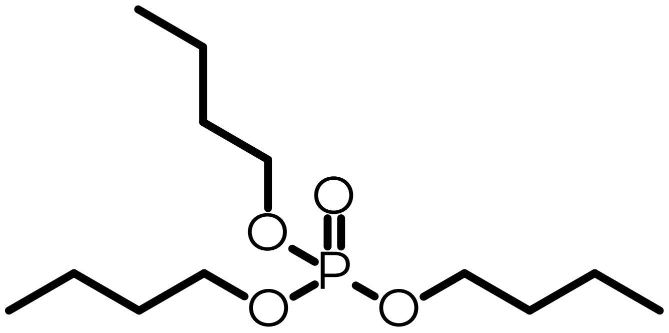


toxicity assessment



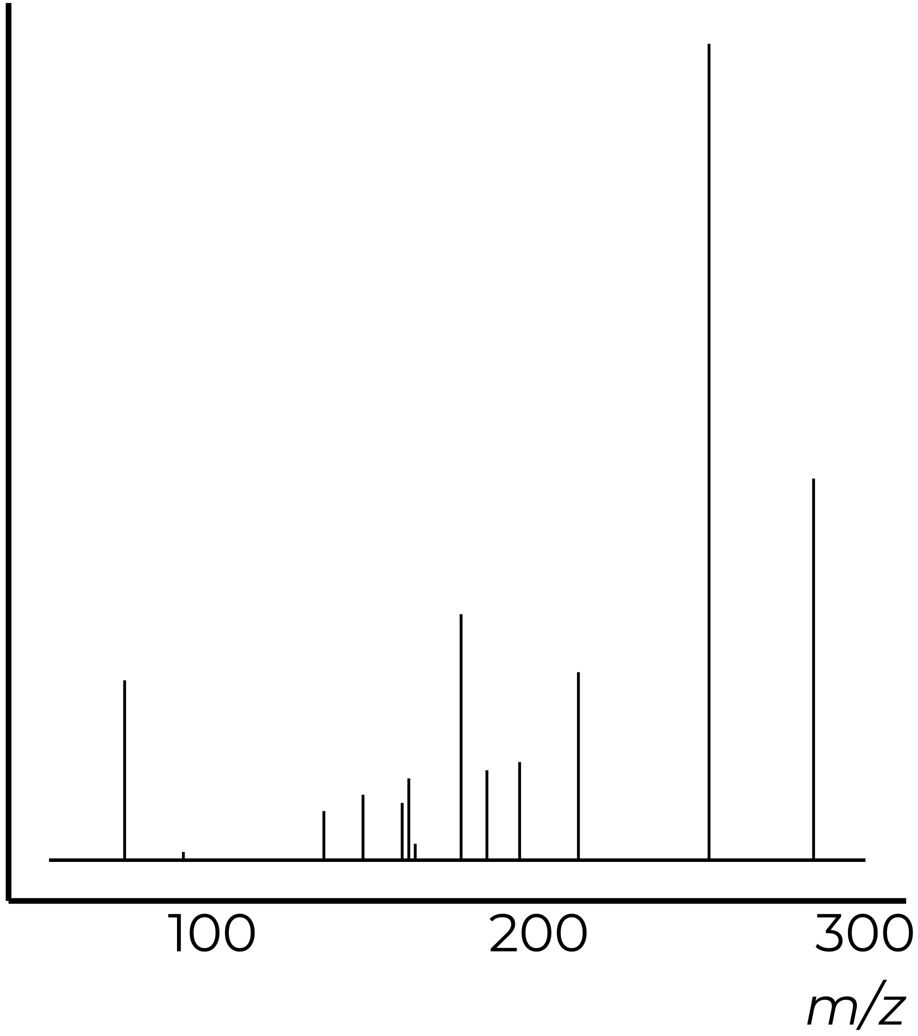


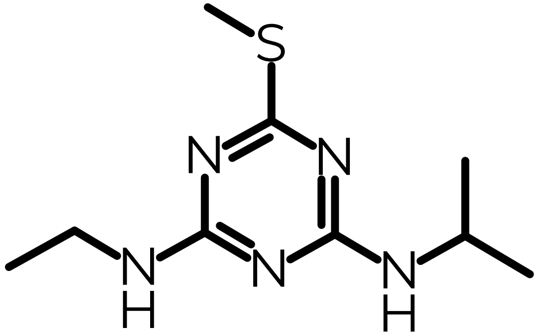
LC_{50} = 9.3 mg/L



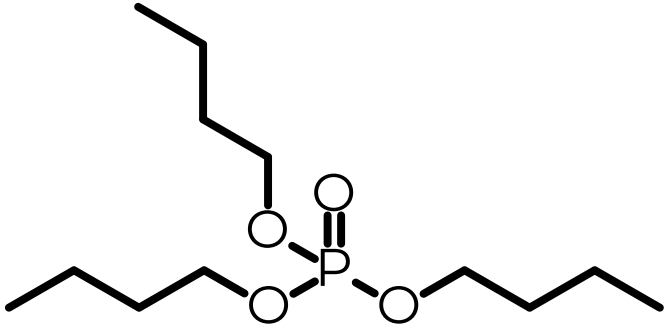
LC_{50} = ? mg/L

toxicity assessment





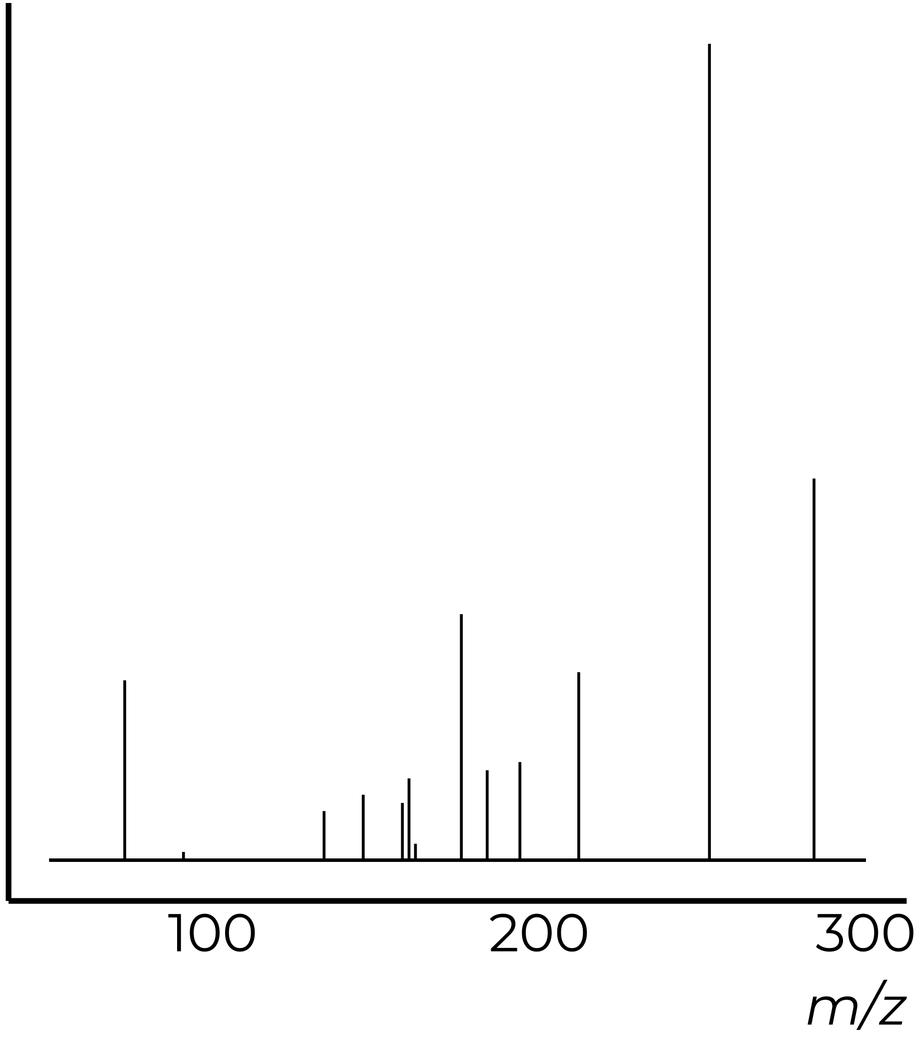
LC_{50} = 9.3 mg/L

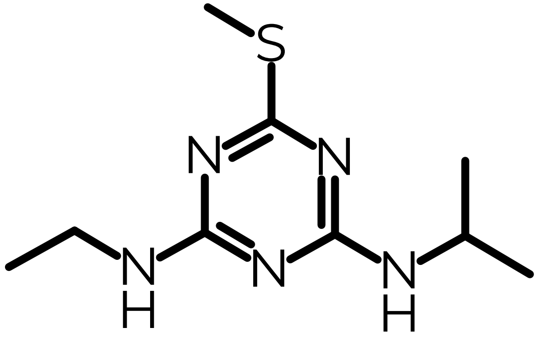


LC_{50} = ? mg/L

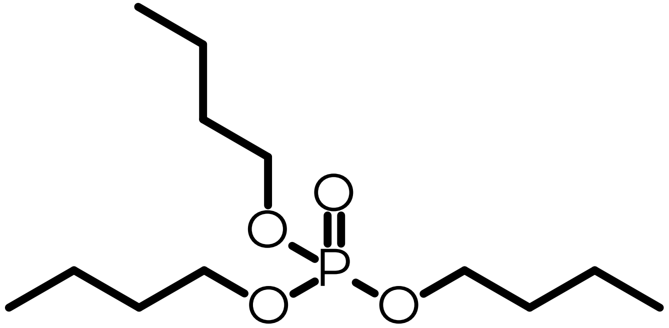
?

toxicity assessment





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

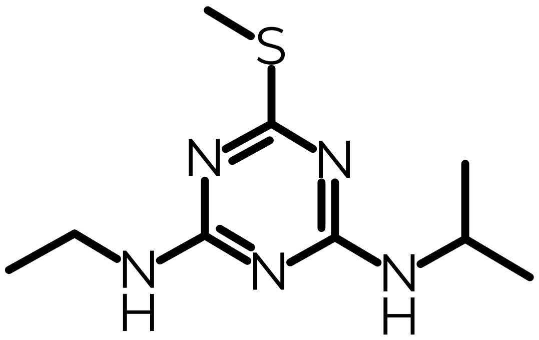


LC$_{50}$ = ? mg/L

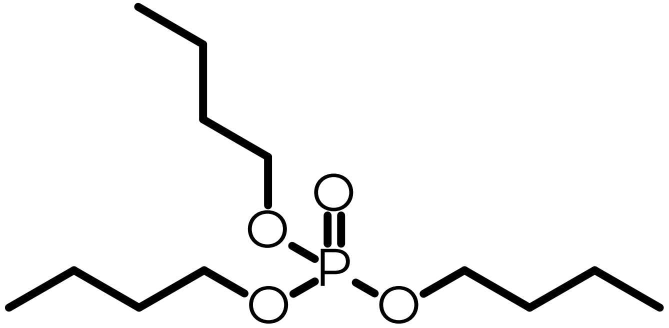
?

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

toxicity assessment



LC_{50} = 9.3 mg/L



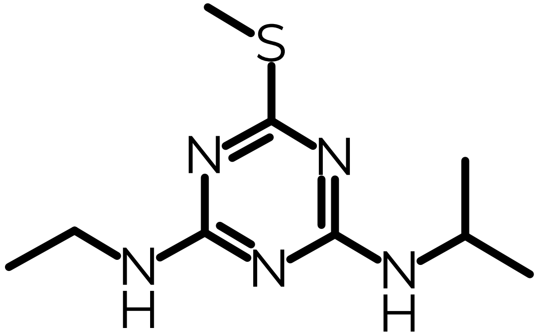
LC_{50} = ? mg/L

?

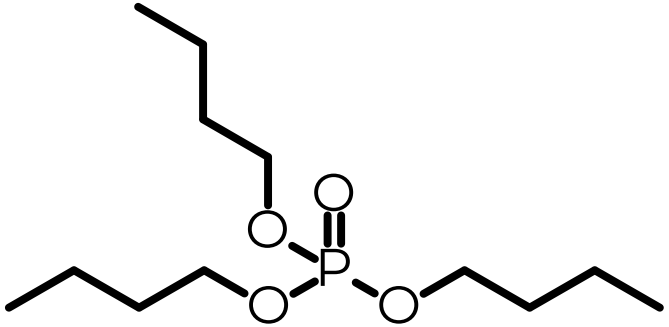
LC_{50} = ? mg/L

toxicity assessment

<1%



LC_{50} = 9.3 mg/L



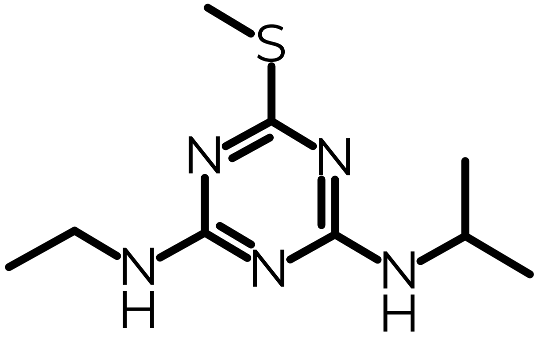
LC_{50} = ? mg/L

?

LC_{50} = ? mg/L

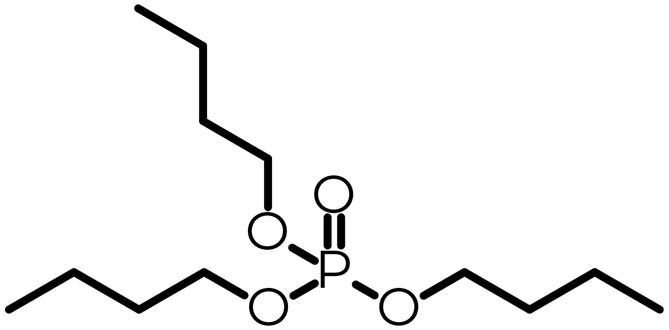
toxicity assessment

<1%



LC_{50} = 9.3 mg/L

<2%



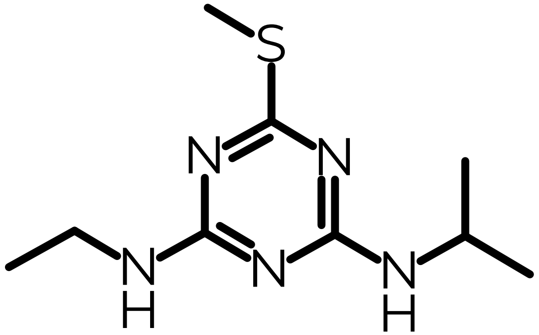
LC_{50} = ? mg/L

?

LC_{50} = ? mg/L

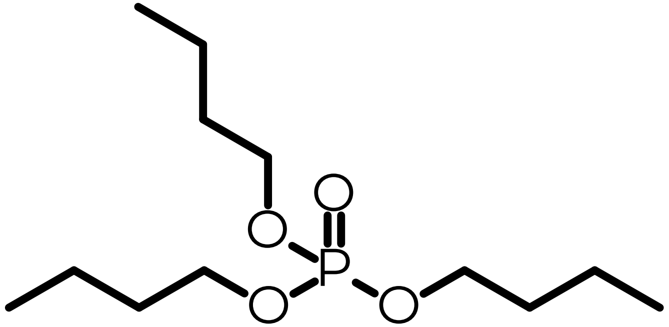
toxicity assessment

<1%



LC_{50} = 9.3 mg/L

<2%



LC_{50} = ? mg/L

~98%

?

