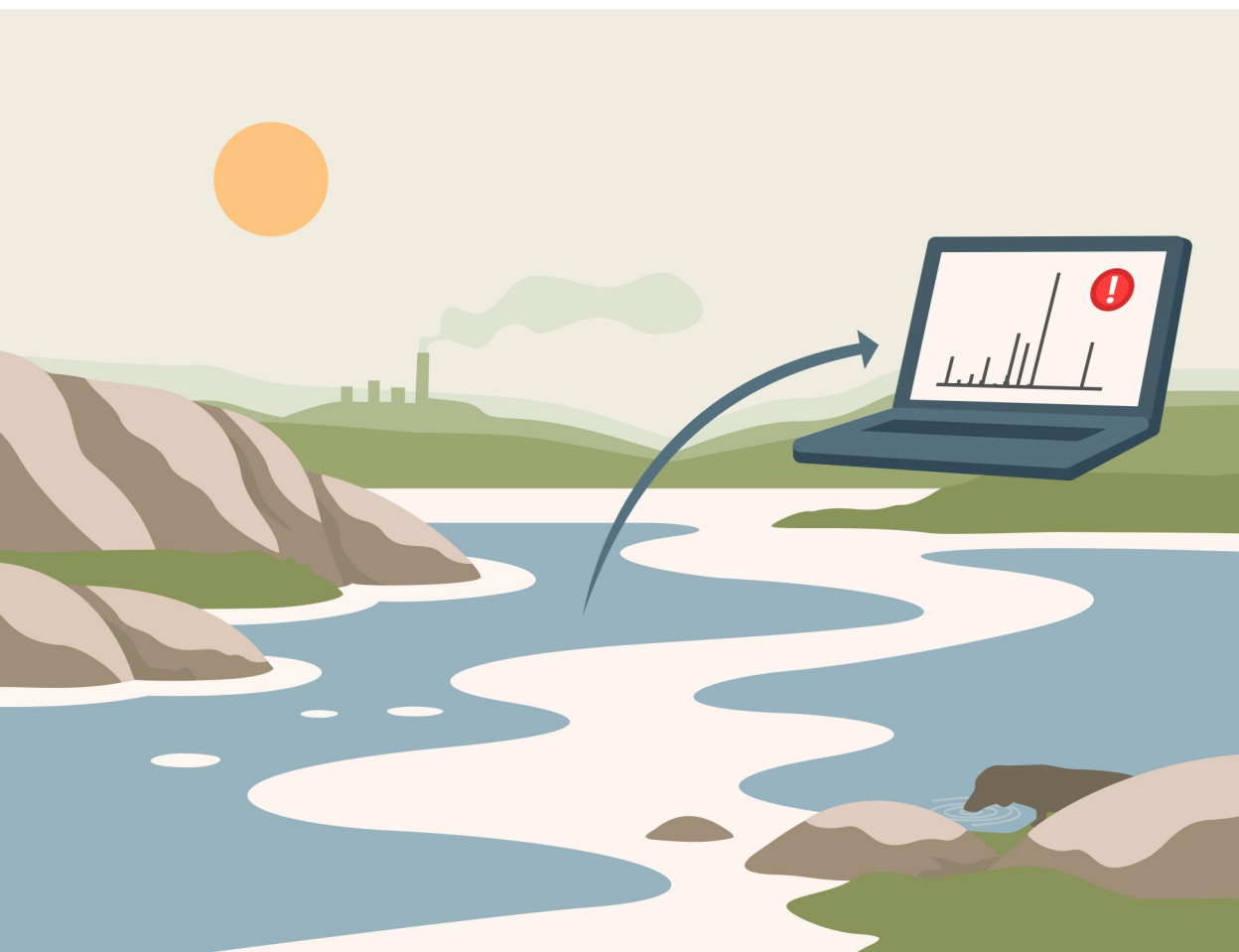


# Machine Learning Tools to Identify Risk Drivers in Water

Helen Sepman





# Machine Learning Tools to Identify Risk Drivers in Water

Helen Sepman

Academic dissertation for the Degree of Doctor of Philosophy in Analytical Chemistry at Stockholm University to be publicly defended on Friday 15 May 2026 at 09.00 in Magnélisalen, Kemiska övningslaboratoriet, Svante Arrhenius väg 16B.

## Abstract

Due to the increasing number of chemicals used in our daily lives, more and more chemicals end up in the environment. Many such contaminants accumulate in water, with thousands of chemicals detected in environmental water samples using liquid chromatography – high-resolution mass spectrometry (LC/HRMS). As a result, all water-dependent organisms are exposed to a large number of low-concentration chemicals, while the health effects of such exposures are unknown. Unfortunately, only a small fraction of the detected chemicals is identified and can be further investigated for their effects on organisms.

This thesis investigated the opportunity to use experimental data of such detected but unidentified chemicals for predicting information regarding their environmental concentration levels, toxicity, and risk - the combination of both. Firstly, in **paper I**, the trends in risk estimation for chemicals detected in water samples were investigated across the years 2019 to 2022. The analysis indicated that risk was considered in only 13% of the papers. In **paper II**, a concentration prediction model, MS2Quant, was developed, allowing concentration prediction for unidentified chemicals based on tandem mass spectra. The experimental data-based concentration predictions were comparable with structure-based predictions. Further, in **paper III**, the predictions from the MS2Quant model were combined with in-house developed MS2Tox model for adult fish acute toxicity predictions in order to prioritize features in wastewater samples. While the feature set of the effluent samples was reduced by 73% to 99%, the subsequent structural assignment with library matching and *in silico* tools could not assign a probable structure for the majority of the prioritized features, highlighting the advantages of incorporating experimental data-based methods in the analysis. Finally, **paper IV** focused on the experimental validation of mixture toxicity predictions. For this, a complementary fish embryo acute toxicity model was developed, and the toxicity values were experimentally validated for eight chemicals. Combined with concentration predictions, the cumulative mixture toxicity was predicted with a  $3\times$  geometric mean error.

The tools developed, investigated, and validated in this thesis showcase the possibility of using available experimental data together with machine learning approaches for exposure and toxicity predictions of unidentified features. They allow looking into a larger subset of detected chemicals for subsequent tandem mass spectra-based prioritization of features that are more likely to cause harm and need immediate attention.

**Keywords:** *non-targeted screening, mass spectrometry, liquid chromatography, machine learning, exposure, toxicity, risk, prioritization.*

Stockholm 2026  
<http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-253770>

ISBN 978-91-8107-574-8  
ISBN 978-91-8107-575-5



Department of Chemistry

Stockholm University, 106 91 Stockholm



MACHINE LEARNING TOOLS TO IDENTIFY RISK DRIVERS IN  
WATER

Helen Sepman





# Machine Learning Tools to Identify Risk Drivers in Water

Helen Sepman

©Helen Sepman, Stockholm University 2026

ISBN print 978-91-8107-574-8

ISBN PDF 978-91-8107-575-5

Cover art adapted from an image generated with DALL-E.

Printed in Sweden by Universitetservice US-AB, Stockholm 2026

# Populärvetenskaplig sammanfattning

Allt fler kemikalier används dagligen. Efter de har uppfyllt sitt syfte, hamnar många av dem i miljön. Till exempel kan bekämpningsmedel spridas till grundvattnet, och läkemedel kan släppas ut i ytvatten på grund av ofullständig avloppsrening. Vattenföroreningar undersöks vanligtvis med hjälp av masspektrometri, vilket kan ge information om elementsammansättningen och funktionella grupper hos de detekterade kemikalier. Dock, är masspektrometri inte tillräckligt för att fullständigt identifiera den unika kemiska substansen. En av de största utmaningarna inom forskning om miljöföroreningar är således att fastställa föroreningarnas struktur – det är inte ovanligt att endast fullständigt identifiera några dussin, eller i bästa fall några hundra av tusentals detekterade kemikalier. Vattenberoende organismer, inklusive människor, exponeras därmed för tusentals okända kemikalier, vars effekter på organismer är okända.

Denna doktorsavhandling fokuserade på utveckling och validering av nya metoder vilka möjliggör förutsägelse av egenskaperna hos detekterade kemikalier baserat på experimentella data, utan att känna till deras struktur. Med hjälp av informationen från masspektra har en maskininlärningsmodell, MS2Quant, utvecklats, som kan förutsäga koncentrationen av en okänd kemikalie med ett femfaldigt geometriskt medelvärde. För riskbaserad prioritering av okända kemikalier i renat avloppsvatten användes sedan MS2Quant tillsammans med en tidigare utvecklad modell MS2Tox. Denna modell förutsäger den akuta toxiciteten för fullvuxna fiskar hos en okänd kemikalie, med hjälp av en liknande arbetsprincip som MS2Quant. Trots att modellerna prioriterade 1% till 23% av de detekterade okända kemikalierna, kunde deras struktur inte fullständigt bestämmas.

På grund av etiska skäl utförs experiment på fisk i allt mindre omfattning. Därför fokuserade det sista projektet på att utveckla en ytterligare modell för att förutsäga akut toxicitet för fiskembryon, och toxicitetsförutsägelsena validerades experimentellt för åtta kemikalier. Tillsammans förutspådde modellerna den totala risken för blandningar bestående av 23 kemikalier med ett trefaldigt geometriskt medelvärde, genom att kombinera koncentrations- och dödliga toxicitetsförutsägelser hos fiskembryon för de kemiska blandningarna.

Sammanfattningsvis fokuserade detta arbete på potentialen av maskininläring för riskprognoser av oidentifierade föroreningar. Modellerna som utvecklades och validerades öppnar upp möjligheten att inkludera fler detekterade föroreningar i riskbaserade studier och hjälper till att prioritera föroreningar som kan utgöra en ökad risk för organismer.

## Populaarteaduslik kokkuvõte

Üha enam kemikaale leiab kasutust inimeste igapäevaelus. Pärast oma eesmärgi täitmist jõuavad paljud neist veeringesse; näiteks võivad taimekaitsevahendid kanduda vihmaveega põhjavette või ravimite toimeained ebapiisava puhastuse tõttu veepuhastusjaamast pinnavette. Veest olevaid saasteaineid uuritakse enamasti vedelikkromatograafia – kõrglahutusmassispektromeetria abil, mis annab teavet detekteeritud kemikaalide võimaliku elementkoostise ja struktuurirühmade kohta, kuid ei ole enamasti piisav kemikaali struktuuri üheseks määramiseks. Seega on üheks suurimaks väljakutseks keskkonnasaaste uuringutes saasteainete struktuuri kindlakstegemine – on tavaline, et tuhandetest detekteeritud kemikaalidest suudetakse üheselt tuvastada vaid mõnikümmend, või heal juhul mõnisada kemikaali. Veest sõltuvad organismid, sealhulgas inimesed, puutuvad seetõttu kokku tuhandete tundmatute kemikaalidega, mille mõju organismidele on teadmata.

Käesolev doktoritöö arendas ja valideeris uusi meetodeid ennustamiseks detekteeritud kemikaalide omadusi nende katseandmetest, vajamata seejuures kemikaalid täpset struktuuri. Kasutades massispektrites sisalduvat informatsiooni, töötati välja masinõppe mudel MS2Quant mis suudab ennustada tundmatu kemikaali kontsentratsiooni viiekordse geomeetrilise keskmise veega. Arendatud mudelit kasutati tundmatute kemikaalide riskipõhiseks prioriseerimiseks puhastatud reovees koos eelnevalt välja töötatud mudeliga MS2Tox, mis ennustab sarnasel tööpõhimõttel tundmatu kemikaali letaalsel kontsentratsiooni täiskasvanud kaladele. Kuigi mudelid prioriseerisid 1% kuni 23% detekteeritud tundmatutest kemikaalidest, ei õnnestunud nende struktuuri kindlaks määrata.

Eetilistel kaalutlustel tehakse kalakatseid üha vähem, seega töötati viimases projektis välja täiendav mudel ennustamiseks letaalsel kontsentratsiooni kala embrüodele, ning letaalsuse ennustused valideeriti eksperimentaalselt kaheksale kemikaalile. Kombineerides kontsentratsiooni ja kalaembrüo letaalse kontsentratsiooni ennustused 23 kemikaalist koostatud keemilistele segudele, ennustasid mudelid segude summarset riski kolmekordse geomeetrilise keskmise veega.

Kokkuvõtvalt keskendus käesolev töö masinõppe võimalikkust riski ennustamiseks identifitseerimata saasteainetele. Välja töötatud ja valideeritud mudelid avavad võimaluse kaasata rohkem detekteeritud saasteained esmasesse riskipõhisesse uuringusse ja aitavad prioriseerida saasteained mis võivad avaldata kõrgendatud riski organismidele.

## List of publications included in the thesis

This is a compilation thesis based on the following scientific papers. Two papers are published, one is submitted, and one is a manuscript.

- I Helen Sepman,\* Louise Malm,\* Pilleriin Peets, Anneli Kruve; **Scientometric Review: Concentration and Toxicity Assessment in Environmental Non-Targeted LC/HRMS Analysis**. *Trends in Environmental Analytical Chemistry*, 2023, 40, e00217.  
\* shared first authorship
- II Helen Sepman, Louise Malm, Pilleriin Peets, Matthew MacLeod, Jonathan Martin, Magnus Breitholtz, Anneli Kruve; **Bypassing the Identification: MS2Quant for Concentration Estimations of Chemicals Detected with Nontarget LC-HRMS from MS<sup>2</sup> Data**. *Analytical Chemistry*, 2023, 95, 33, 12329–12338.
- III Helen Sepman, Lisa Jonsson, Malte Posselt, Theodor Johansson, Magnus Breitholtz, Jonathan Martin, Matthew MacLeod, Anneli Kruve; **High-Risk Contaminants Detected in Wastewater Effluent Samples Can Be Prioritized Prior to Structural Assignment Using Machine Learning Tools**. *Manuscript submitted to Water Research*.
- IV Helen Sepman, Frederic Been, Lisa Baumann, Magnus Breitholz, Matthew MacLeod, Jonathan W. Martin, Anneli Kruve; **Toxic Unit Prediction of Mixtures Using High-Resolution Mass Spectrometric Data Combined with *in vivo* Exposure of Zebrafish Embryos**. *Manuscript*.