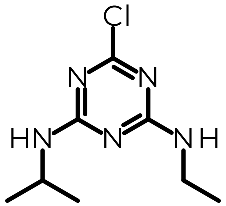
LC_{50} = ? mg/L

predicting toxicity

for detected chemicals

workflow

MS^2 spectra

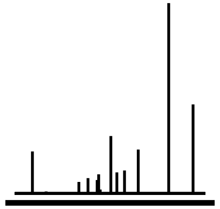


structure as SMILES

molecular descriptors

predict toxicity

workflow



MS^2 spectra

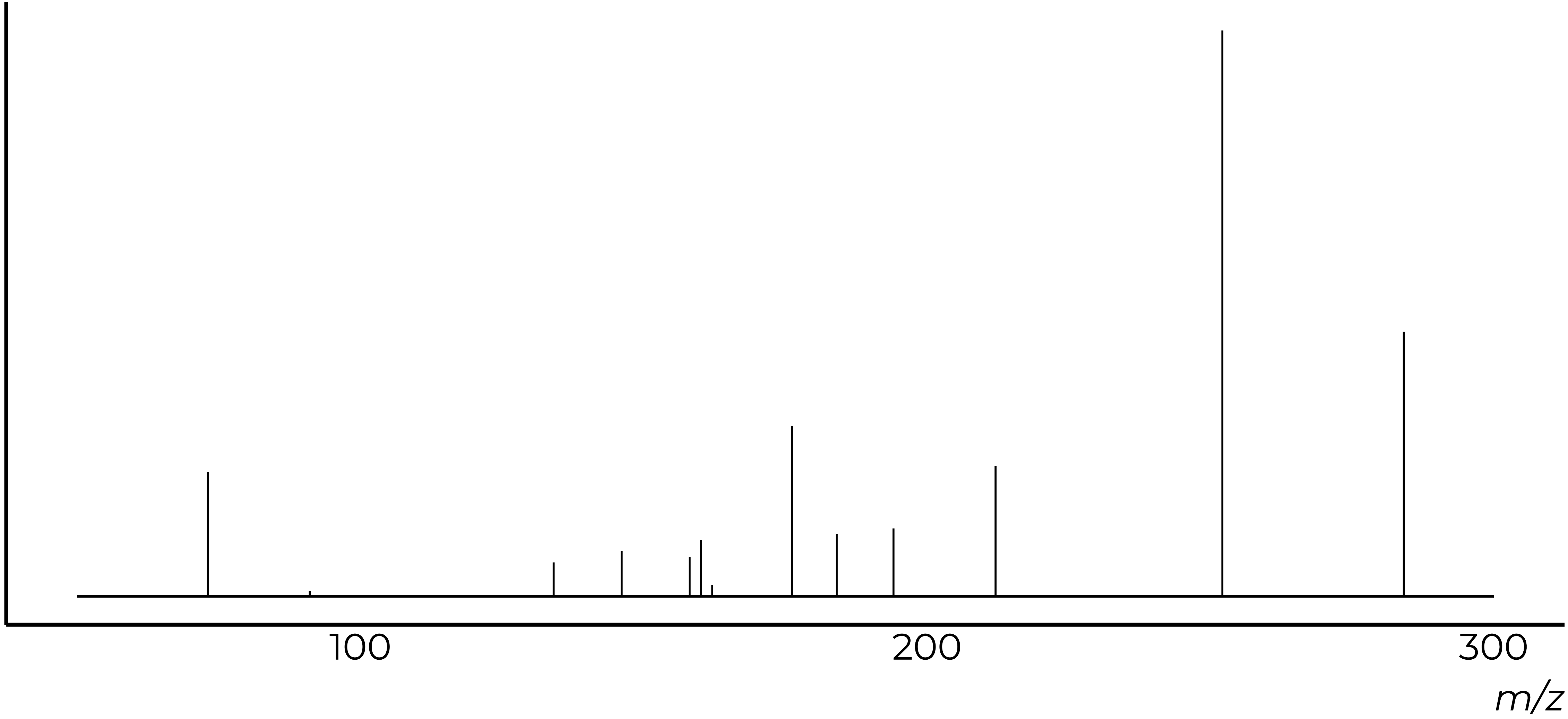
molecular descriptors

predict toxicity

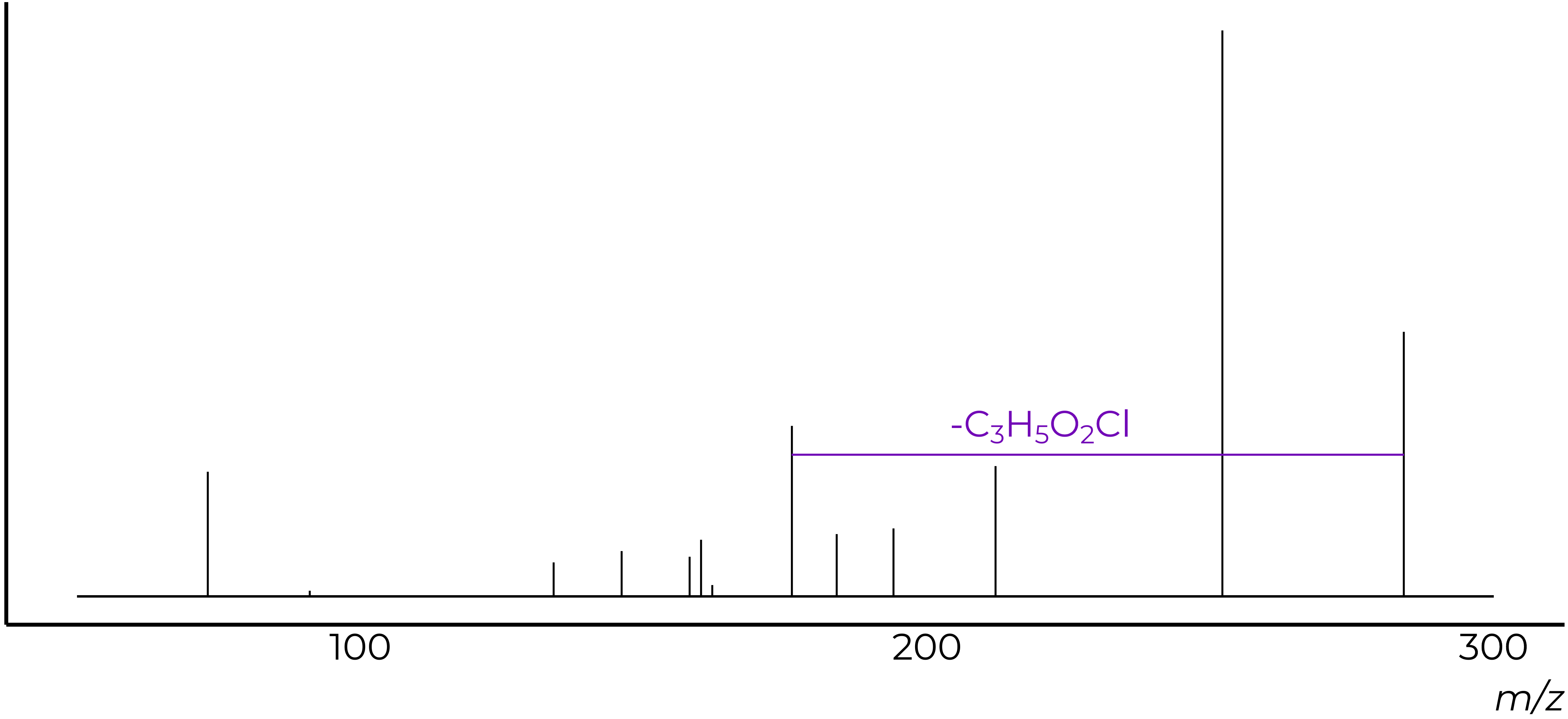
information available

in MS^2 spectra

MS^2 spectra

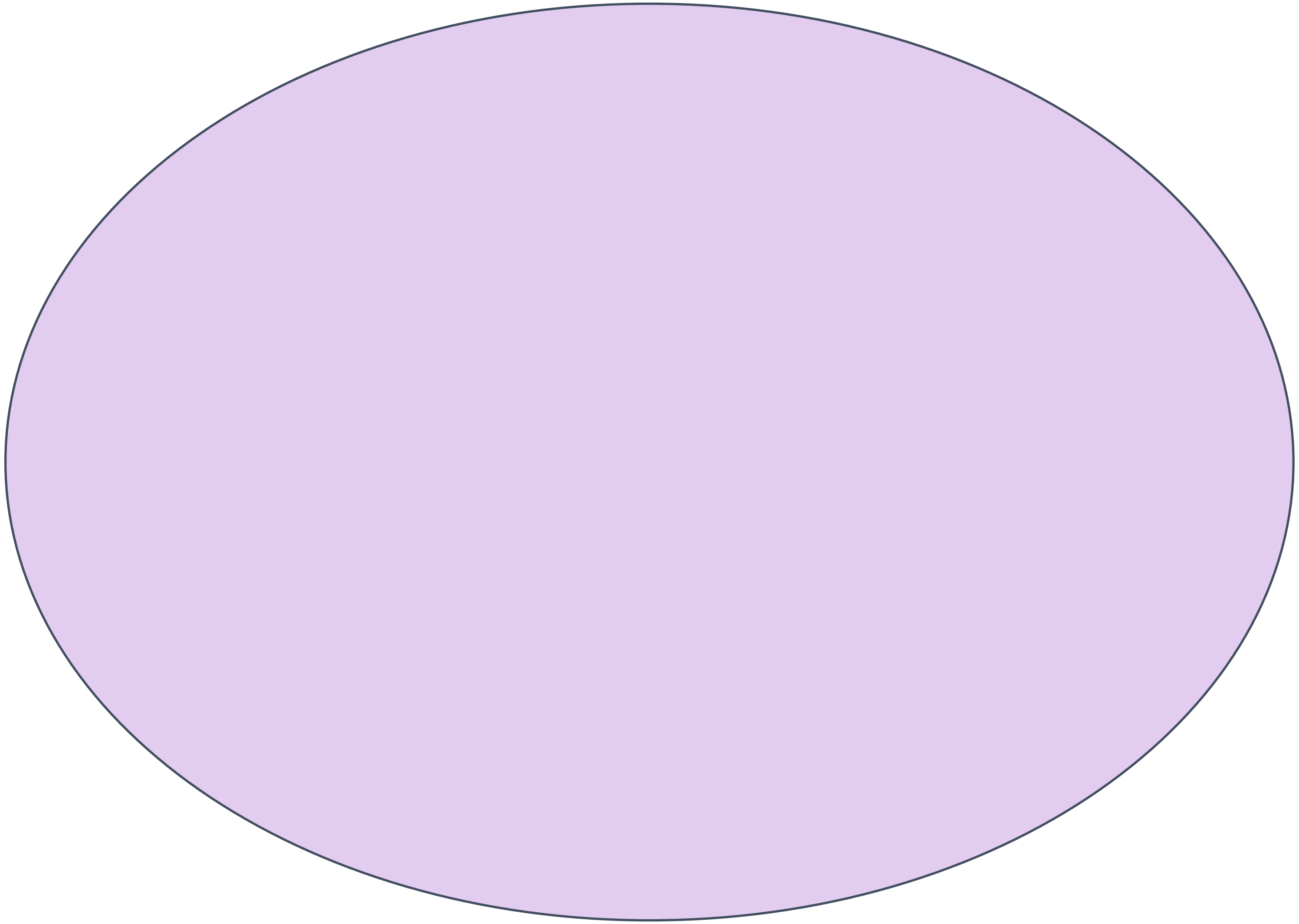


MS^2 spectra



data for machine learning models

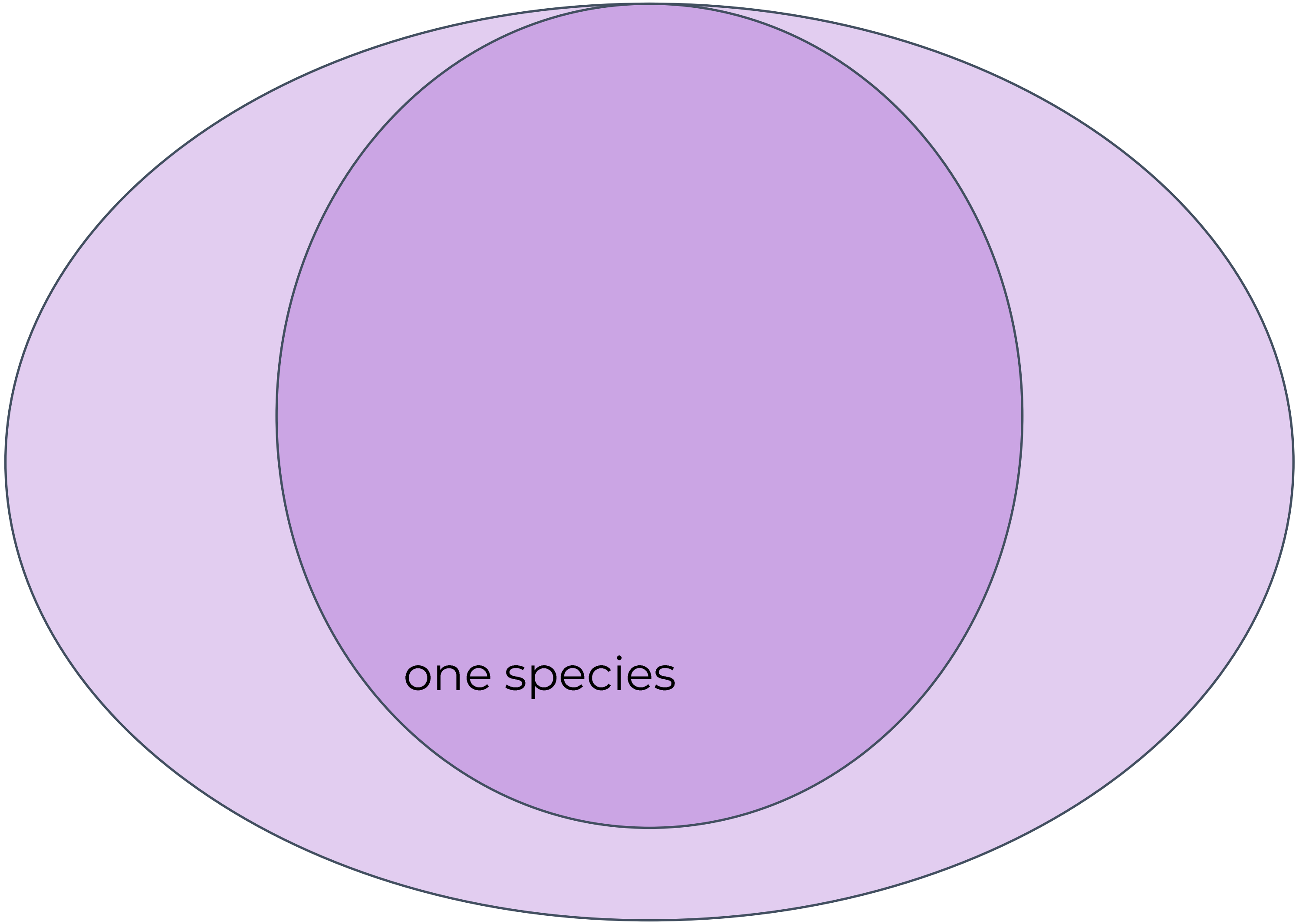
data for machine learning models



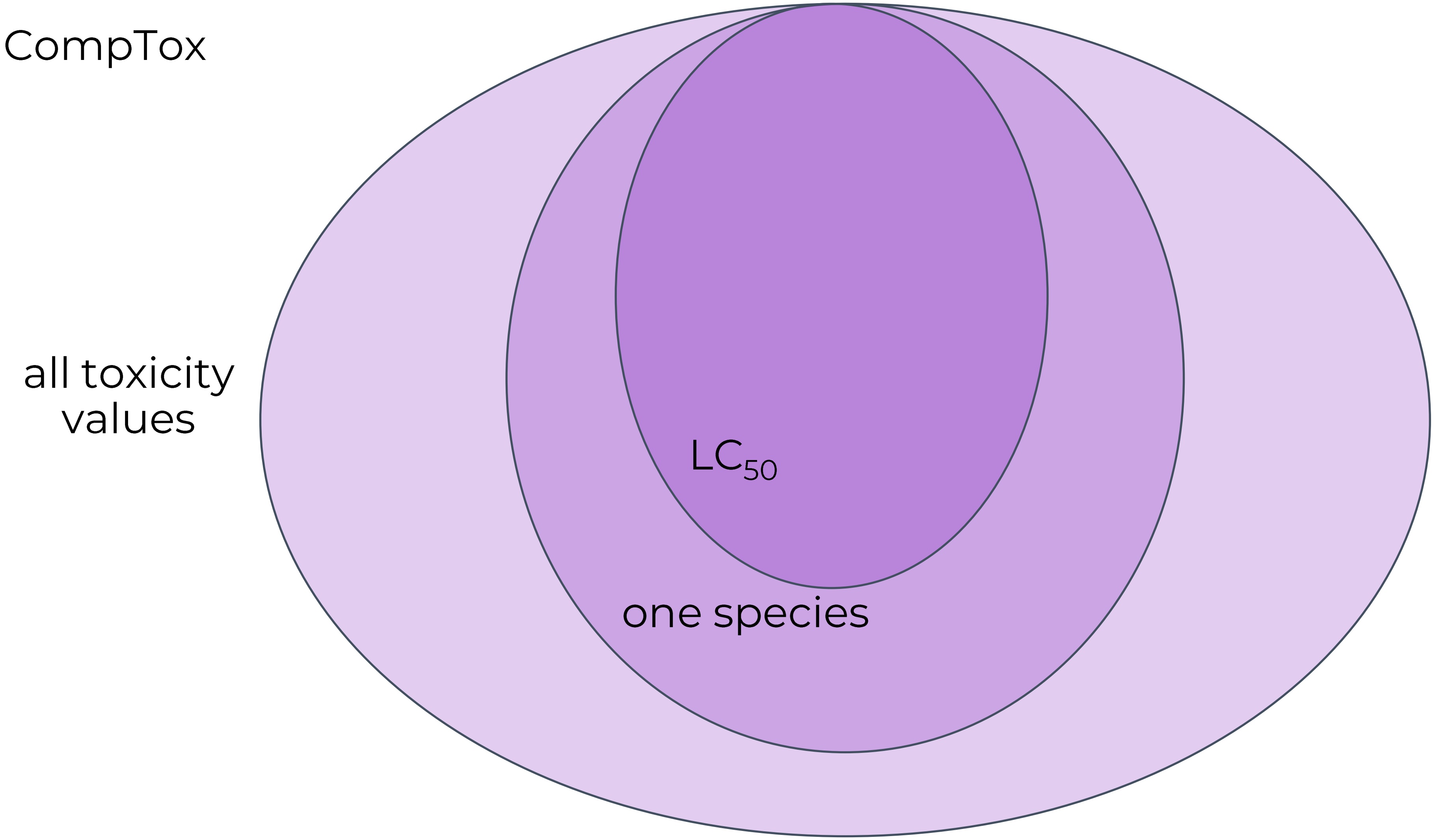
CompTox

all toxicity values

data for machine learning models



data for machine learning models



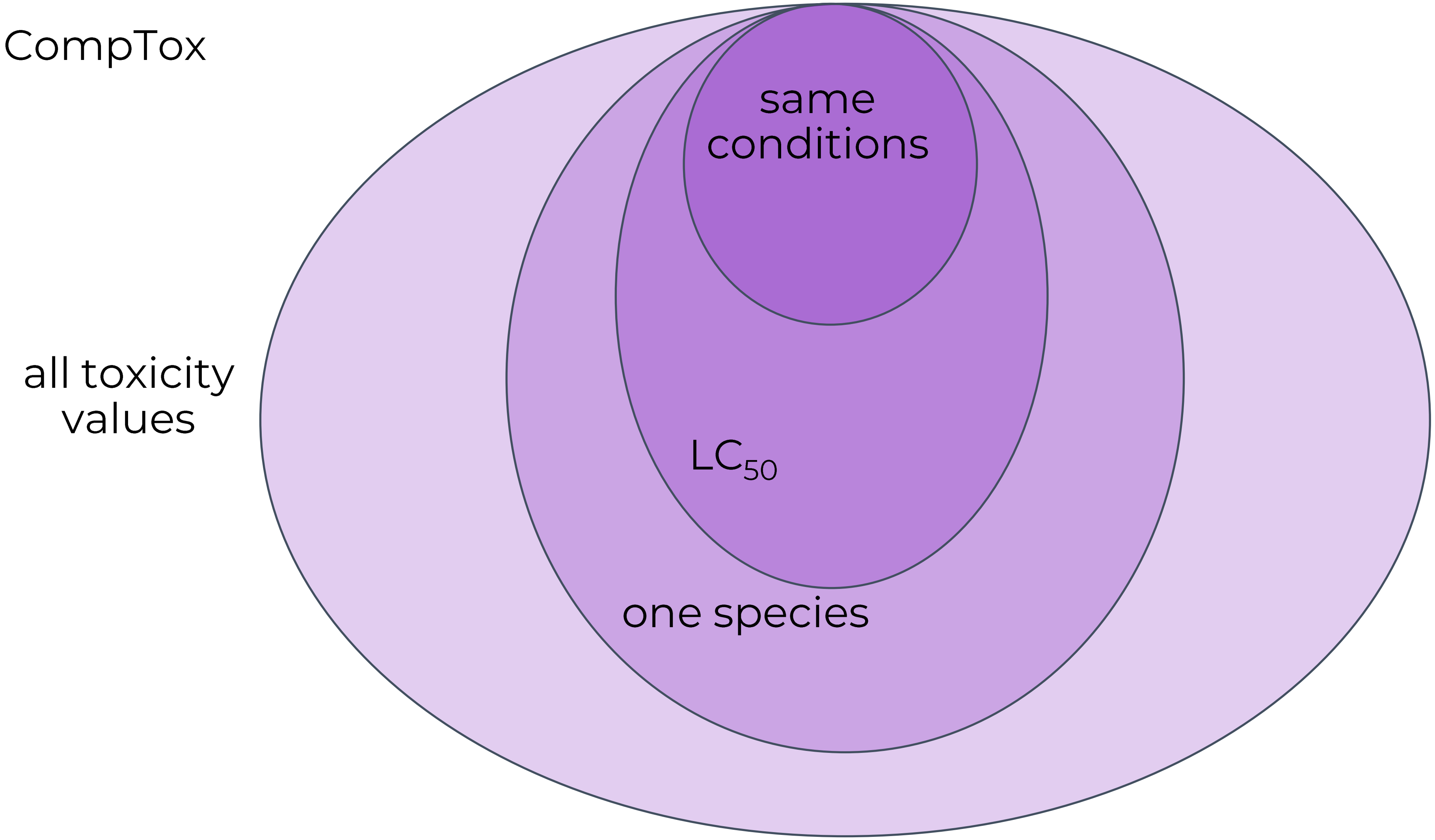
CompTox

all toxicity values

one species

LC_{50}

data for machine learning models



CompTox

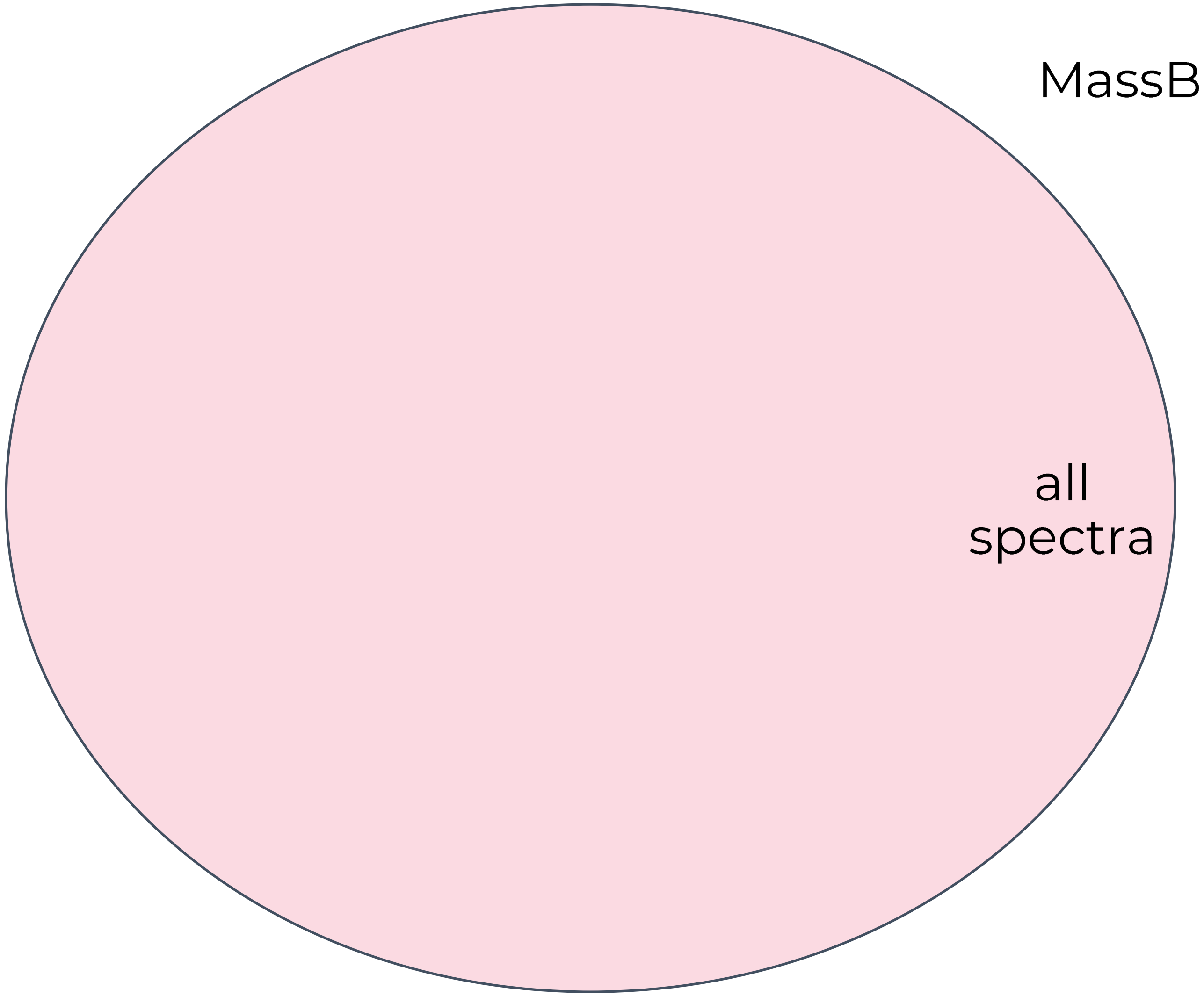
same conditions

all toxicity values

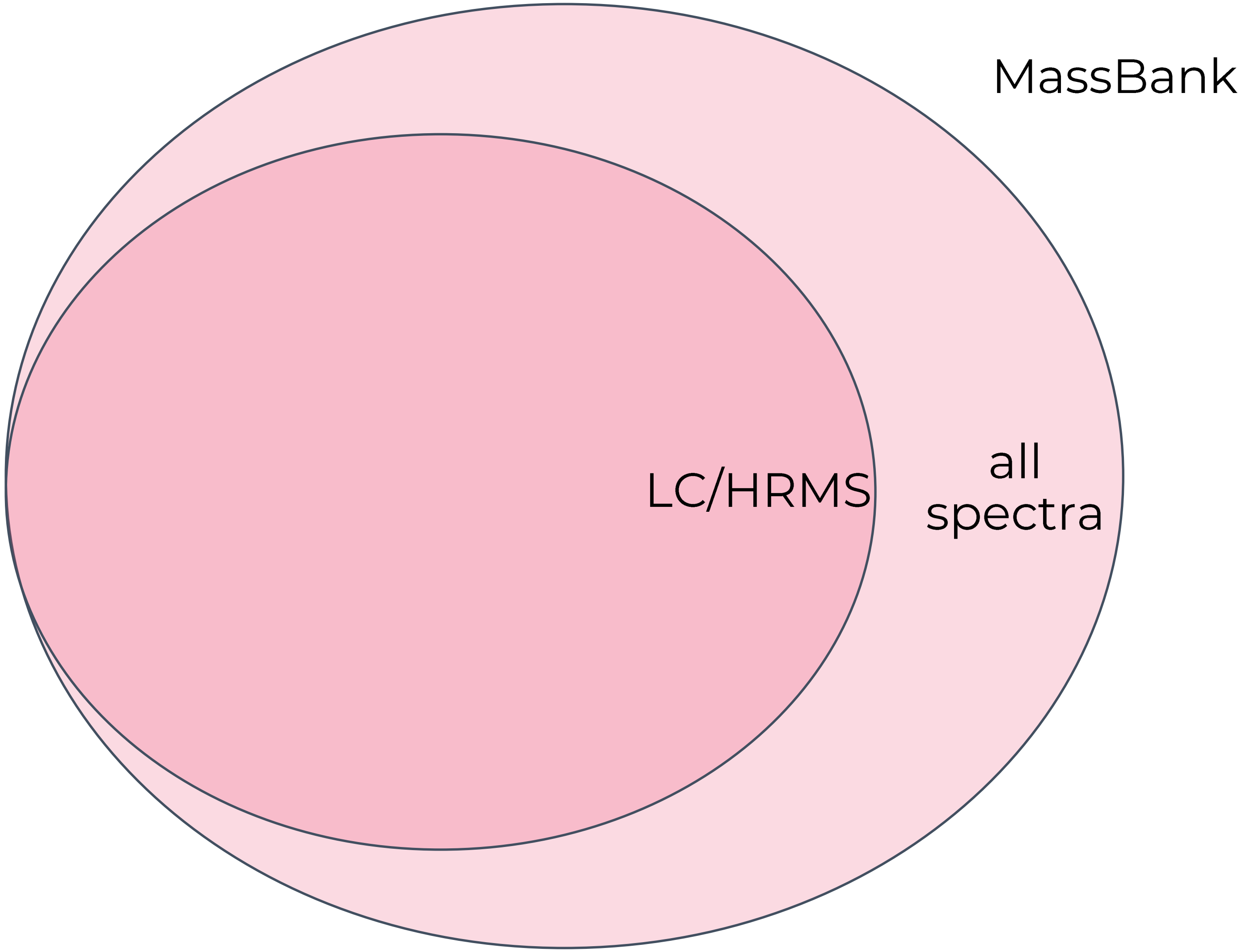
LC_{50}

one species

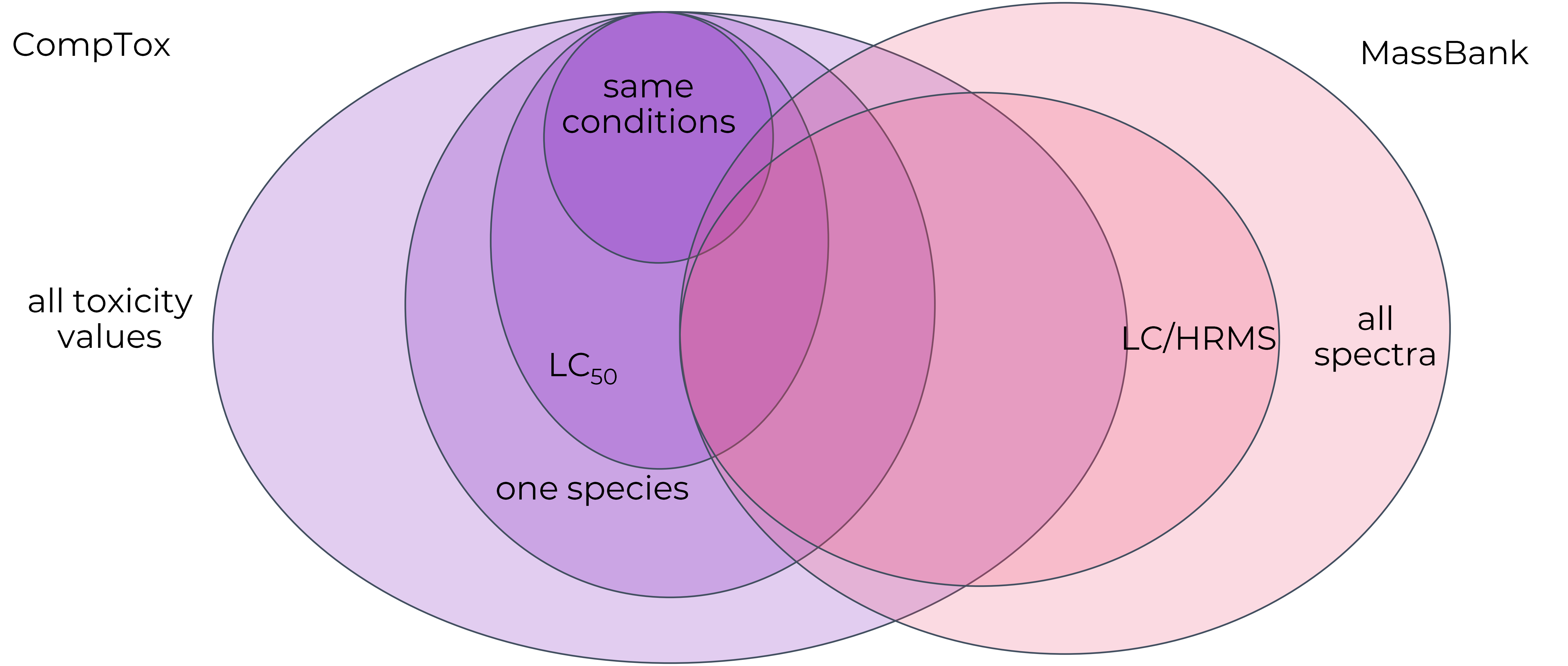
data for machine learning models



data for machine learning models



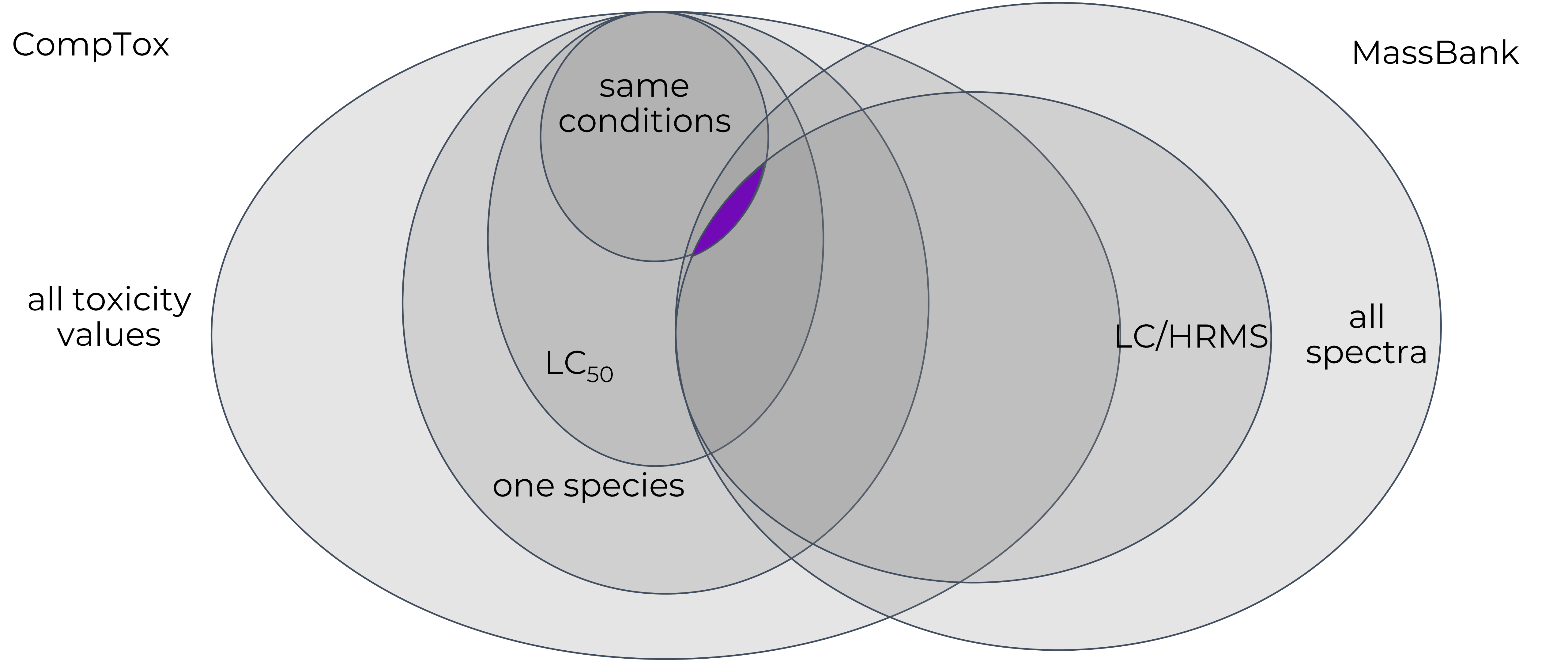
data for machine learning models



anneli kruve

anneli.kruve@su.se

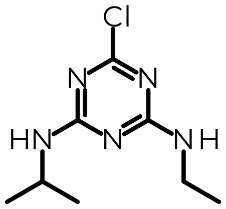
data for machine learning models



predicting toxicity

from the structure

workflow



structure as SMILES

molecular fingerprints

machine learning for predicting toxicity

selected endpoint

selected endpoint

🐟 fathead minnow, bluegill, and rainbow trout

selected endpoint

fathead minnow, bluegill, and rainbow trout

water flea

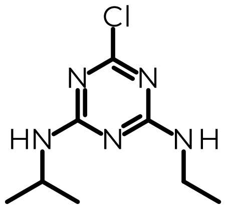
selected endpoint

fathead minnow, bluegill, and rainbow trout

water flea

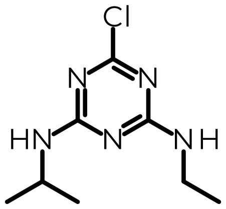
algae

workflow



structure as SMILES

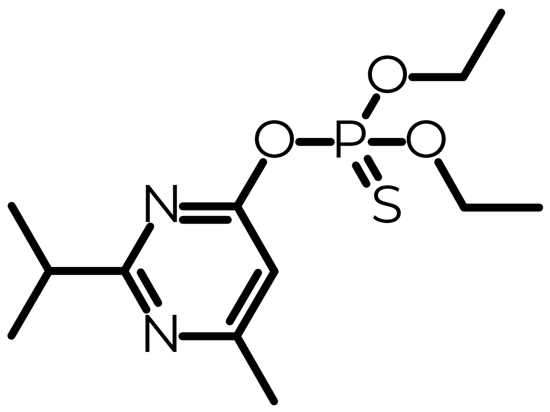
workflow



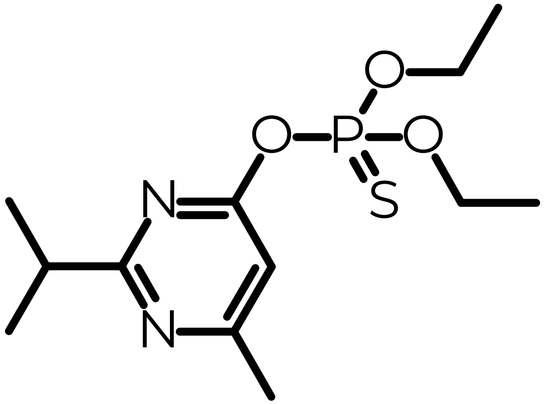
structure as SMILES

molecular fingerprints

structural fingerprints



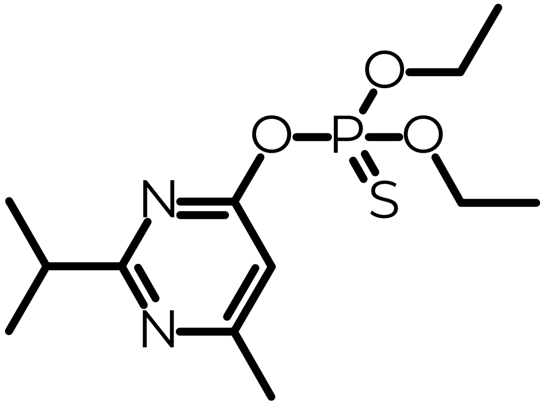
structural fingerprints



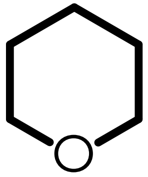
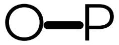
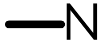
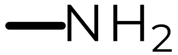