Author's contributions:

- I Participated in the study design, performed literature search and data analysis, participated in the manuscript writing and reviewing.
- II Participated in the study design, performed data unification, modelling, analysis of validation data, and data analysis. Led the writing of the manuscript.
- III Participated in the study design, performed LC/HRMS measurements, data processing, identification and data analysis. Led the writing of the manuscript.
- IV Participated in the study design; prepared the mixtures, performed LC/HRMS measurements, fish embryo acute toxicity measurements, and data analysis. Led the writing of the manuscript.

## List of publications not included in the thesis

- V Riccardo Costalunga, Sofja Tshepelevitsh, Helen Sepman, Meelis Kull, Anneli Kruve; **Sodium adduct formation with graph-based machine learning can aid structural elucidation in non-targeted LC/ESI/HRMS.** *Analytica Chimica Acta*, 2022, 1204, 339402.
- VI Helen Sepman, Sofja Tshepelevitsh, Henrik Hupatz, Anneli Kruve; **Protomer Formation Can Aid the Structural Identification of Caffeine Metabolites.** *Analytical Chemistry*, 2022, 94, 10601–10609.
- VII Mélanie Z. Lauria,\* Helen Sepman,\* Thomas Ledbetter, Merle Plassmann, Anna M. Roos, Malene Simon, Jonathan P. Benskin, Anneli Kruve; **Closing the Organofluorine Mass Balance in Marine Mammals Using Suspect Screening and Machine Learning-Based Quantification.** *Environmental Science & Technology*, 2024, 58, 5, 2458–2467.  
\* shared first authorship
- VIII Shirley Pu, James P. McCord, Rebecca A. Dickman, Nickolas A. Sayresmith, Helen Sepman, Anneli Kruve, Diana S. Aga, Jon R. Sobus; **Examining Environmental Matrix Effects on Quantitative Non-Targeted Analysis Estimates of Per- and Polyfluoroalkyl Substances.** *Analytical and Bioanalytical Chemistry*, 2025, 417, 2097–2110.

## Abbreviations

AFT	Adult fish toxicity
Acc	Accuracy
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
EDA	Effect-directed analysis
ESI	Electrospray ionization
FET	Fish embryo toxicity
GC	Gas chromatography
HRMS	High-resolution mass spectrometry
<i>IE</i>	Ionization efficiency
InChI	International chemical identifier
LC	Liquid chromatography
ML	Machine learning
NTS	Non-targeted screening
PaDEL	Pharmaceutical data exploration laboratory
PFAS	Per- and polyfluoroalkyl substances
ppm	Parts per million
Prec	Precision
QSAR	Quantitative structure-activity relationship
Rec	Recall
RF	Response factor
RMSD	Root mean square difference
RMSE	Root mean square error
RQ	Risk quotient
SMILES	Simplified molecular input line entry system
SWS	Surface water suspect list
WWTP	Wastewater treatment plant

# Contents

Populärvetenskaplig sammanfattning.....	i
Populaarteaduslik kokkuvõte.....	ii
List of publications included in the thesis .....	iii
List of publications not included in the thesis.....	iv
Abbreviations.....	v
1. Introduction.....	1
1.1. Water contaminants .....	3
1.1.1. Risk posed by water contaminants .....	3
1.1.2. Toxicity testing .....	4
1.1.3. Extent of the problem .....	5
1.2. Screening tools and approaches.....	5
1.3. Prioritization .....	7
1.3.1. Online prioritization .....	7
1.3.2. Offline prioritization.....	8
1.3.3. Identification as prioritization .....	9
1.3.4. Properties of the identified features.....	10
1.3.5. Experimental mass spectrometry data-based approaches .....	13
2. Objectives.....	15
3. Experimental.....	16
3.1. Literature search .....	16
3.2. Data, descriptors, and modelling.....	16
3.3. Chemicals .....	17
3.4. Samples and preparation.....	18
Instrumentation.....	18

3.5. Experimental fish embryo acute toxicity.....	19
3.6. Software.....	19
3.7. Evaluation metrics.....	19
4. Results and discussion.....	21
4.1. Risk estimation trends in non-targeted screening (paper I)' .....	21
4.2. Development of MS <sup>2</sup> -based quantification model (paper II) .....	23
4.3. MS <sup>2</sup> -based prioritization of wastewater features (paper III) .....	27
4.4. Experimental validation of MS <sup>2</sup> -based predictions (paper IV) .....	33
5. Conclusions and future perspectives .....	37
Acknowledgements.....	40
References.....	42



# 1. Introduction

Increasing numbers of chemicals are produced and used in our daily lives.<sup>1</sup> Many of these chemicals, such as pesticides and pharmaceuticals, end up in the environment after fulfilling their intended purpose.<sup>2,3</sup> Water captures many such unwanted chemicals,<sup>4</sup> called contaminants, and can carry them to surface and ground water through the water cycle. Under environmental conditions, contaminants can further transform into their transformation products through processes such as photodegradation and oxidation, creating complex mixtures with new chemical structures.<sup>5</sup> These mixtures are further consumed by living organisms, which depend on water for survival, including humans. Currently, only a small fraction of environmental contaminants and their potential health effects have been investigated.<sup>6</sup>

The global and poorly reversible distribution of chemicals can have disruptive effects on the Earth's ecosystem.<sup>7,8</sup> It is estimated that the limit for safe operating space has already been exceeded for novel human-made chemicals overall,<sup>9</sup> and even by four per- and polyfluoroalkyl substances (PFAS) alone,<sup>10</sup> while many other persistent contaminants can potentially pose a threat.

Simultaneously, increasing trends in adverse health outcomes associated with environmental pollution have been observed,<sup>11</sup> yet the exact contaminants responsible for these effects are difficult to identify. Chemical pollution is considered one of the drivers of biodiversity loss.<sup>12-14</sup> In human populations, environmental contaminants have been connected to diverse sublethal effects, including neurodevelopmental disorders such as autism and attention-deficit hyperactivity disorder,<sup>15,16</sup> mental disorders,<sup>17</sup> asthma and allergic diseases,<sup>18</sup> and decreased fertility rates.<sup>19,20</sup> Furthermore, due to the high number of low-concentration chemicals in the environment, the focus of toxicity studies has shifted from single chemicals to mixtures of chemicals with potential synergistic or antagonistic effects.<sup>21,22</sup>

State-of-the-art water monitoring tools, such as liquid chromatography – high-resolution mass spectrometry (LC/HRMS), enable the detection of thousands of contaminants by providing information on their detected mass-to-charge ratio and

fragments. The acquired data is further used to identify the detected chemical through comparison with reference information. Still, the identification yields are generally low (<10%)<sup>23-27</sup>; thus, high-resolution experimental data are currently underutilized and many detected chemicals are often discarded from further risk analysis due to the absence of structural annotation.

This thesis aims to develop and validate a workflow for predicting the risk of water contaminants detected with LC/HRMS based on experimental mass spectrometric data. It is hypothesized that with the help of machine learning, such high complexity data can be used for toxicity and exposure predictions prior to identification, providing a useful prioritization step for further analysis and identification of the potentially harmful detected contaminants.

**Paper I** investigates the recent trends in toxicity, concentration, and risk assessments in non-targeted screening of environmental water samples analyzed with LC/HRMS over the years 2019-2022 as a scientometric review.

**Paper II** focuses on the development of the concentration prediction tool MS2Quant, which enables exposure estimation for detected chemicals prior to structural assignment.

**Paper III** investigates the applicability of MS2Quant together with a complementary adult fish acute toxicity prediction model, MS2Tox, for the prioritization of high-risk contaminants in the influent and effluent samples from three wastewater treatment plants in Sweden.

**Paper IV** focuses on the experimental validation of the risk prediction models. In order to experimentally validate the toxicity predictions, a complementary fish embryo acute toxicity model is developed.

## 1.1. Water contaminants

Organic water contaminants largely arise from man-made chemicals that enter the aquatic environment through diverse pathways, which reflect their intended use.<sup>28</sup> This group of chemicals captures a wide variety of chemical structures, and includes pesticides, chemicals in consumer products, and industrial chemicals which can be carried to surface and ground water by the water cycle. Environmental processes, such as photodegradation and abiotic and biotic oxidation reactions, can further alter the structure of the original water contaminant, resulting in new chemical structures, so-called transformation products.<sup>5,29-31</sup> Contrary to expectations, complexity is also introduced by the water cleaning processes, where biodegradation and advanced treatments, such as ozonation and chlorination, also result in new transformation products.<sup>32-36</sup>

Suspect and non-target screening enables the detection of thousands of chemicals in environmental water samples. For example, monitoring of European rivers has revealed the presence of hundreds of known contaminants, many of which have been estimated to be present at concentrations potentially causing chronic long-term effects.<sup>37,38</sup> Still, only a fraction of the detected chemicals with experimental data available are successfully annotated, leaving the unidentified chemicals completely out of any chemical-based risk estimations. Simultaneously, monitoring environmental water samples with bioassays provides a comprehensive overview of the effects, such as endocrine disruption,<sup>39-43</sup> oxidative stress,<sup>44</sup> and mutagenicity,<sup>41</sup> caused by surface water and wastewater effluents. Yet, the chemicals identified in the analysis rarely fully explain the observed effects. Thus, pinpointing the chemicals causing the observed effects is a major current scientific challenge for ecotoxicologists.

### 1.1.1. Risk posed by water contaminants

Any chemical in excess can cause harmful effects. For example, one tablet containing 500 mg of paracetamol administered to an adult can relieve pain and reduce fever, while exceeding the recommended daily dose of 4000 mg can cause severe liver damage.<sup>45</sup> The same chemical can have widely different effects on different species. For example, paracetamol is highly toxic to cats as they lack the key enzyme responsible for its safe metabolism.<sup>46</sup> In other words, every chemical will cause an

adverse effect if its dose/concentration exceeds the safe dose/concentration of that chemical, which in turn depends on the species and the considered effect.

The aquatic toxicity of a chemical is expressed as a concentration at which a certain effect can be observed. Often, lethal concentration to 50% of the species ( $LC_{50}$ ) or the concentration causing toxic effects to 50% of the species ( $EC_{50}$ ) are used to quantitatively describe the toxicity of a chemical.<sup>47</sup> The chemical can cause harm if the organism is exposed to this chemical at a concentration close to or exceeding the  $LC_{50}/EC_{50}$  values. This allows to express the risk quotient (RQ) of a chemical as follows:

$$RQ = \frac{\text{exposure}}{\text{toxicity}} \quad (1)$$

In case  $RQ \geq 1$ , the exposure levels exceed effect levels, meaning that adverse outcomes may occur, indicating a high-risk scenario. The exposure to aquatic organisms can be difficult to estimate as it depends on the uptake, bioavailability, and bioaccumulation; however, environmental concentration is often used in initial risk screening as an approximation.<sup>48,49</sup>

In the case of environmental contaminants, many chemicals are present at the same time, and the toxic effects can already be seen at very low concentrations.<sup>48</sup> Neither the toxicity nor the exposure of these contaminants is simple to determine. The classical approach for risk assessment requires the concentration to be determined with target chemical analysis that can further be compared with experimental or guideline effect values. However, this approach fails in the absence of analytical standards or ambiguity in structural assignment, which is the case for the majority of the environmental contaminants detected with non-targeted screening.

### 1.1.2. Toxicity testing

In order to determine the acute or chronic aquatic toxicity, test organisms of one species are exposed to various concentration levels of the test chemical to obtain the concentration-response curve. The sigmoidal curve is used to derive the effect concentration values, such as  $LC_{50}$  and  $EC_{50}$ . Historically, aquatic toxicity testing was performed *in vivo* on whole organisms such as fish and invertebrates. For ethical considerations, unicellular and small organisms or *in vitro* molecular and cell-based bioassays are increasingly used.<sup>49</sup> For example, acute toxicity tests on adult fish have

been replaced with more resource-effective and ethical fish embryo toxicity testing, which correlates well with toxicity values derived from adult fish.<sup>50</sup>

As the toxicity testing of single chemicals is resource-consuming and requires analytical standards, this approach does not provide a comprehensive picture of complex environmental mixtures with low identification yields. Additionally, single chemical testing does not account for the mixture effects of the chemicals, even though at low concentration levels, concentration addition is considered a suitable approximation for mixture toxicity estimation.<sup>51,52</sup> Instead, the effect-based monitoring of the whole sample can be used to provide an overview of the combined effects of all contaminants, albeit without the information on the toxicity drivers.<sup>49,51</sup>

### 1.1.3. Extent of the problem

Over 350,000 chemicals and mixtures are registered for production globally,<sup>1</sup> which can further undergo structural transformations, increasing the chemical complexity. The largest chemical databases, such as PubChem, currently consist of more than 119,000,000 chemicals.<sup>53</sup> Not all of these chemicals are equally likely to be detected in environmental water samples, yet databases focusing on chemicals that organisms are more likely to be exposed to, such as PubChemLite<sup>54,55</sup> and SusDat,<sup>56</sup> with approximately 120,000 and 560,000 entries, respectively, highlight the scale of the problem. There is a clear need for high-throughput methods to speed up the risk estimation for emerging contaminants.