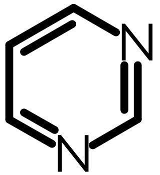
R: rcdk

→

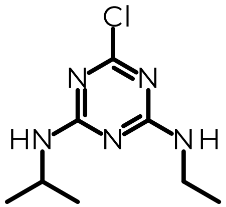
structural fingerprints



R: rcdk

0	
1	
1	
0	
1	

workflow



structure as SMILES

molecular fingerprints

machine learning for predicting LC_{50}

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

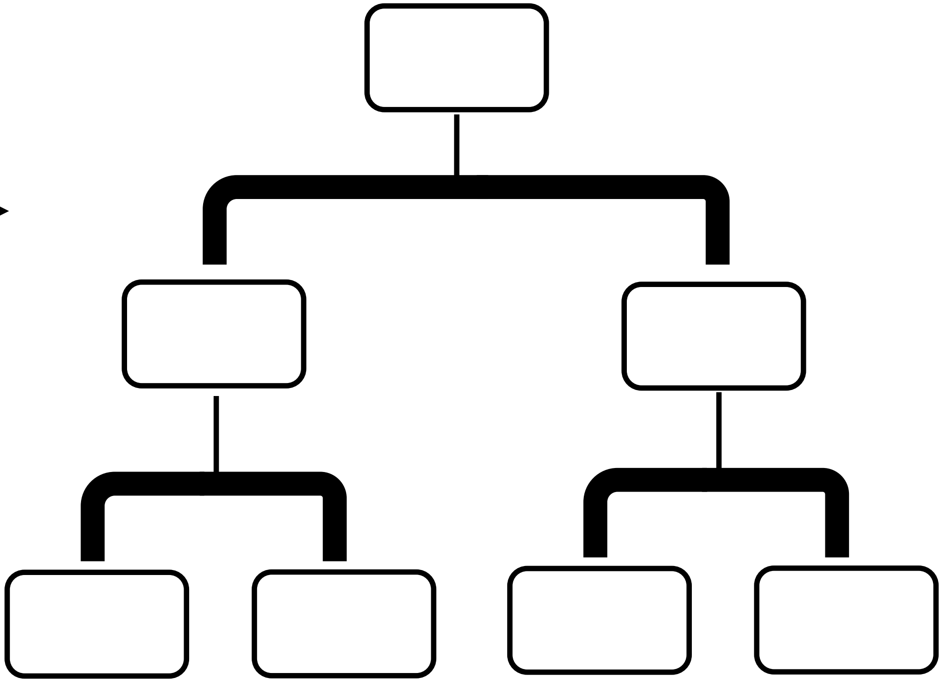
model training

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

$\xrightarrow[\text{boosting}]{\text{gradient}}$



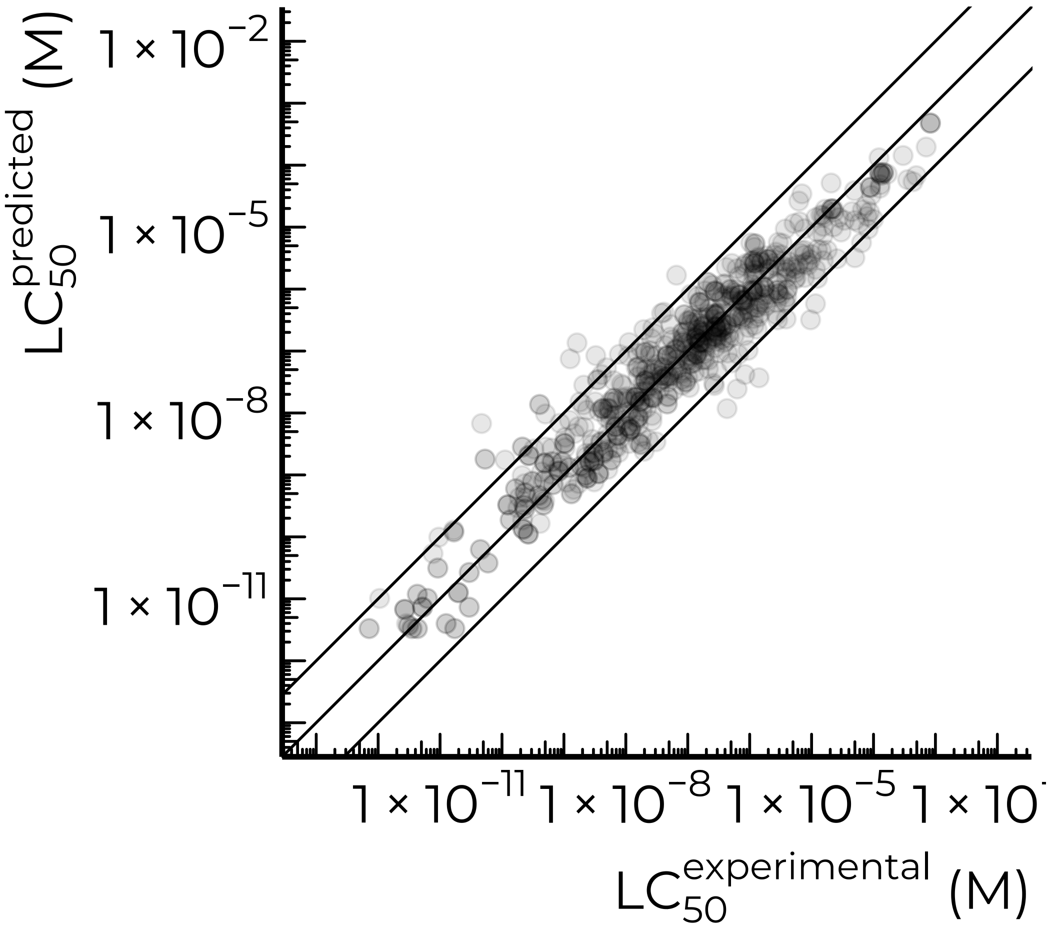
performance

of LC_{50} predictions with molecular fingerprints

LC_{50} predictions

Peets et al. ES&T 2022

fish LC_{50}



training set

RMSE 0.52 log(M)

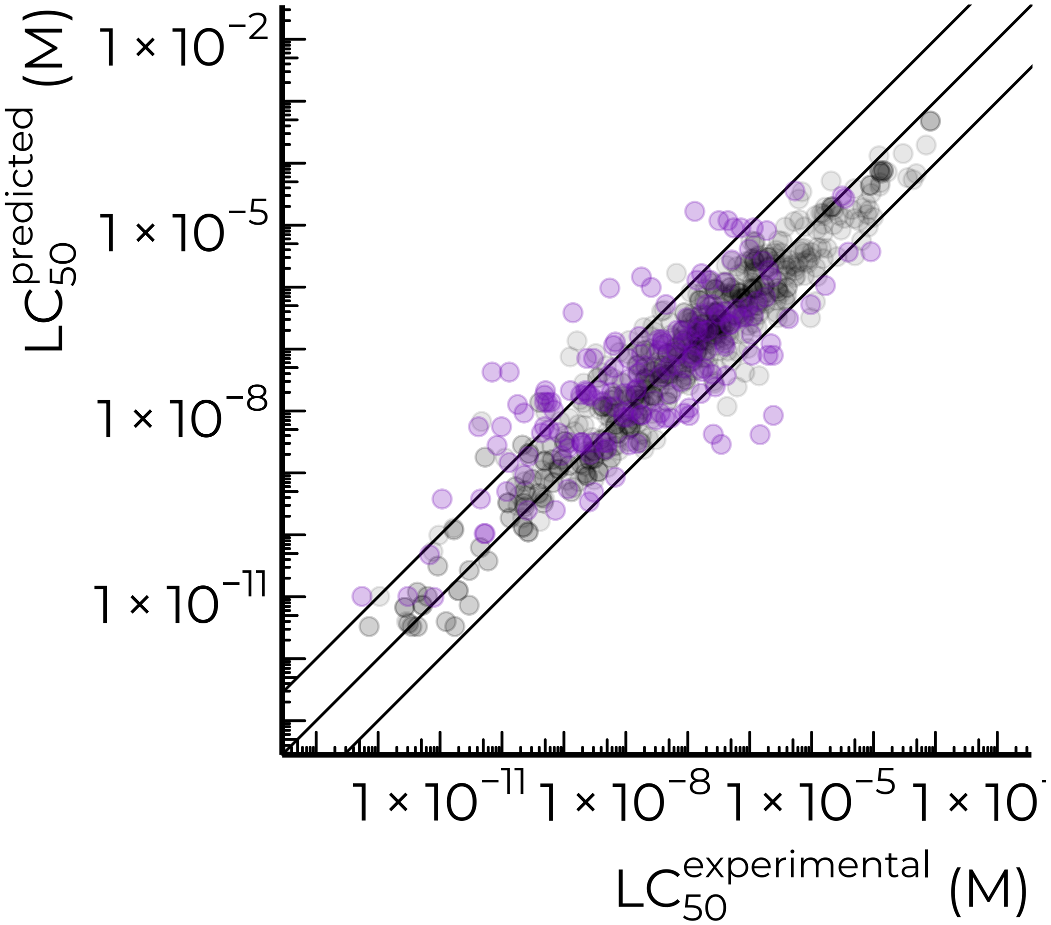
anneli kruve

anneli.kruve@su.se

LC_{50} predictions

Peets et al. ES&T 2022

fish LC_{50}



training set

RMSE 0.52 log(M)

test set

RMSE 0.78 log(M)

unidentified chemicals

from MS^2 spectra

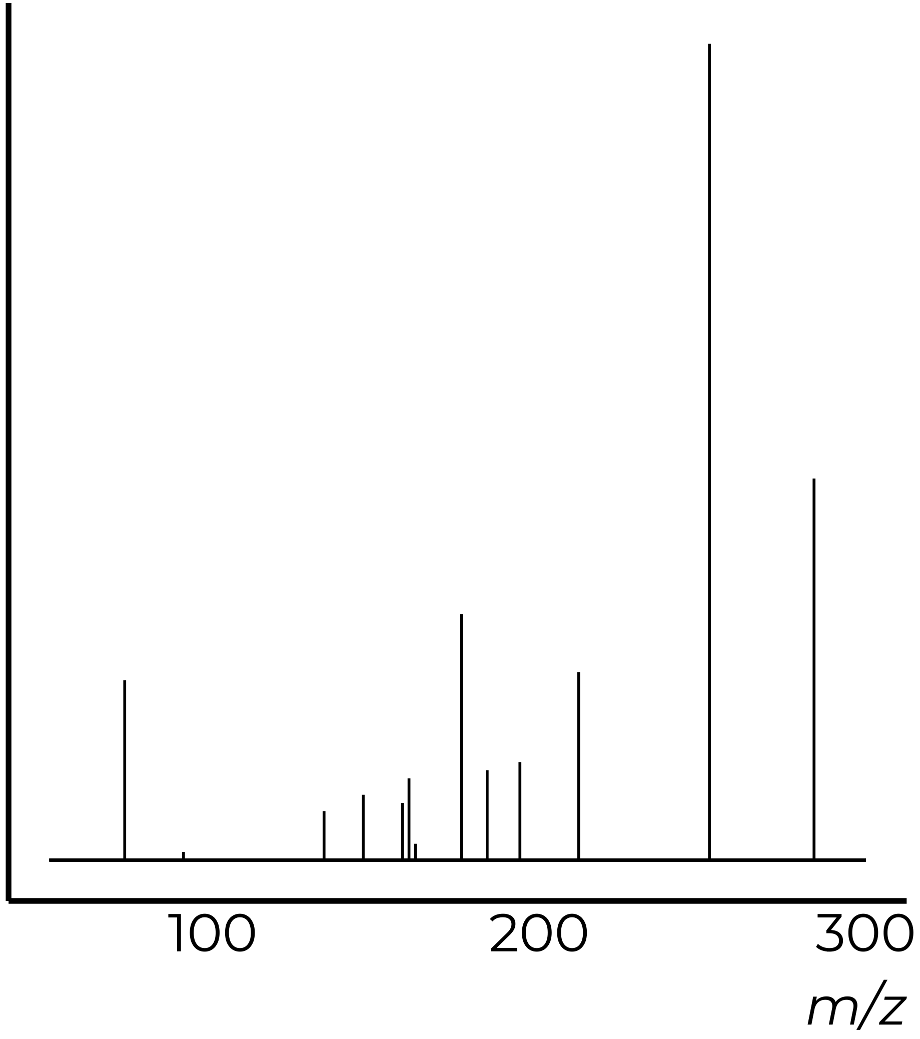
workflow

MS^2 spectra

molecular fingerprints with SIRIUS+CSI:FingerID

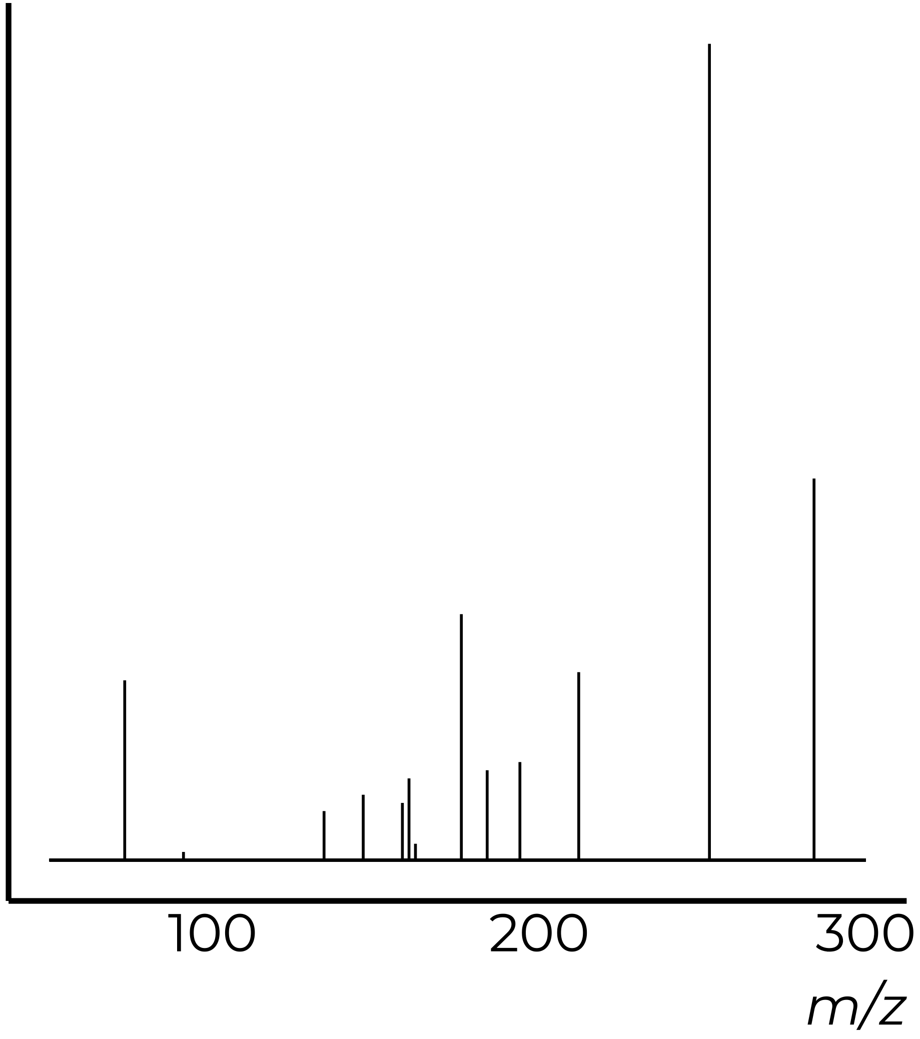
predict LC_{50} with pretrained gradient boosting