## 1.2. Screening tools and approaches

Water analysis and monitoring are predominantly carried out using a chromatographic system for the separation of chemicals coupled to a mass spectrometer for their detection. For separation, liquid chromatography (LC) is suitable for monitoring non-volatile or semivolatile contaminants. LC is often the preferred method for water samples as it requires minimal sample preparation (filtering, enrichment), allows (almost) direct injection, and is suitable for separating polar water contaminants.<sup>57</sup> For the separation of volatile and non-polar chemicals, gas chromatography (GC) is a suitable technique.<sup>58,59</sup> For the detection of the separated chemicals, high-resolution mass spectrometers, such as time of flight and Orbitrap,

are widely used due to their excellent resolving power (resolution from 60,000 to 1,000,000) and high mass accuracy (<5 ppm).<sup>59</sup>

In order to analyze the chemicals with HRMS, the chemicals separated with LC need to be transferred from the liquid phase to the gas phase and be positively or negatively charged. For this, electrospray ionization (ESI) is commonly used to couple the LC to HRMS. ESI is considered a soft ionization technique where the chemical is either protonated in positive mode (ESI+) or deprotonated in negative mode (ESI). Excess charge is produced by electrochemical reactions under high voltage, and charged droplets are emitted from the ESI capillary. As the solvent evaporates, the droplets concentrate, and the gas-phase ions of the analyte are formed.<sup>60,61</sup> In the presence of other charged species, such as Na<sup>+</sup>, K<sup>+</sup>, NH<sub>4</sub><sup>+</sup>, and Cl<sup>-</sup>, corresponding adducts beyond protonation/deprotonation can be formed.<sup>60,61</sup>

The charged analyte can be detected, but also separated from other ions for further fragmentation by colliding the selected ion with neutral gas molecules. The fragmentation spectrum (MS<sup>2</sup>) of a chemical carries structural information, as the high-resolution measurements enable the suggestion of the elemental composition of the fragment ions, which in turn can indicate the presence of functional groups in the structure. Additional ionization techniques used with LC include atmospheric pressure chemical ionization (APCI) and atmospheric pressure photoionization (APPI), which are suitable for less polar chemicals but are limited to thermally stable analytes with molecular masses below 3000 Da. Thus, ESI remains the most common ionization method of choice in non-targeted screening with LC/HRMS.

Prior to data analysis, the acquired LC/HRMS data needs to be pre-processed to obtain a set of relevant features with retention time, mass-to-charge ratio ( $m/z$ ), signal intensity, and MS<sup>2</sup> (if acquired) information. This step usually includes feature detection and alignment over replicates, subtraction of blank features, and componentization, i.e., grouping different adducts, isotopes, and in-source fragments corresponding to the same chemical.

Even after pre-processing, thousands of chemical features are detected per sample, resulting in large amounts of data. Specific chemicals of interest can be monitored using optimized analytical methods and confirmed with reference standards – a strategy known as targeted screening. A broader way of analyzing the sample is by looking for  $m/z$  matches with a list of expected contaminants, also known as suspect screening. This approach is especially useful for monitoring previously reported or

suspected, potentially emerging contaminants. Still, completely unknown chemicals, such as new transformation products, would be overlooked with target and suspect screening methods. The broadest screening method, non-target screening, considers all detected features as potentially important. Due to the large number of features analyzed in NTS, experimental data-based prioritization is generally performed to focus the resources on the features of highest importance. The choice of prioritization approach in an NTS study is critical for determining the subset of features that are further analyzed in more detail.<sup>59,62,63</sup>

### 1.3. Prioritization

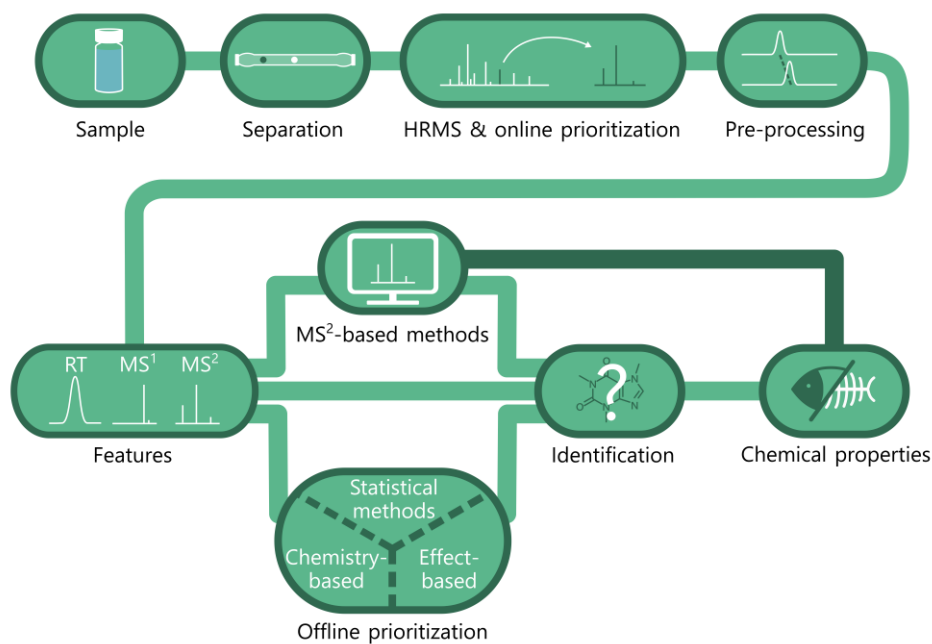


Figure 1. Non-targeted screening workflow.

#### 1.3.1. Online prioritization

Due to instrumental limitations, it is currently impossible to gather high-quality MS<sup>2</sup> data for all detected features. Therefore, the first prioritization step is already

performed when the approach for MS<sup>2</sup> acquisition is chosen (Figure 1). In general, the approaches can be divided into two subcategories – data-dependent acquisition (DDA) and data-independent acquisition (DIA).<sup>59,64</sup>

In the case of DDA, specific ions are selected based on each full scan, which are further isolated and fragmented, resulting in high-quality MS<sup>2</sup> spectra. The disadvantage of the DDA approach is that, due to time limitations, only a fraction of the detected ions is fragmented. The ions that will further be fragmented are selected either based on signal intensity, e.g. top N *m/z* with the highest signal intensity per full scan, or the *m/z* match with the chemicals in the user-provided inclusion list. Often, the inclusion list approach is complemented by the intensity-based acquisition when no precursors match the inclusion list.<sup>59,62,64</sup>

The DIA aims to fragment all ions detected in the full scan, followed by the deconvolution of the MS<sup>2</sup> spectra based on the retention time profile and *m/z*. The fragmentation in DIA can either be performed for all ions detected in full scan at the same time (all ion fragmentation), or by using the Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH), where sequential *m/z* ranges are used after each full scan to reduce the complexity of the data for deconvolution. The frequently observed disadvantages of DIA are the lower quality and reliability of the deconvoluted MS<sup>2</sup> spectra.<sup>64,65</sup>

### 1.3.2. Offline prioritization

After data acquisition and pre-processing, thousands of features detected in the sample can remain relevant for the data analysis. Numerous offline prioritization techniques can be applied depending on the objectives and the research question.<sup>59,62,63</sup>

Chemistry-based methods include suspect screening by matching detected *m/z* with suspect lists or databases, depending on the aim of the research.<sup>66,67</sup> Additionally, unique compound properties can be searched for in a full scan, such as isotopic patterns and negative mass defects for halogenated chemicals.<sup>68–70</sup> Similarly, MS<sup>2</sup> data can be used. For example, the patterns in the fragmentation spectra can indicate the presence of toxicity-related structural alerts, i.e., functional groups that have been related to higher toxicity of a chemical.<sup>71</sup>

The detected features can also be prioritized based on the known process, event, or site, often with the help of statistical analysis. In such cases, additional data about the samples must be available, for example the influent and effluent sample of the same wastewater treatment plant<sup>23</sup> or site-specific samples over certain time intervals.<sup>72</sup>

Effect-based methods are guided by the experimental toxicity of the sample. While the effect-based monitoring of the whole sample helps to prioritize samples of interest, the further analysis of such samples can also be guided by toxicity. Effect-directed analysis (EDA) is a commonly used approach for the prioritization of chemical features.<sup>49</sup> The contaminants are separated with chromatography and simultaneously analyzed with a mass spectrometer and fractionated for toxicity analysis. Fractions with adverse effects are related back to the chemicals detected with the mass spectrometer in the retention time range. While EDA is a powerful method for narrowing down the features of interest, the observed toxicity is generally not explained by the contaminants identified in the chemical analysis due to low identification yields and higher sensitivity of the bioassays that can result in LC/HRMS signals below the limit of detection.<sup>40,41,43,73-76</sup>

### 1.3.3. Identification as prioritization

Prioritization helps to direct focus to a subset of detected features, which need to be identified in order to draw conclusions regarding chemical-specific properties or risk. While mass spectrometric data is informative, this alone is generally insufficient to unequivocally determine the structure of the contaminant; therefore, different identification approaches can be used, often in parallel.<sup>77,78</sup> Firstly, the fragmentation spectrum of a detected feature can be matched against experimental spectral libraries. Many such libraries, such as MassBank,<sup>79</sup> GNPS,<sup>80</sup> METLIN,<sup>81</sup> etc., have been compiled through collaborative efforts and include experimental spectra for thousands of unique chemicals. Gathering experimental data is resource-intensive; thus, experimental spectral libraries only cover a fraction of known chemical space. In the effort to expand the annotatable chemical space, *in silico* generated spectral libraries have been created, which aim to imitate experimental spectra.<sup>82-84</sup> The *in silico* spectra can suffer from generating noise and missing important fragments and are thus less reliable than the experimental reference spectra. Alternatively, identification approaches extracting structural information from the MS<sup>2</sup> spectra with the help of machine learning (ML) have proven effective for compound identification. Software such as

SIRIUS+CSI:FingerID<sup>85</sup> and MIST<sup>86</sup> allow the comparison between MS<sup>2</sup>-based predicted structural information with database structures for which the same information can be calculated, removing the need for experimental data and expanding the explorable chemical space. As these methods still need structural databases to suggest candidate structures, the recent trends in chemical space expansion include generative modelling for elucidating the structure of novel chemicals.<sup>87,88</sup>

Even though the identification methods are constantly being improved, currently, only a fraction of features with available MS<sup>2</sup> data are structurally annotated, and many suggested candidate structures can have low confidence.<sup>89</sup> Thus, the identification process that depends on the existing databases can also be interpreted as prioritization based on current knowledge of potential environmental contaminants. As further steps rely on structural annotation, large numbers of detected features are completely discarded from further analysis.

#### 1.3.4. Properties of the identified features

In order to investigate the risk (or other relevant chemical properties) of the detected features, different approaches can be applied depending on the reliability of structural annotation. The highest confidence in structural annotation, that is, level 1 on the widely used NTS annotation confidence scale,<sup>90</sup> can only be achieved if the experimental data of the detected feature matches the measured data of an analytical standard. On the same scale, MS<sup>2</sup> match with the experimental spectral libraries can bring the annotation confidence to level 2, while the *in silico* and ML-based methods provide a list of candidate structures with varying confidence scores and can therefore only achieve level 3 on the 5-level annotation confidence scale.

In the best-case scenario, when an analytical standard is available, the relevant toxicity can be experimentally determined, and the environmental concentration of a chemical can be determined by constructing a calibration graph. However, this approach is resource-intensive and impossible if standards are not available. Furthermore, for the majority of the detected features in environmental samples, the annotation confidence remains at level 3 or lower. With the expanding list of emerging contaminants that require investigation, standard-free approaches are needed to speed up the process.

### *Toxicity (standard-free)*

For structurally annotated features, toxicity databases<sup>91,92</sup> can be used to screen for toxic effects; however, the data is largely missing for newly detected contaminants and transformation products. Alternatively, computational tools offer a possibility to estimate toxicity from the structure in a rapid and high-throughput manner. Many prediction models have been trained for numerous species and endpoints.<sup>93</sup> For example, the widely used Ecological Structure Activity Relationships (ECOSAR) Predictive Model predicts acute and chronic effect levels for fish, daphnid, and algae.<sup>94</sup> The prerequisite of training such models is relevant training data availability, preferably including a wide variety of chemical structures.

### *Concentration (standard-free)*

Determining the concentration of a chemical detected with LC/ESI/HRMS is not trivial. The chemicals need to be ionized for the analysis and detection with a mass spectrometer, and the fraction of the analyte molecules that will be ionized, also known as ionization efficiency (*IE*), differs widely between chemicals.<sup>95-98</sup> The *IE* is affected by both the chemical properties, such as hydrophobicity,<sup>99</sup>  $pK_a$ ,<sup>61</sup> and gas-phase proton affinity,<sup>100</sup> and the conditions during elution, such as aqueous pH,<sup>101</sup> organic modifier percentage,<sup>102</sup> and presence of  $NH_4^+$  ion.<sup>103</sup> Differences of more than 7 orders of magnitude in *IE* have been observed between different chemicals, and thus, the signal/peak area alone cannot be used for concentration estimation of detected chemicals.<sup>98</sup>

Several approaches have been developed to estimate the concentration of a detected *a priori* unknown chemical without an analytical standard. A recent interlaboratory study investigated commonly applied strategies – using a close eluting calibrant, surrogate standard, and *IE* prediction models for quantification.<sup>104</sup> The best concentration estimations were achieved by using the *IE* prediction models, showcasing the utility of machine learning modelling in understanding chemical processes. All tested approaches relied on the structural annotation, except for the closest eluting calibrant approach, which also performed the worst out of all tested approaches.

The surrogate standard approach has proven useful in specific cases, such as PFAS quantification with a standard from the homologue series. Usually, the surrogate

standard would consist of the same hydrophilic head group but would differ in the fluorinated alkyl chain length. The differences in the performance of the surrogate quantification and *IE* prediction model approach were shown to be statistically insignificant for PFAS.<sup>105</sup> The advantage of the latter is the applicability in the absence of suitable surrogate standards.

In order to train a general *IE* prediction model, high-quality calibration data for a wide range of chemical structures is needed. Several *IE* prediction models, typically focusing on specific chemical classes, have been developed using training datasets of up to 100 chemicals.<sup>106–115</sup> Only a limited number of more general models have been trained on datasets containing a few hundred chemicals.<sup>96,116–120</sup> As the functional groups that play a crucial role in the ionization process differ between positive and negative mode ESI, separate models are generally trained for the two modes.