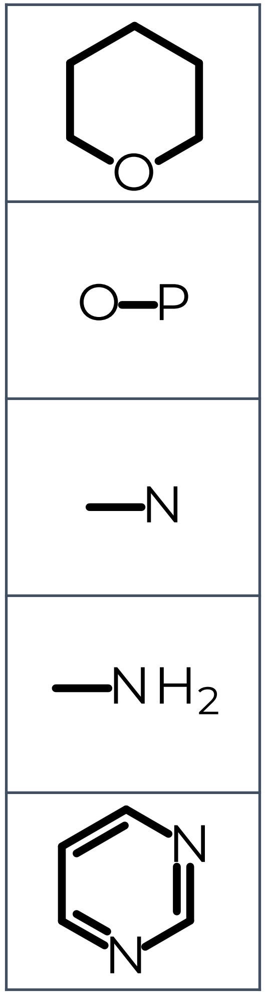
predict for unknown chemicals



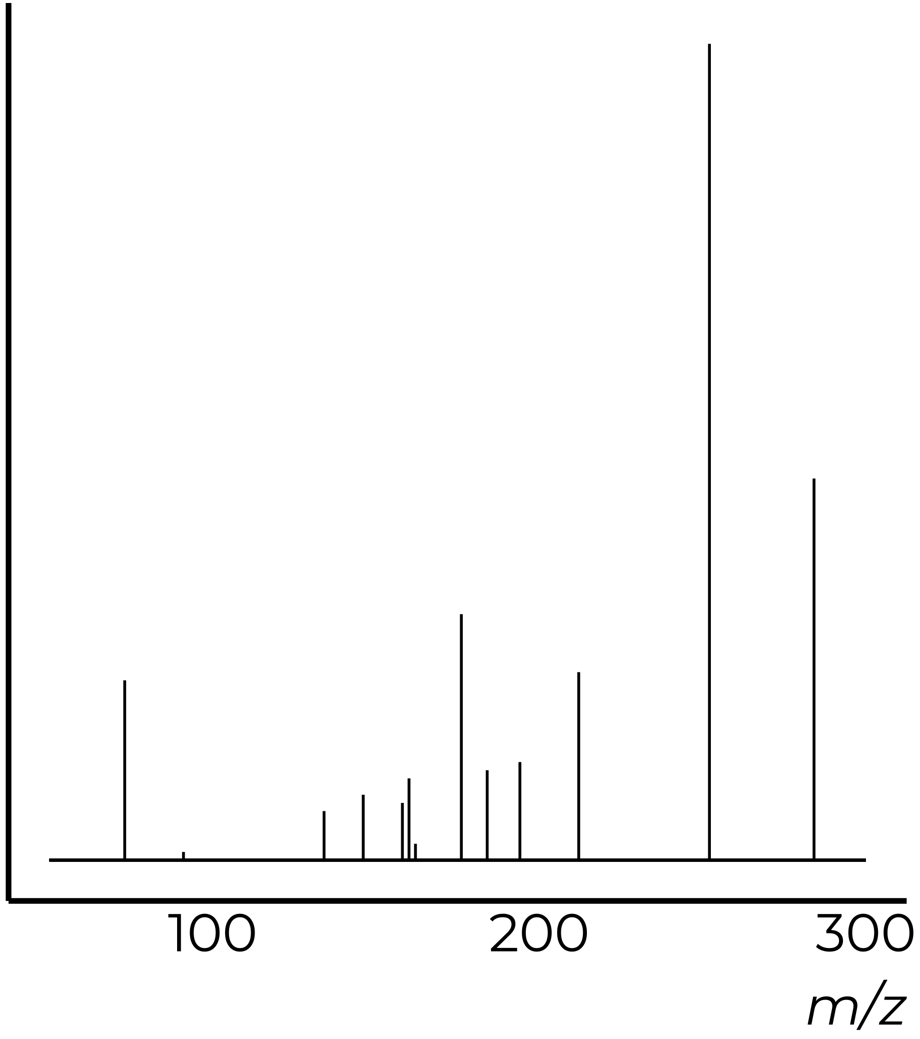
→ ?

predict for unknown chemicals

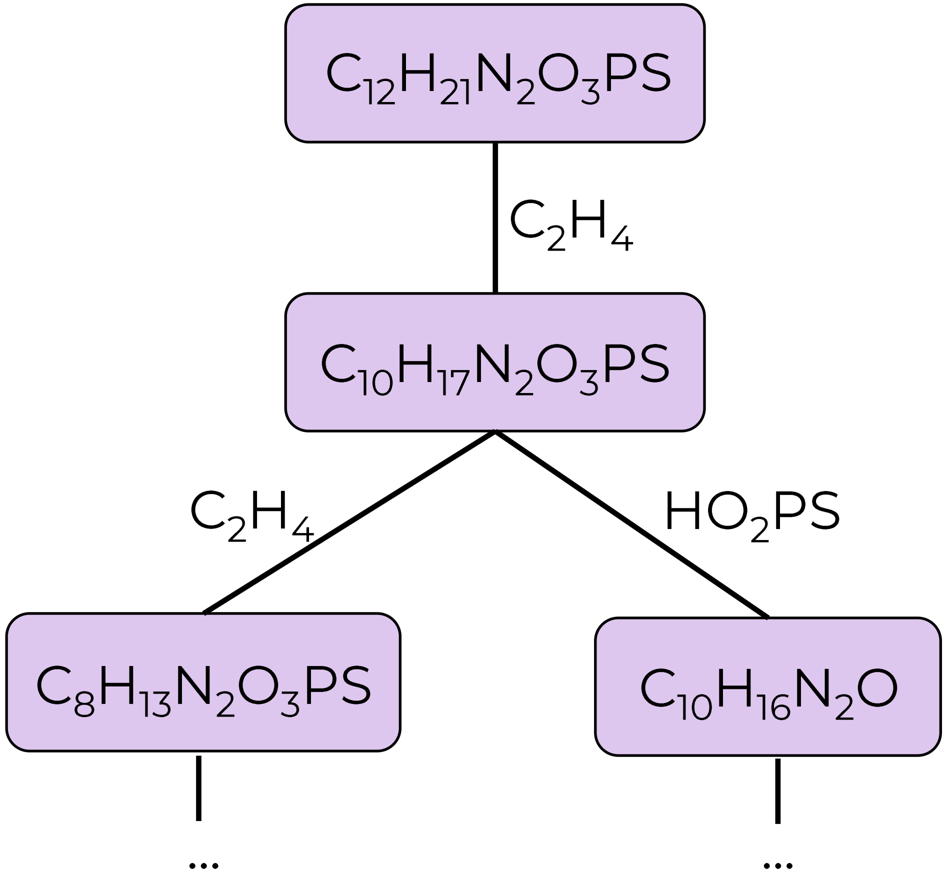




predict for unknown chemicals



SIRIUS+
CSI:FingerID →



$C_{12}H_{21}N_2O_3PS$

C_2H_4

$C_{10}H_{17}N_2O_3PS$

C_2H_4

HO_2PS

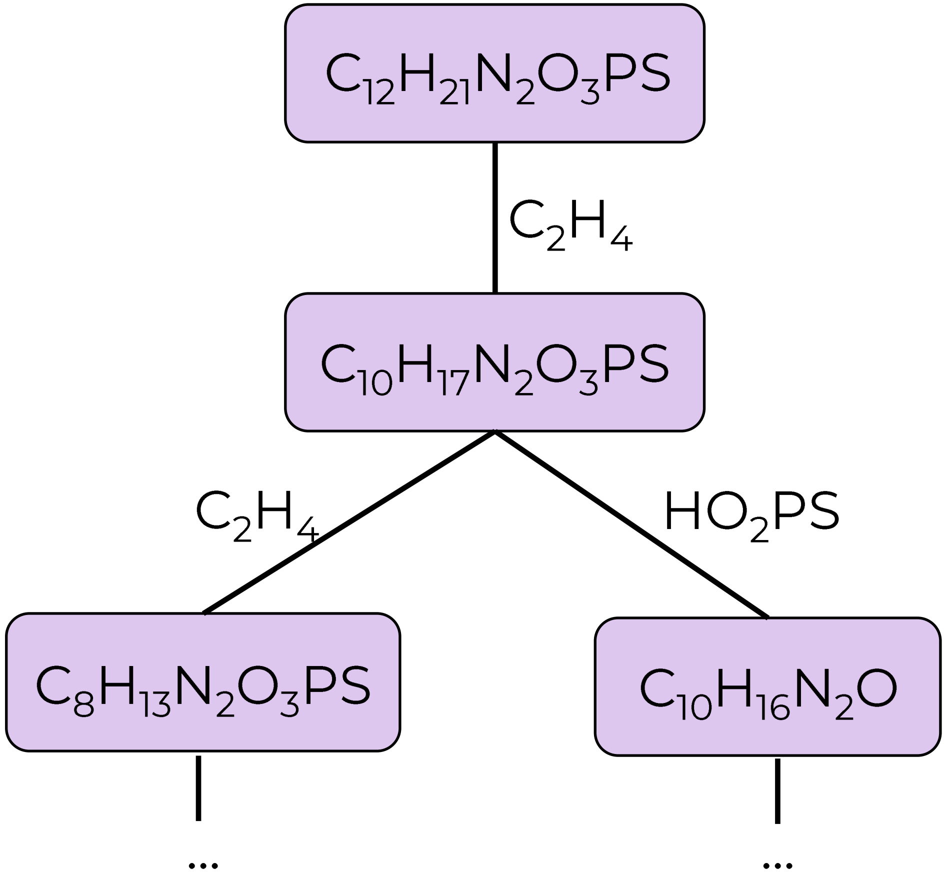
$C_8H_{13}N_2O_3PS$

$C_{10}H_{16}N_2O$

…

…

predict for unknown chemicals


$C_{12}H_{21}N_2O_3PS$

C_2H_4

$C_{10}H_{17}N_2O_3PS$

C_2H_4

HO_2PS

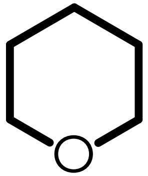
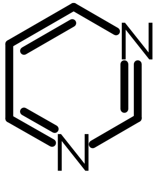
$C_8H_{13}N_2O_3PS$

$C_{10}H_{16}N_2O$

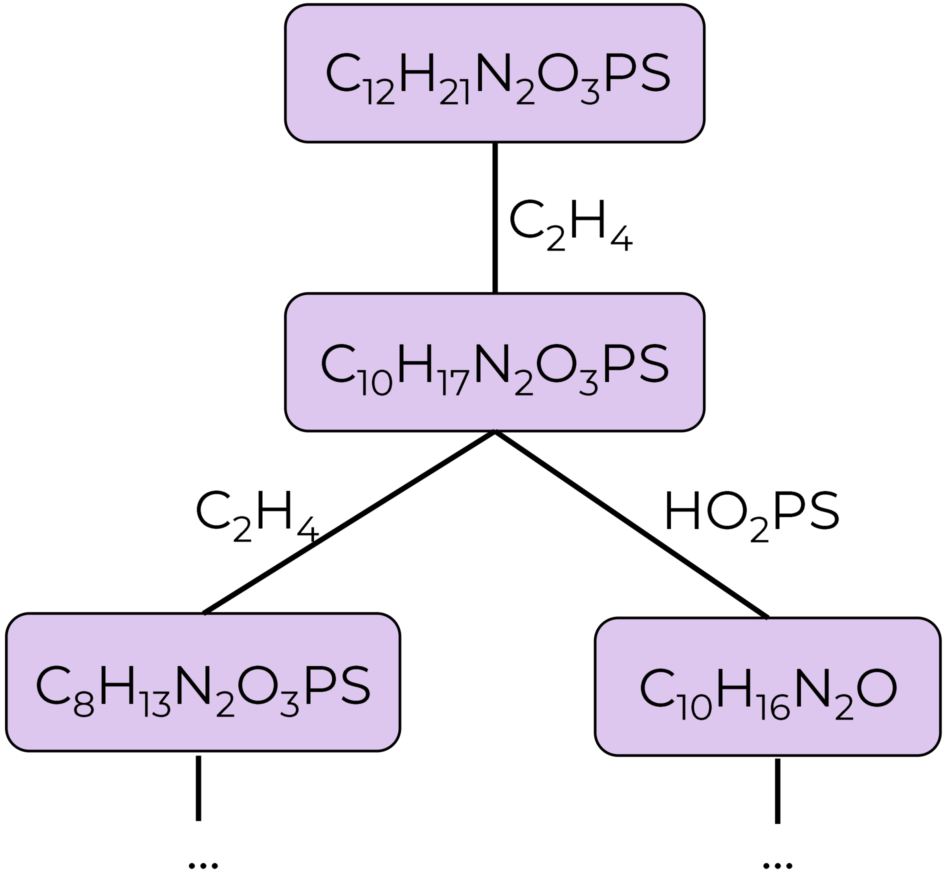
...

...

$\xrightarrow{\text{SIRIUS+ CSI:FingerID}}$

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

predict for unknown chemicals



$C_{12}H_{21}N_2O_3PS$

C_2H_4

$C_{10}H_{17}N_2O_3PS$

C_2H_4

HO_2PS

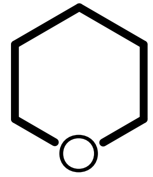
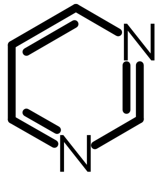
$C_8H_{13}N_2O_3PS$

$C_{10}H_{16}N_2O$

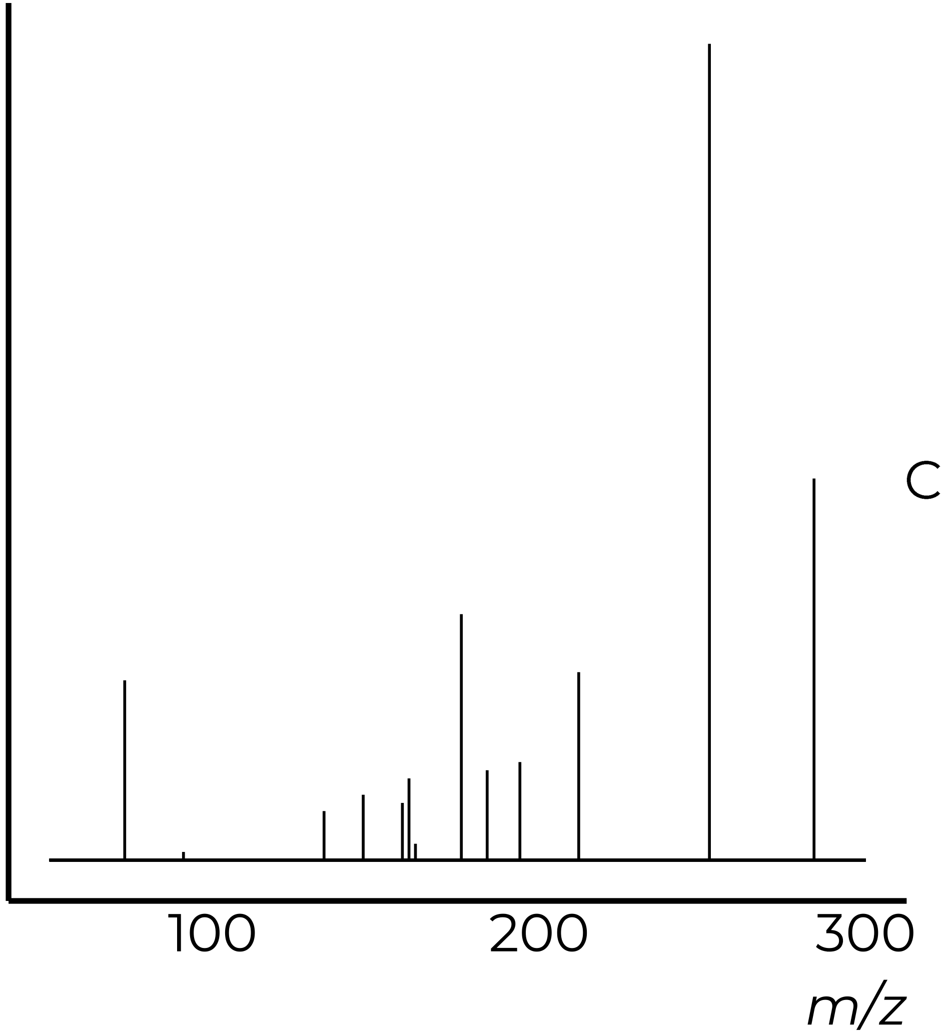
...

...

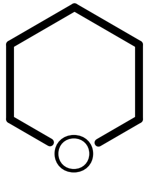
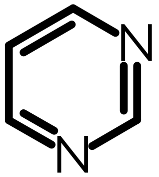
$\xrightarrow{\text{SIRIUS+ CSI:FingerID}}$

0	
1	O—P
1	—N
0	—NH_2
1	

predict for unknown chemicals



SIRIUS+
CSI:FingerID →

0	
1	O—P
1	—N
0	—NH$_2$
1	

gradient
boosting →

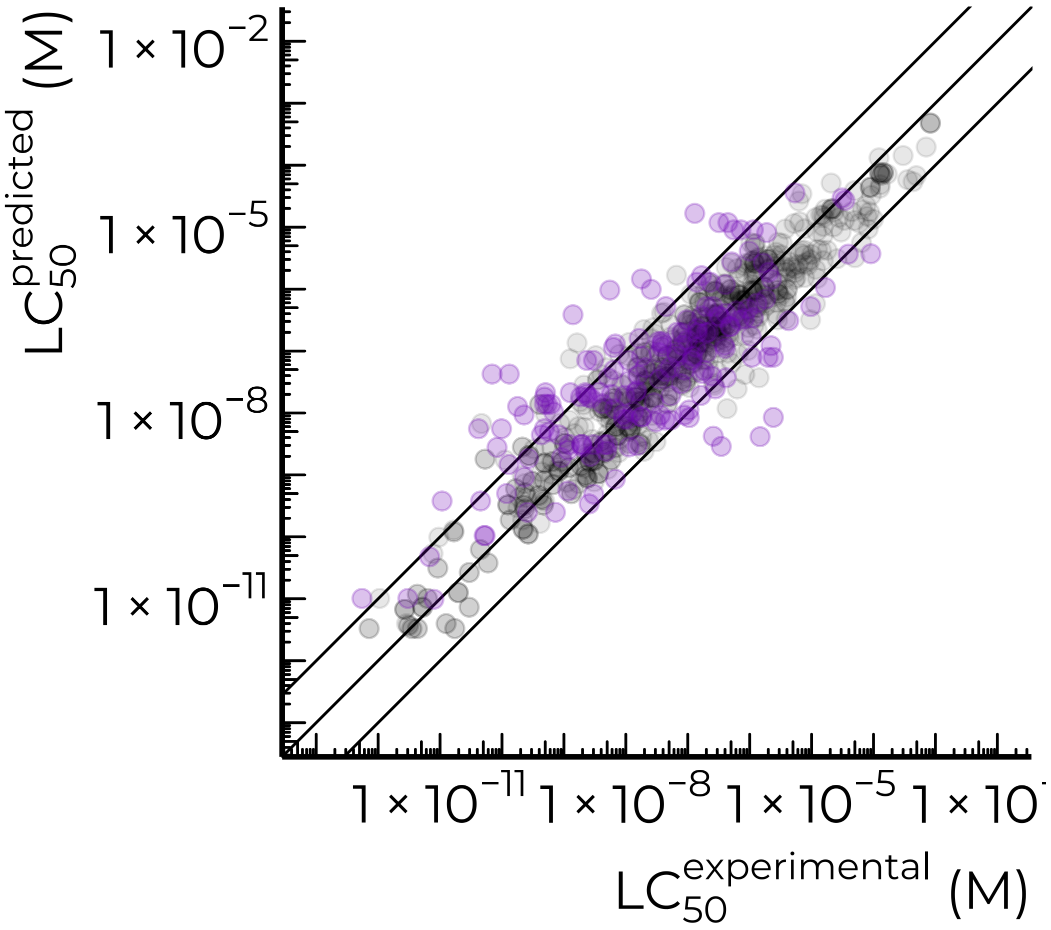
LC_{50} =
-2.2 log(mM)

LC_{50} predictions

LC_{50} predictions

Peets et al. ES&T 2022

fish LC_{50}



test set on structures

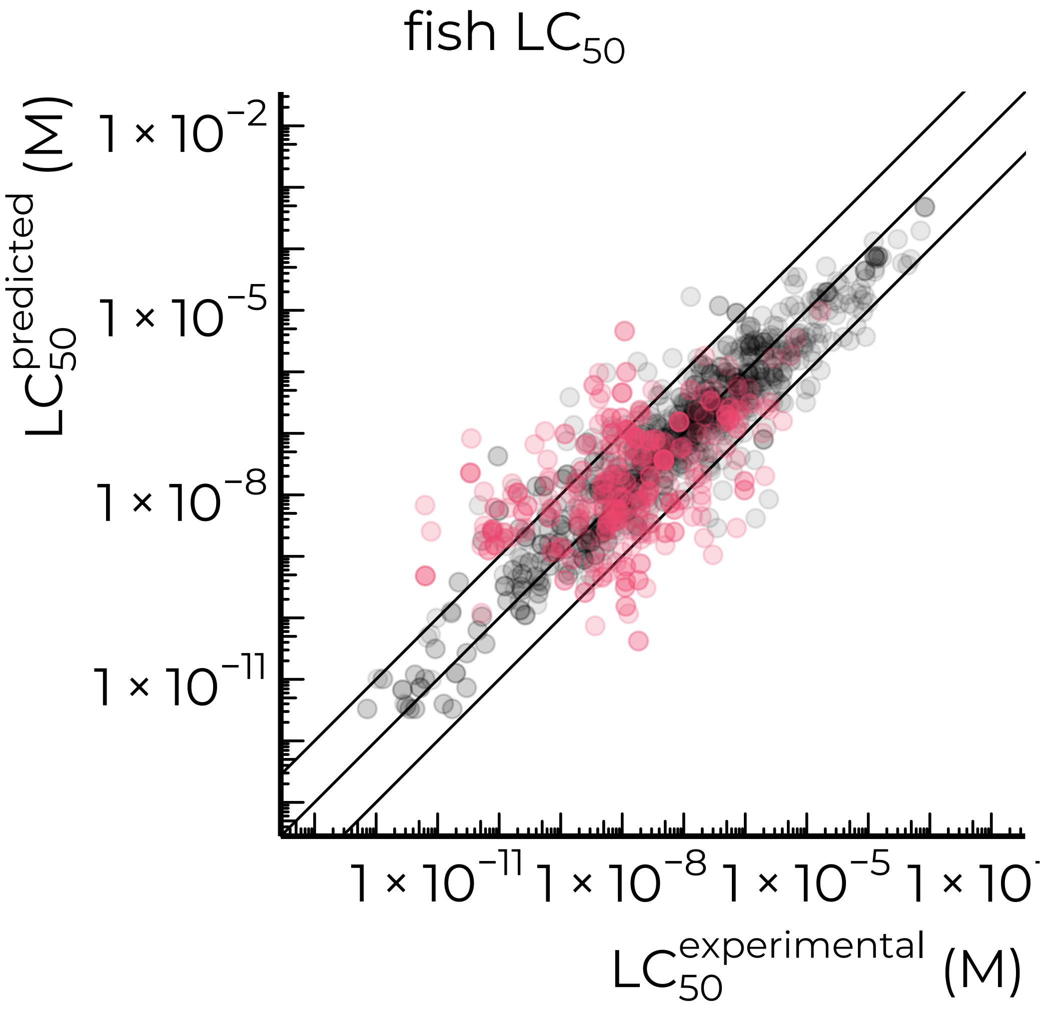
RMSE 0.78 log(M)

anneli kruve

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

$\text{RMSE}_{\text{model}}$ 0.88 log(M)

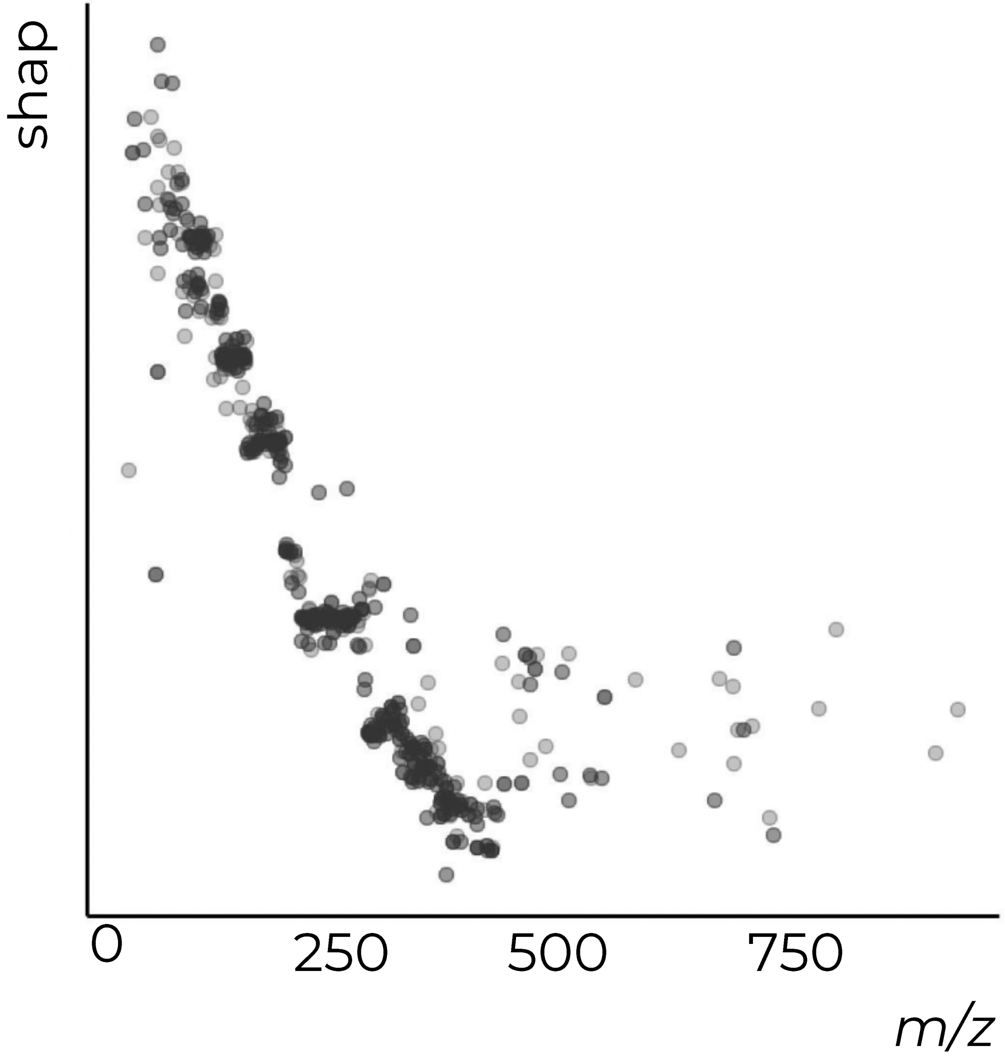
$\text{SD}_{\text{experimental}}$ 0.44 log(mM)