### *Concentration prediction models and workflow*

The *IE* prediction models predict the extent of chemical ionization relative to a reference chemical. In other words, the predicted *IE* value is on a relative scale and does not consider the experimental conditions, such as source cleanliness, instrument, software, etc., which affect absolute signal intensities in the LC/ESI/HRMS run. In order to predict the experiment-specific response factor (RF), i.e. the slope of the calibration graph, for the chemical of interest, the *IE* predictions need to be calibrated. For this, calibration graphs for a set of calibrants with varying *IE*s are measured in the same run with the chemical(s) subject to quantification, and the RFs in the linear range are determined. As the structures of the calibrants are known, the *IE* values can be predicted, and the linear regression is used to relate the predicted *IE* values to experiment-specific RFs. The latter is used to convert the predicted *IE* to the corresponding RF for the chemical subject to quantification. The concentration estimation is obtained by dividing the signal area by the predicted RF (Figure 2).<sup>96</sup>

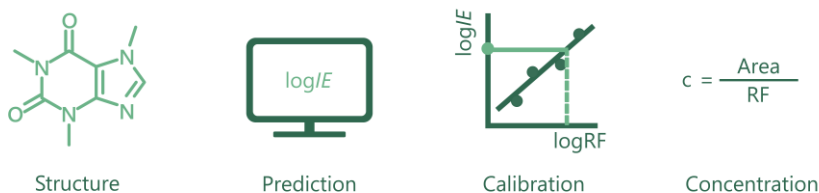


Figure 2. Concentration prediction workflow using an ionization efficiency prediction model (left to right).

### 1.3.5. Experimental mass spectrometry data-based approaches

Currently, the main requirement for predicting the risk and chemical properties of the detected features in NTS without an analytical standard is the structural annotation of the chemical. If the structure cannot be assigned, identification becomes a bottleneck in risk evaluation. In most cases, the unidentified features are discarded from further analysis, and high-quality mass spectrometry (MS) data that include structural information of the detected features are ignored.

Thus, this thesis focuses on the development and application of tools that bypass the identification bottleneck. These tools enable estimating properties of the still unidentified chemical corresponding to the detected feature based on relevant structural information extracted from the MS<sup>1</sup> and MS<sup>2</sup> spectra. Importantly, this work focuses on investigating the utility of such information for risk prediction rather than developing new tools for extracting structural information from MS<sup>2</sup> spectra.

The high-resolution MS data is high-dimensional (every potentially separable  $m/z$  can be thought of as one dimension), with a sparse data population. Thus, dimensionality reduction for capturing the most relevant structural information is often needed to reduce noise and computational cost. In addition to direct spectrum-based feature extraction, such as fragments and neutral losses, ML algorithms, including deep learning models and fragmentation tree-based approaches, have been shown to successfully capture chemically relevant information from MS data (Figure 3). For example, SIRIUS+CSI:FingerID<sup>85</sup> software predicts the presence or absence of substructures in a detected chemical. These substructures are predefined, and for each substructure, a support vector machine model has been trained to predict the probability of a substructure from the MS data.<sup>121,122</sup> The vector with presence/absence information of substructures, known as the structural fingerprint of a chemical, can also be calculated for a known structure.

These approaches have been used for mapping chemical space,<sup>123</sup> improving spectral matching,<sup>124</sup> predicting chemical properties such as  $\log P$ ,<sup>125</sup> and allowing identification based on structural databases and *de novo* generated structures.<sup>85,86</sup> In the context of environmental monitoring and risk predictions, the structural fingerprints extracted from MS<sup>2</sup> spectra have been successfully used for predicting adult fish acute toxicity,<sup>126</sup> and bioassay activity for endocrine disruption.<sup>127-129</sup> Additionally, fragments and neutral losses have been used to pinpoint potentially hazardous chemicals.<sup>130</sup> A few

studies have combined the MS<sup>2</sup>-based toxicity and concentration predictions for risk predictions.<sup>66,131</sup>

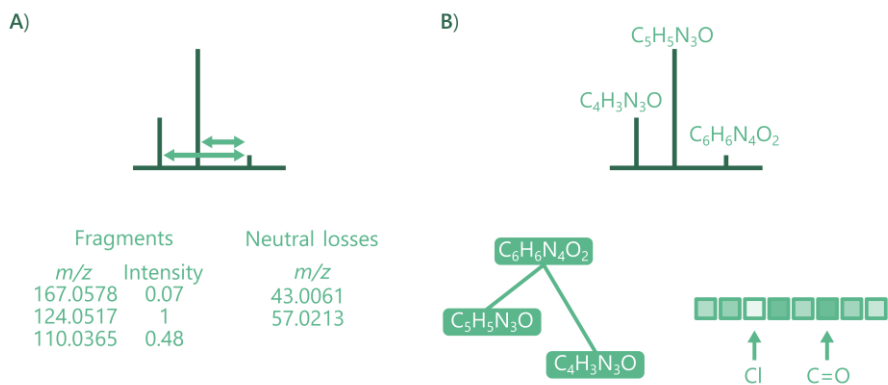


Figure 3. Methods for extracting information from MS data. A) Spectrum-based feature extraction summarizing all fragments and neutral losses, and B) fragmentation tree-based structural fingerprint prediction (working principle of SIRIUS+CSI:FingerID).

## 2. Objectives

The state-of-the-art analytical tools provide high-quality data for thousands of detected contaminants in the environmental water samples. At the same time, only a fraction of this information is used, as the strategies applied in NTS require a successful structural assignment as a basis for risk estimation. It is hypothesized that the mass spectrometric data that carry structural information can be used for risk prediction, even without successful structural assignment. Therefore, the objectives of this doctoral thesis were to:

- 1) Review the recent trends in assessing the toxicity, concentration, and risk for chemicals detected in environmental water samples with the NTS workflow.
- 2) Develop and improve experimental mass spectrometric data-based quantification and toxicity prediction tools that increase the number of features subject to risk predictions.
- 3) Apply and validate the developed tools in NTS workflows for the prioritization of high-risk contaminants.

## 3. Experimental

### 3.1. Literature search

For the scientometric review (**paper I**), the literature search was conducted using the Web of Science search engine. We aimed to gather all research papers that applied non-target screening on water samples published over the years 2019 to 2022. The exact search query was: “(TS=(non-target) OR TS=(nontarget) OR TS=(untarget)) AND (TS=(water) OR TS=(wastewater) OR TS=(groundwater)) AND DT=(Article) AND PY=(2019–2022)”. As the Web of Science search engine screens abstracts and not full texts, we excluded keywords relating to LC and MS from the search, as they would have excluded many relevant papers. The usage of the LC system was manually checked for all papers resulting from the query.

### 3.2. Data, descriptors, and modelling

In **paper II**, 13 datasets of calibration graph slopes from different sources were unified to compile the training data for the *IE* prediction model in ESI+ mode. In total, the dataset consisted of 6049 datapoints corresponding to 1191 unique chemicals. The data was divided into training and test sets with a ratio of 80/20 based on the unique International Chemical Identifier (InChI) of chemicals, ensuring that all the datapoints corresponding to the same chemical would be in either training or test set. The structural fingerprints were calculated in R with an in-house function corresponding to fingerprints of the SIRIUS+CSI:FingerID (version 4.9.15). For comparison with the structure-based quantification method, Pharmaceutical Data Exploration Laboratory (PaDEL) descriptors, Mordred, extended-connectivity fingerprints (ECFP), and minHashed atom-pair fingerprints (MAP4) were calculated and used in *IE* model training. Different extreme gradient boosting-based algorithms (xgbTre, xgbLinear, xgbDART) were used for model training with five-fold cross-validation for hyperparameter tuning.