anneli kruve

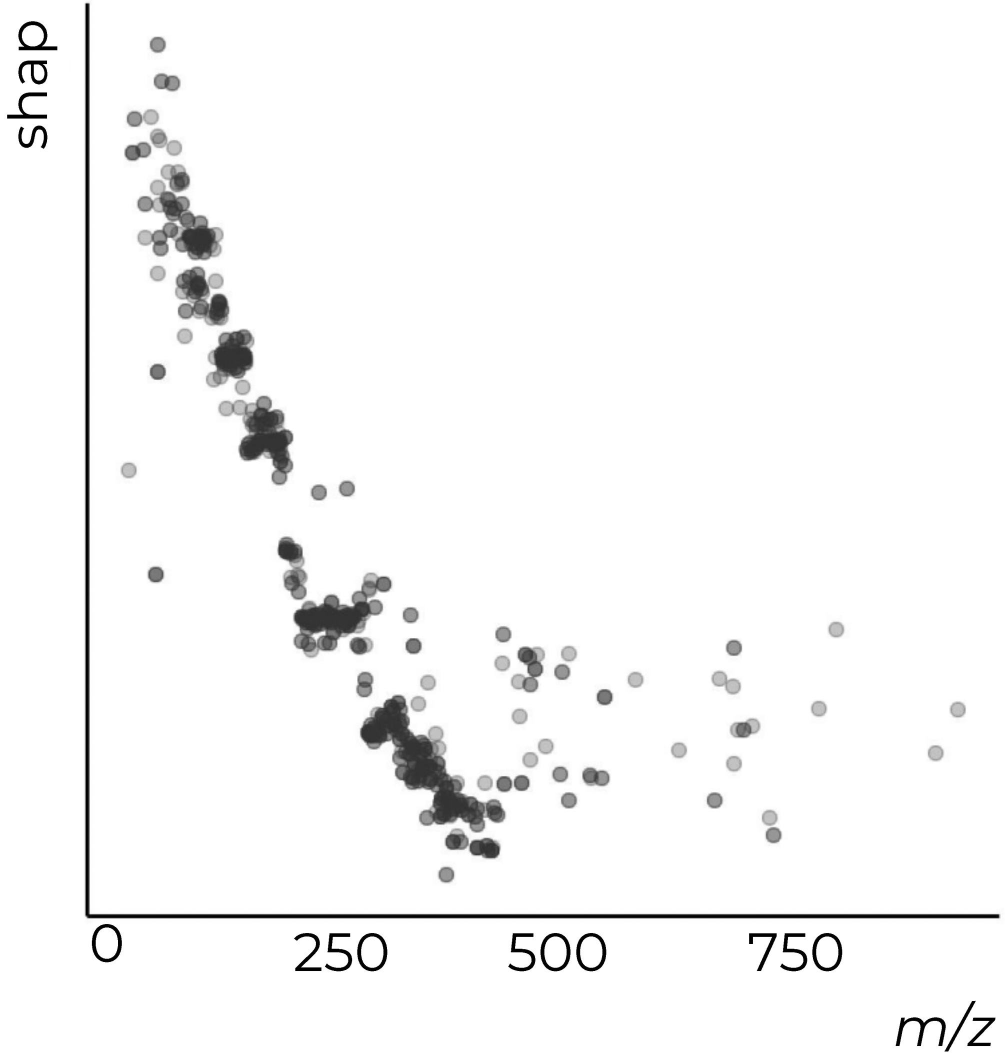
anneli.kruve@su.se

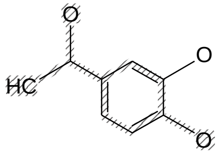
model interpretation

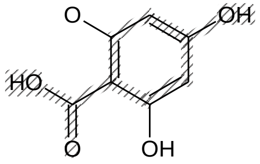
model interpretation

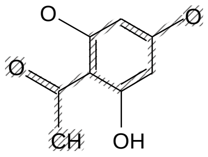


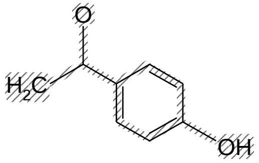
model interpretation







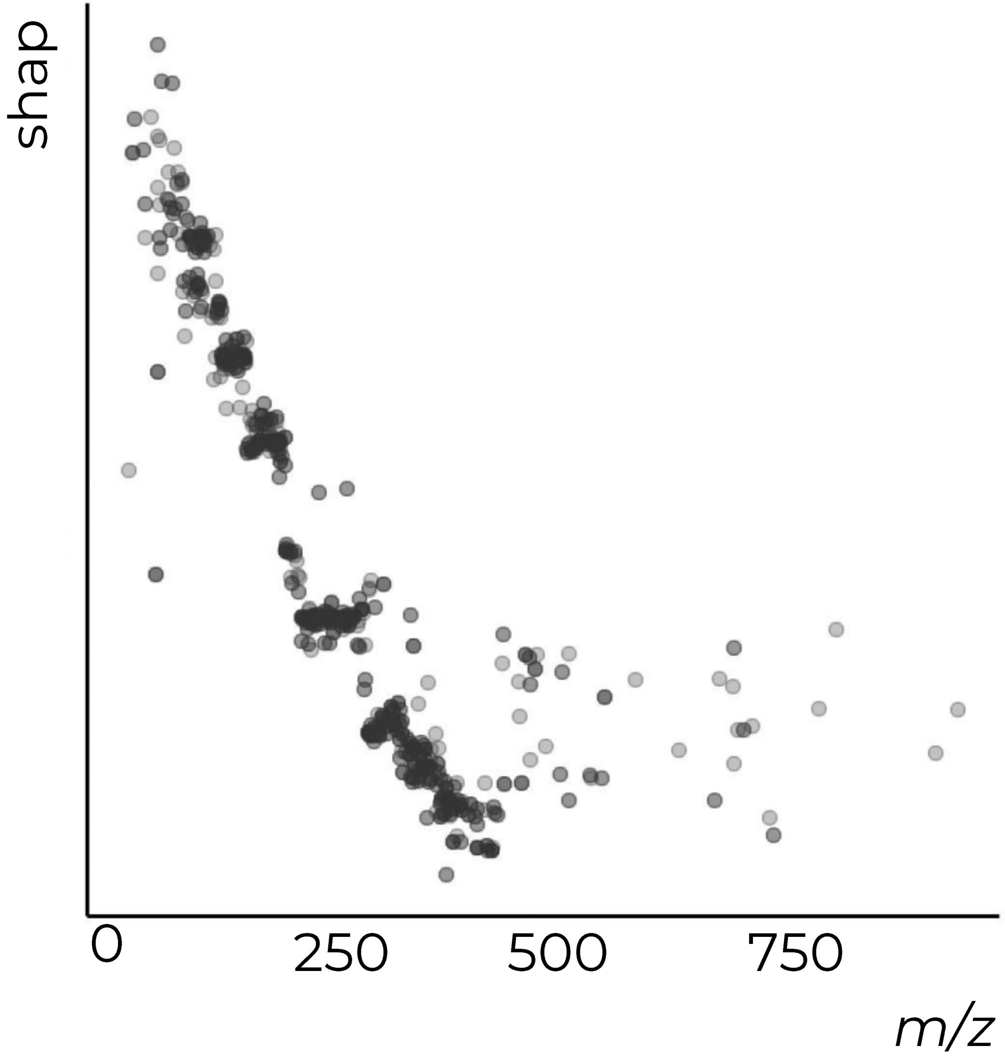


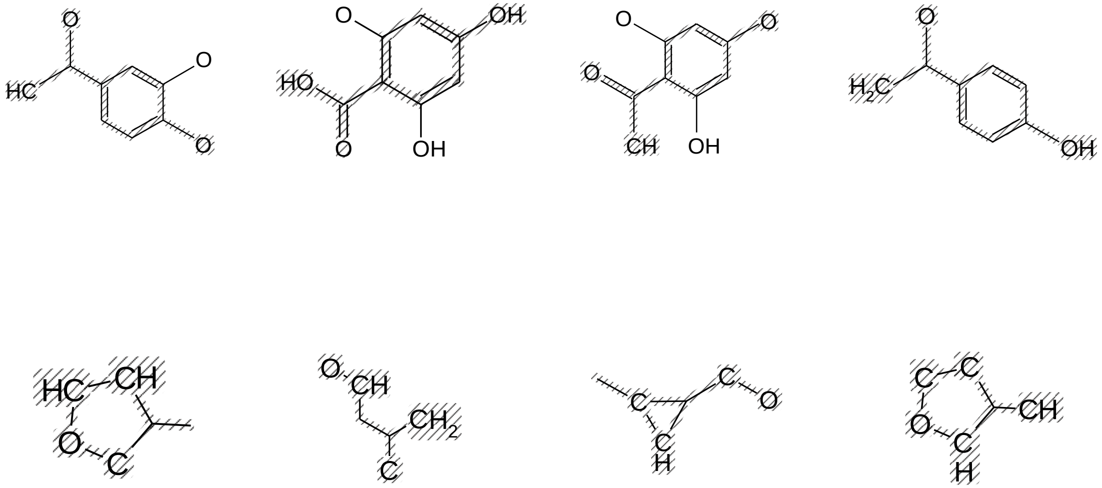


anneli kruve

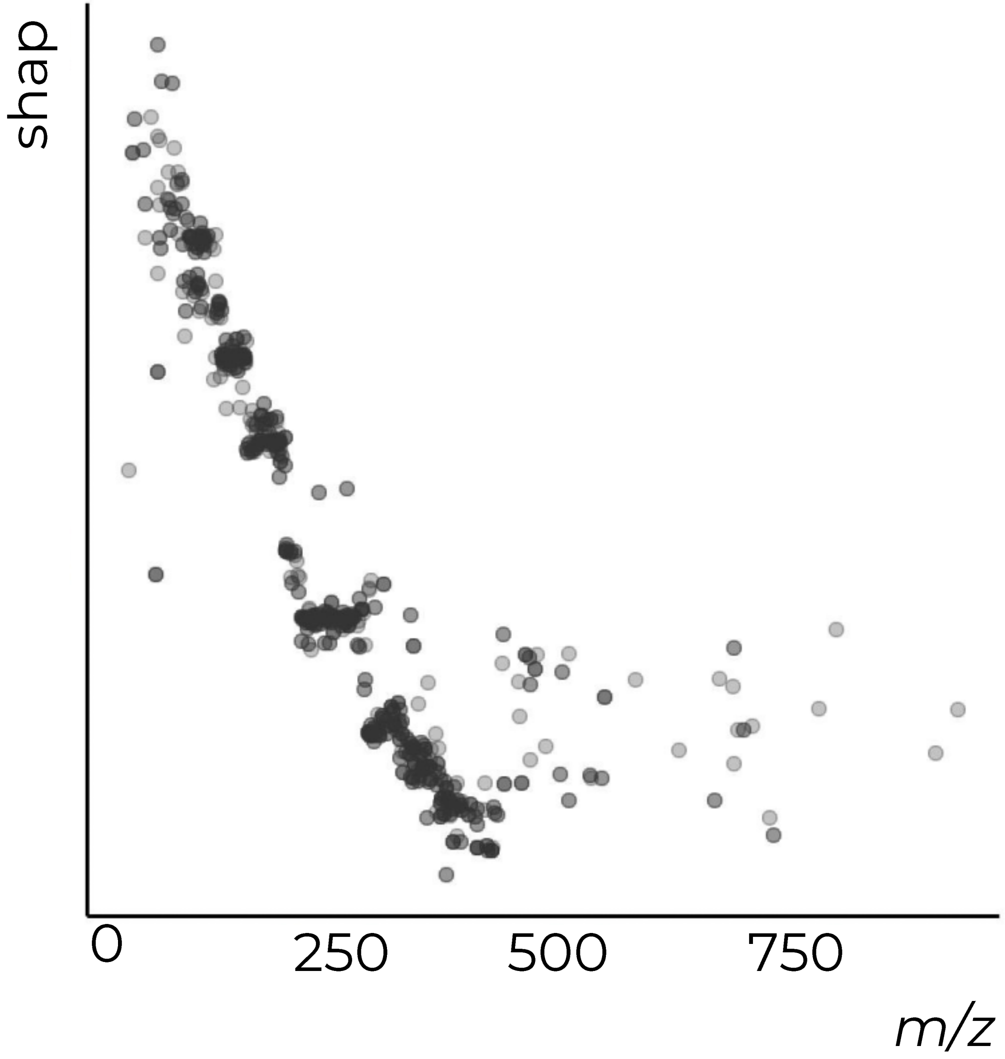
anneli.kruve@su.se

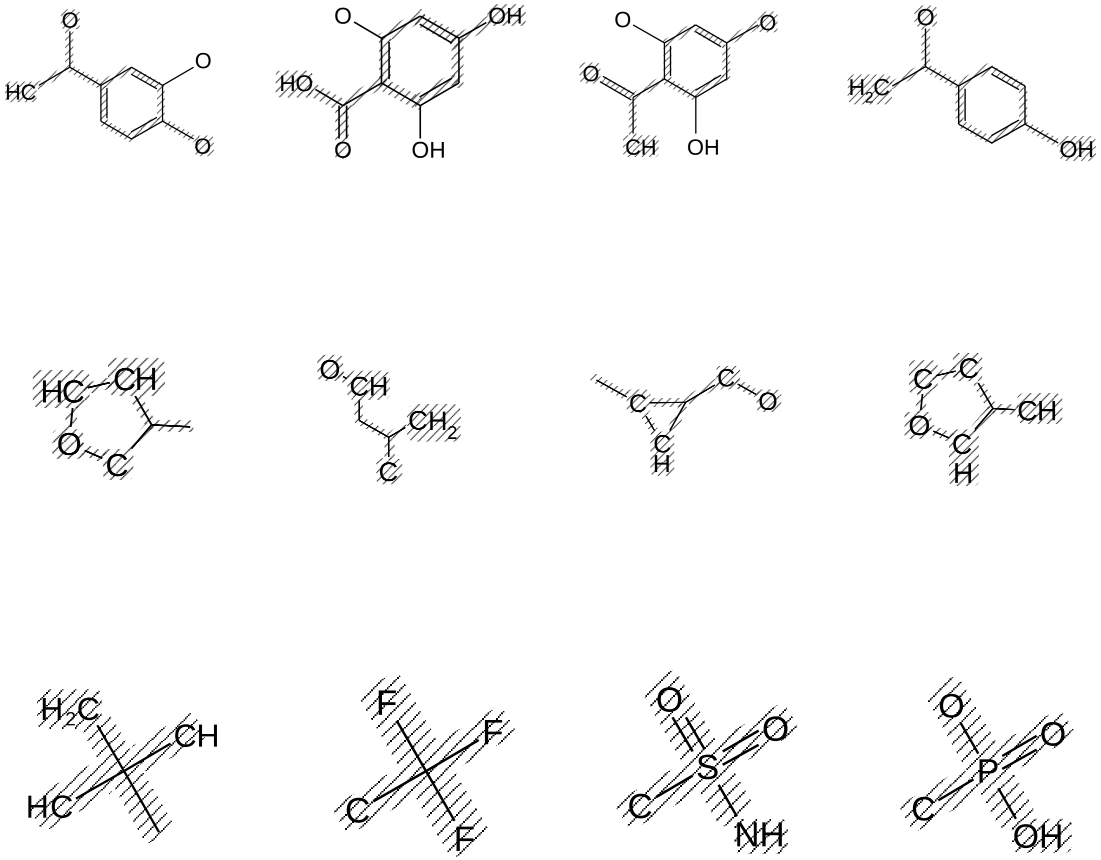
model interpretation





model interpretation





anneli kruve

anneli.kruve@su.se

toxic chemicals

in wastewater

case study on wastewater

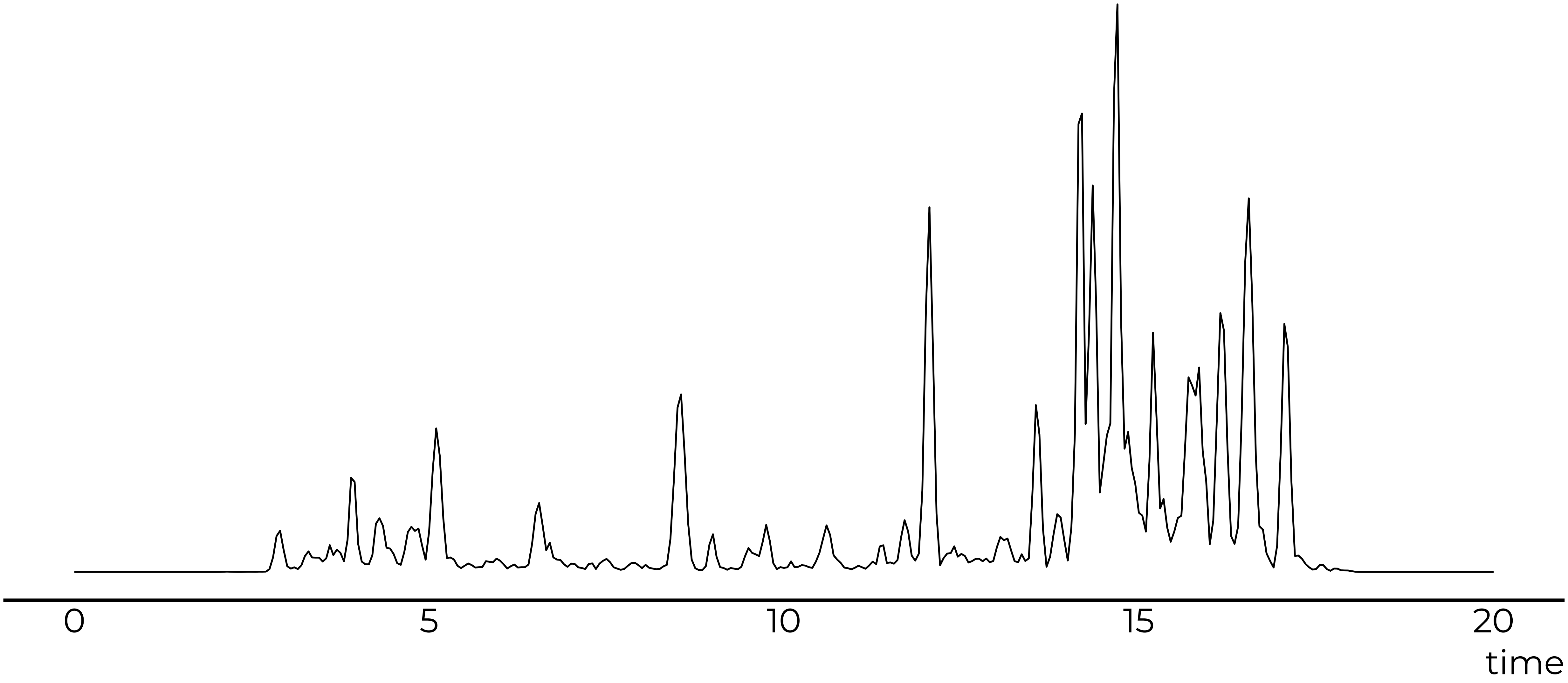
wastewater samples

case study on wastewater

wastewater samples

LC/HRMS analysis

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

case study on wastewater

wastewater samples

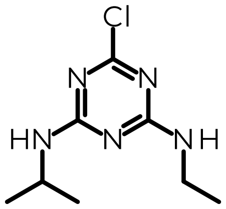
LC/HRMS analysis

molecular fingerprints with SIRIUS+CSI:FingerID

predict LC_{50} with pretrained gradient boosting

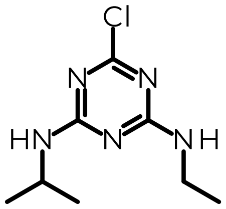
quality control

quality control

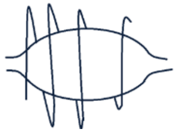


216 analytical standard

quality control

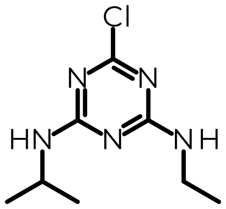


216 analytical standard



DIA and DDA MS^2 data

quality control

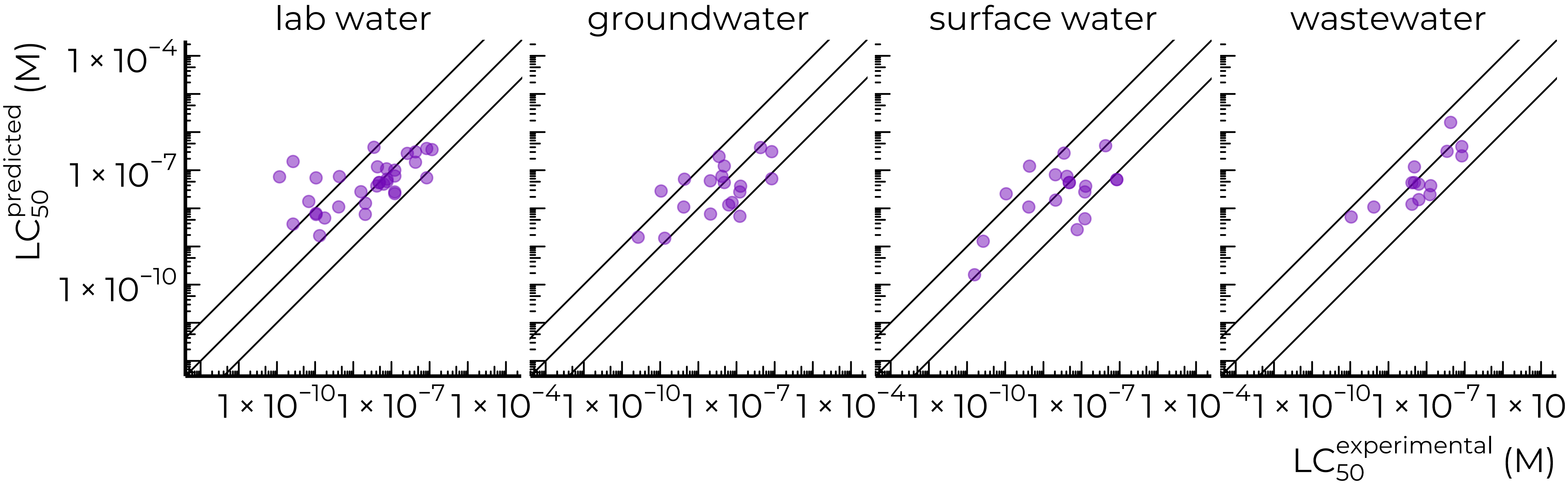


216 analytical standard

DIA and DDA MS^2 data

comparison with experimental LC_{50}

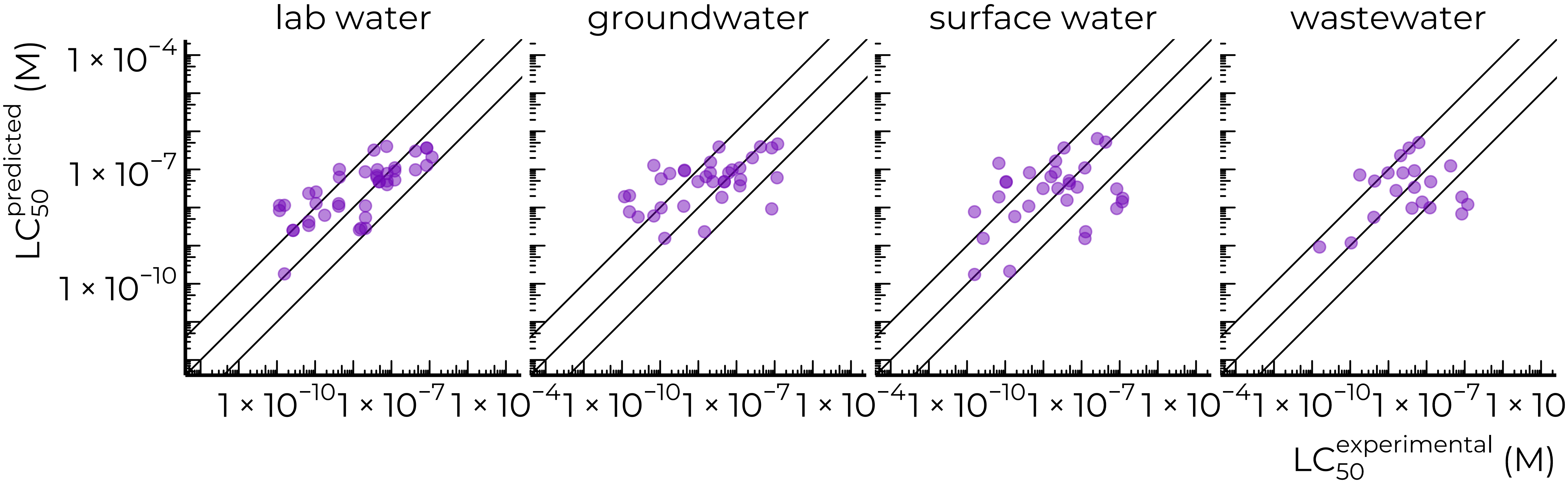
DDA



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

anneli kruve anneli.kruve@su.se

DIA



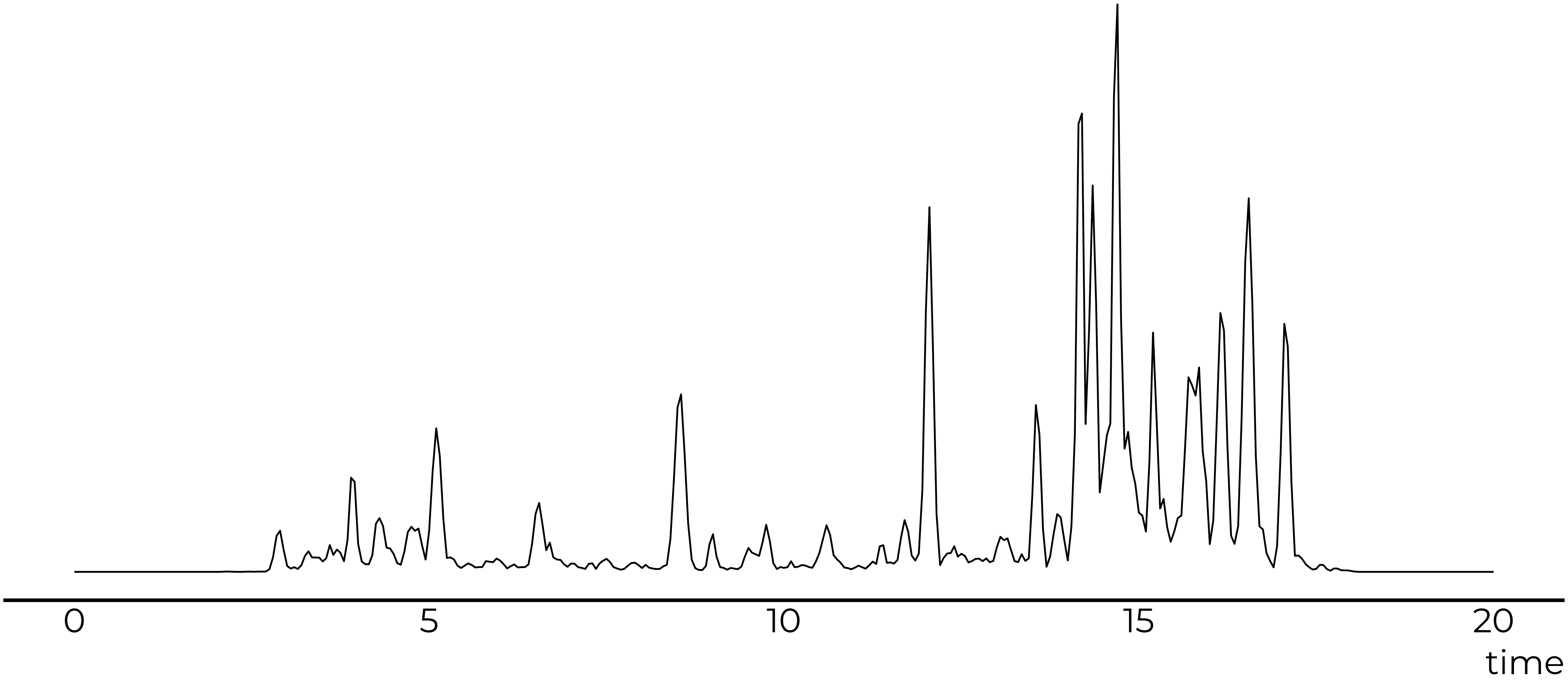
RMSE = 0.85 log-mM 1.09 log-mM 1.18 log-mM 1.03 log-mM

anneli kruve

anneli.kruve@su.se

pinpointing toxic chemicals

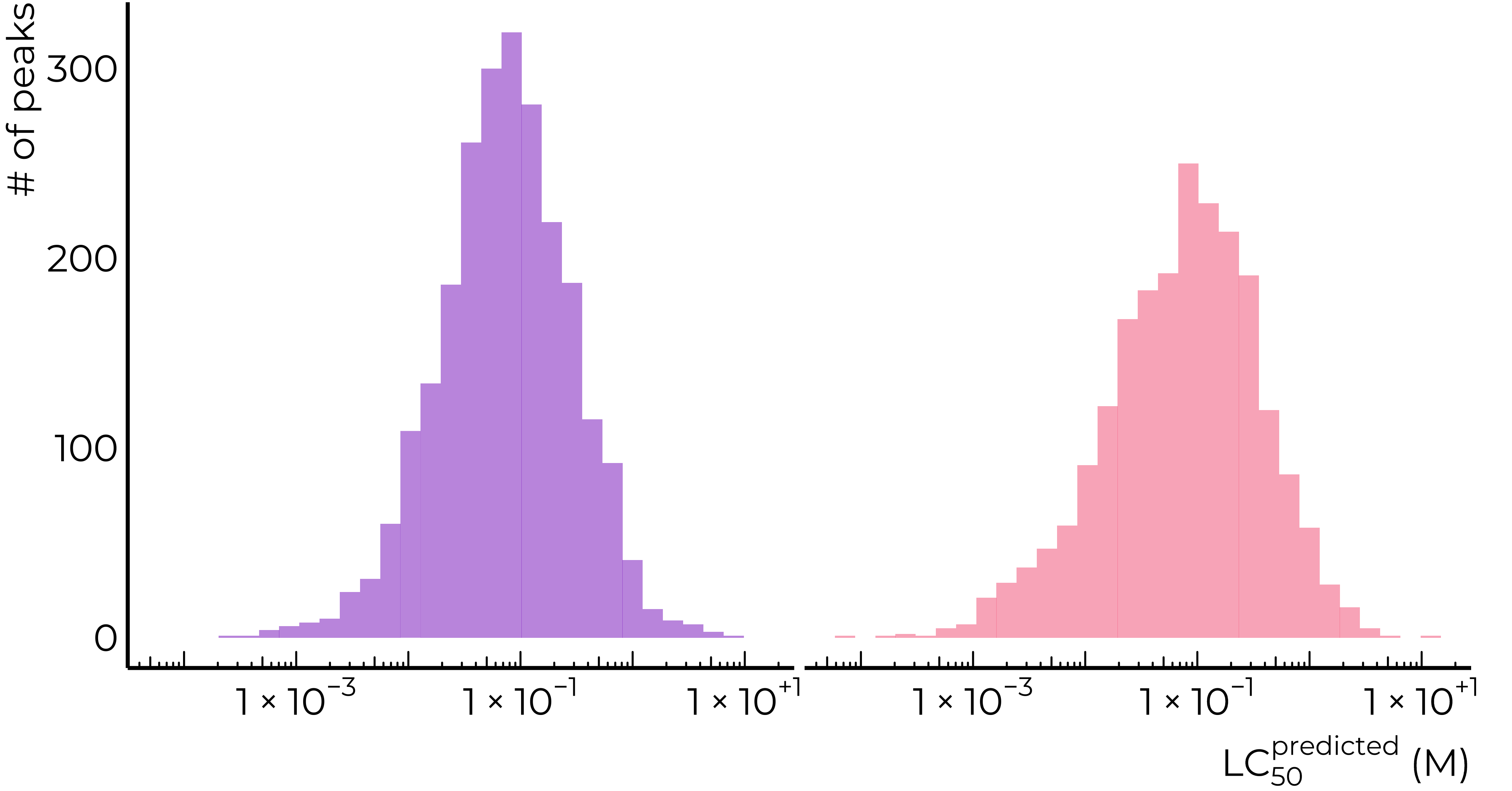
case study on wastewater



anneli kruve

anneli.kruve@su.se

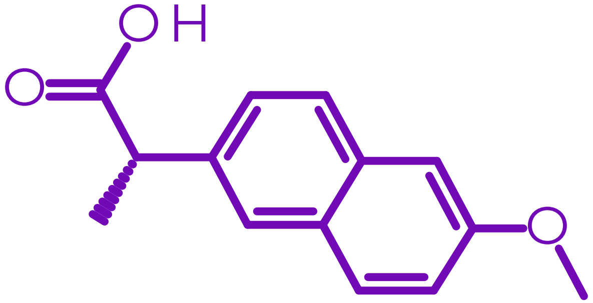
LC_{50} distribution

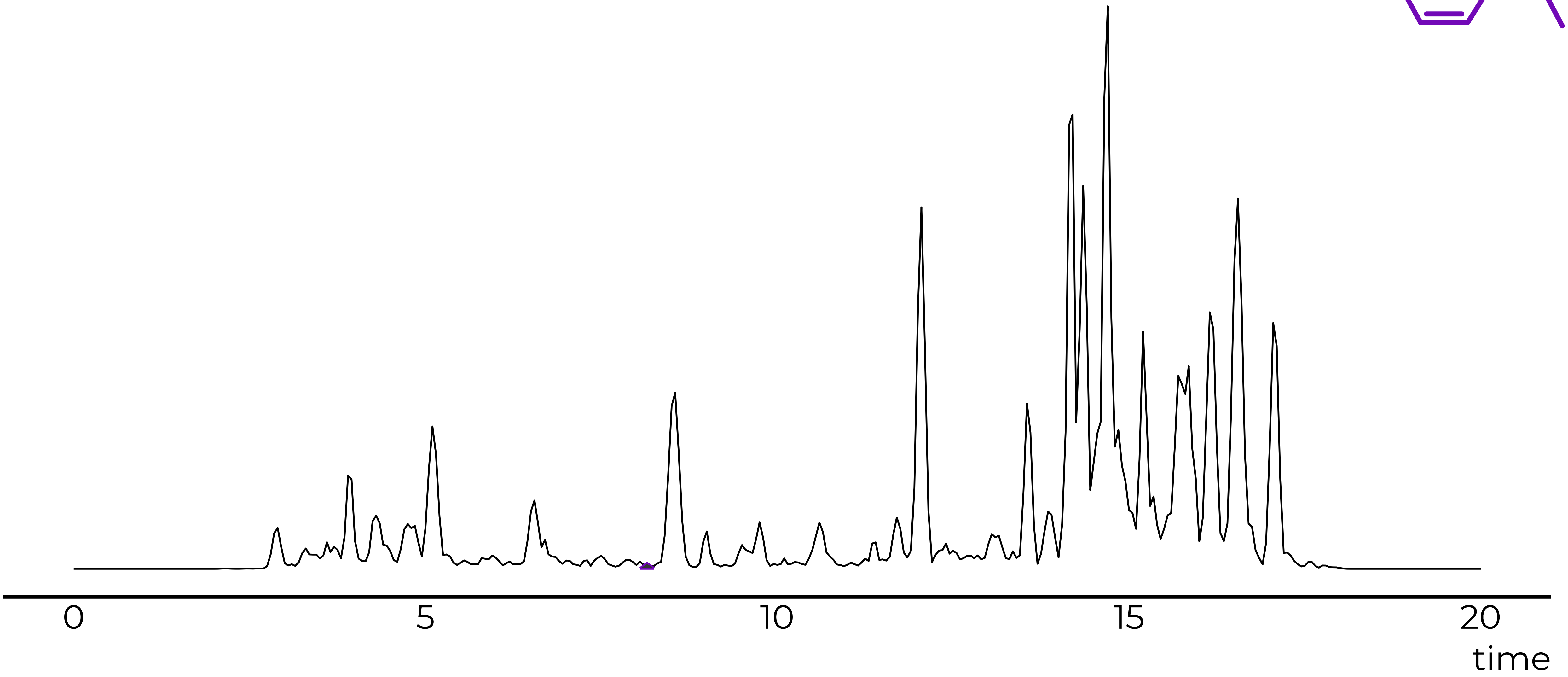


anneli kruve

anneli.kruve@su.se

naproxen

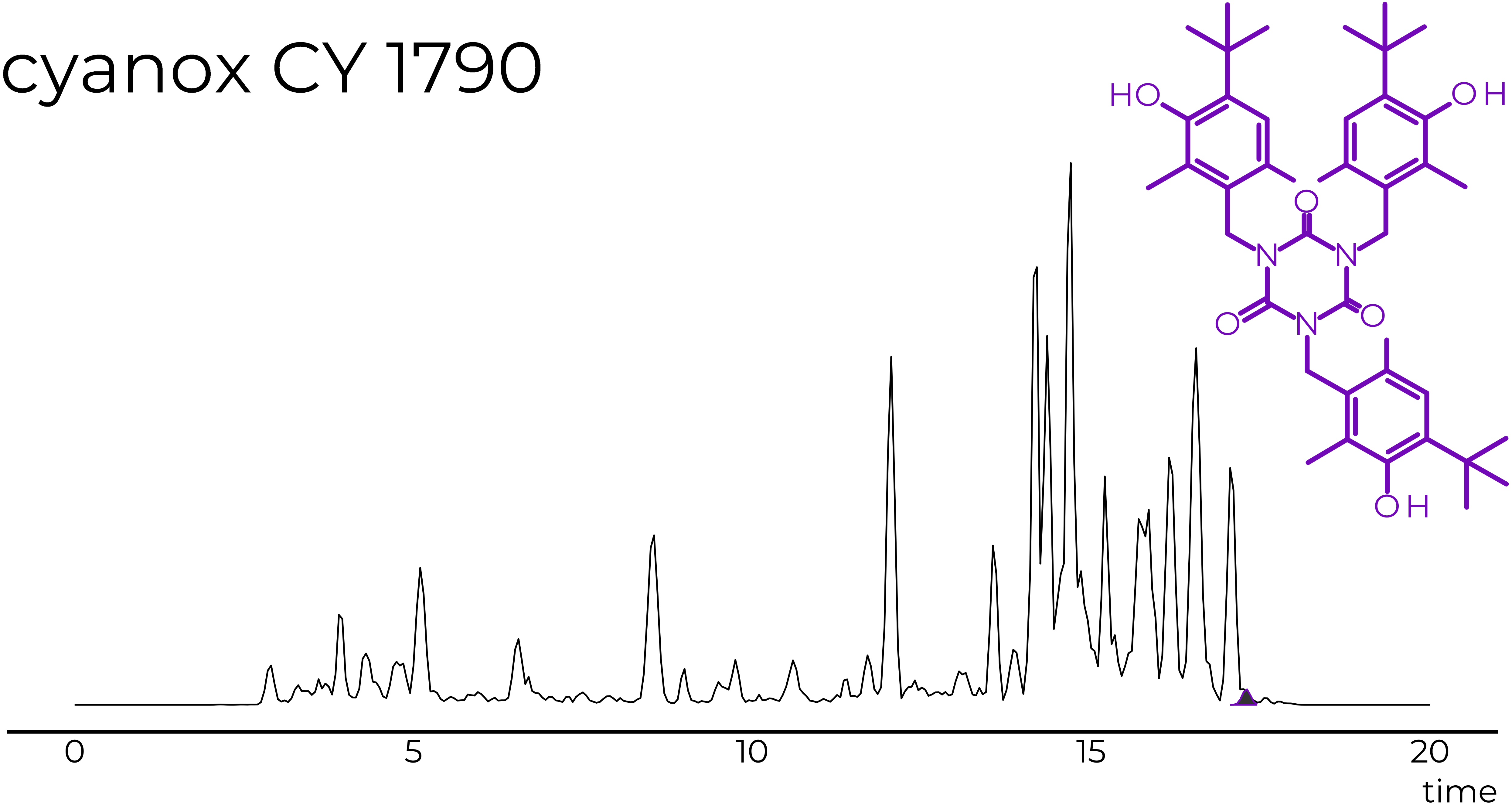




anneli kruve

anneli.kruve@su.se

cyanox CY 1790



summary

prioritization in NTS

toxicity

concentration

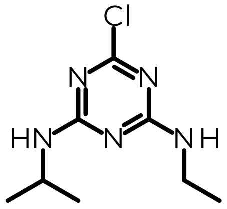
risk

prioritization in NTS

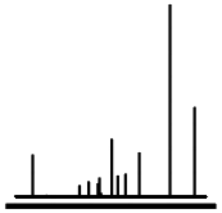
toxicity

concentration

risk



structure



MS^2 spectrum

kruvelab.com

anneli.kruve@su.se