In **paper IV**, the dataset from the ECHA report analyzing relevance of fish embryo acute toxicity test guidance<sup>132</sup> was used as the basis for model training. After data cleaning, 752 chemicals with LC<sub>50</sub> values measured under differing experimental conditions remained in the dataset. To enable experimental validation of the model

based on OECD Test Guideline 236, a dataset of LC<sub>50</sub> values was compiled by transferring data to experimental conditions corresponding to the test guidelines based on linear regression analysis, resulting in 209 unique chemicals. Additionally, the adult fish LC<sub>50</sub> data used in MS2Tox modelling were added to the dataset based on the linear regression of common chemicals (N=51,  $R^2 = 0.83$ ) to increase the data for model training. The 23 chemicals used for validation were removed from the training data. Structural fingerprints were calculated for all chemicals based on Simplified Molecular Input Line Entry System (SMILES) using the *fingerprinter* function from SIRIUS+CSI:FingerID software (version 5.8.7). The final dataset consisted of 988 unique chemicals, and the final toxicity prediction model was trained with xgbTree algorithm using five-fold cross-validation. Additionally, the MS2Quant model for ESI+ was retrained on datasets where the chemicals used in validation were removed from the training. The data was re-standardized, removing some duplicate chemicals. Final training data consisted of 1140 unique chemicals. The MS2Quant for ESI- was trained on a dataset of 268 unique chemicals. The models were trained on the whole dataset using the xgbTree algorithm and five-fold cross-validation.

### 3.3. Chemicals

In **paper II**, chemicals used in a NORMAN interlaboratory study<sup>104</sup> were used for model validation. For model calibration, the calibration graph slopes of 36 chemicals were used, and the concentration was predicted for 39 chemicals spiked in lake water at two concentrations.

For the development of the prioritization approach in **paper III**, a mixture of 159 environmentally relevant contaminants was prepared and spiked into methanol/water (20/80, v/v) with a final concentration of 50 µg/L, and into influent and effluent samples from one WWTP at three concentration levels (final concentrations of 50, 5, and 0.5 µg/L).

The risk predictions in **paper IV** were validated on the set of 23 chemicals, which were used to prepare different designed mixtures. These chemicals were selected from the dataset of the ECHA report for Zebrafish embryo LC<sub>50</sub> values and were removed from model training. The MS2Quant positive and negative mode *IE* models were calibrated by using calibration graph slopes of 28 and 19 calibrants for ESI+ and ESI-mode, respectively.

### 3.4. Samples and preparation

In **paper III**, influent and effluent samples from three WWTPs in Sweden were investigated. The samples were stored at -20 °C until the analysis. Before the measurements, the samples were allowed to thaw at room temperature and filtered using 0.45 µm pore size syringe filters.

### Instrumentation

The measurements in **papers II, III, and IV** were performed using Dionex UltiMate™ 3000 UHPLC liquid chromatography system (Thermo Fischer Scientific) coupled to Q Exactive Orbitrap HRMS (Thermo Fischer Scientific).

For **paper III**, ACQUITY UPLC HSS T3 column (100 Å, 1.8 µm, 2.1 mm × 100 mm, 1/pk, Waters) and ACQUITY UPLC HSS T3 VanGuard pre-column (100 Å, 2.1 mm × 5 mm, 3/pk, Waters) were used. Both the aqueous mobile phase and methanol as organic mobile phase contained 10 mM acetic acid. The flow rate was held at 0.4 mL/min, column temperature at 40 °C, and autosampler temperature at 5 °C. The content of organic mobile phase was increased linearly from 5% to 95% over 10 min, was held constant for 2 min, decreased back to 5% in 0.1 min and kept at 5% for 2.9 min for equilibration before the next injection. The injection volume of 100 µL was used for all samples. The measurements were performed in ESI+ mode and fragmentation spectra were acquired using three acquisition methods: top 5 highest intensity features per full scan, inclusion list based on expert curated suspect list, and data-independent acquisition where every full scan measurement was followed by five sequential window acquisitions of all theoretical mass spectra (SWATH-MS).

For **papers II and IV**, Kinetex EVO column (2.6 µm, C18, 100 Å, 150 mm × 3.0 mm, Phenomenex) was used. Aqueous phase with 0.1% formic acid and acetonitrile as organic mobile phase were used in ESI+ mode measurements. Additionally, ESI- mode measurements were performed in **paper IV** with the aqueous phase with 5 mM ammonium acetate. The flow rate was kept at 0.35 mL/min, column temperature at 40 °C, and autosampler temperature at 5 °C. The content of organic phase was increased linearly from 5% to 100% over 20 min, was held constant for 5 min, decreased back to 5% in 0.1 min and was kept at 5% for 2.9 min for equilibration before the next injection. The injection volume of 10 µL was used for all samples. The fragmentation spectra were acquired using the inclusion list combined with acquiring

MS<sup>2</sup> for top 5 most intensive peaks in the absence of precursor mass matches with the inclusion list.

### 3.5. Experimental fish embryo acute toxicity

In **paper IV**, fish embryo acute toxicity testing was performed for 8 chemicals following the OECD Test Guideline No. 236. The fertilized wild-type zebrafish (*Danio rerio*) embryos were placed into pre-incubated 24-well plates, one embryo per well, and exposed to the test chemical. All 24 embryos per plate were exposed to the same concentration level of the test chemical. Similarly, positive and negative control plates (24 embryos each) were analyzed in parallel. Lethality was assessed at 96 hours post fertilization based on endpoints listed in the OECD guideline (coagulation of the fertilized egg, lack of somite formation, lack of detachment of the tail-bud from the yolk sac, and lack of heartbeat).

### 3.6. Software

In **papers II, III, and IV**, MS-DIAL<sup>133,134</sup> (versions 4.80 and 5.5.250820) was used for feature alignment, integration, and deconvolution of DIA data. Further processing of the data was performed using R (versions 4.1.3, 4.2.2, 4.3.2, and 4.4.2) and RStudio (versions 2022.12.0 and 2024.12.0). SIRIUS+CSI:FingerID<sup>85</sup> software (versions 4.9.15 and 5.8.7) was used for the prediction of probabilistic fingerprints for the detected features and *in silico* structural assignment. MetFrag<sup>82</sup> (version 2.6.1) was used for *in silico* structural assignment in **paper III**.

### 3.7. Evaluation metrics

In **papers II, III, and IV**, the squared Pearson correlation coefficient  $R^2$  was used for correlation analysis. Additionally, in **paper III**, the squared Spearman correlation coefficient was used for MS<sup>2</sup>-based priority score and risk quotient analysis. The prediction errors were analysed using root mean square errors (RMSE), and prediction fold errors were calculated as follows:

$$\text{error}_{\text{prediction}} = \max \left\{ \begin{array}{l} \frac{\text{predicted}}{\text{real}} \\ \frac{\text{real}}{\text{predicted}} \end{array} \right. \quad (2)$$

Additionally, for the candidate structures assigned by *in silico* models in **paper III**, the root mean square difference (RMSD) was calculated for MS<sup>2</sup> and candidate structure-based predictions.

The validation of priority score approach using binary classification in **paper III** was evaluated based on precision (Prec), recall (Rec), and accuracy (Acc) calculated as follows:

$$\text{Prec} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

The metrics are calculated using the number of predicted true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

## 4. Results and discussion

### 4.1. Risk estimation trends in non-targeted screening (paper I)

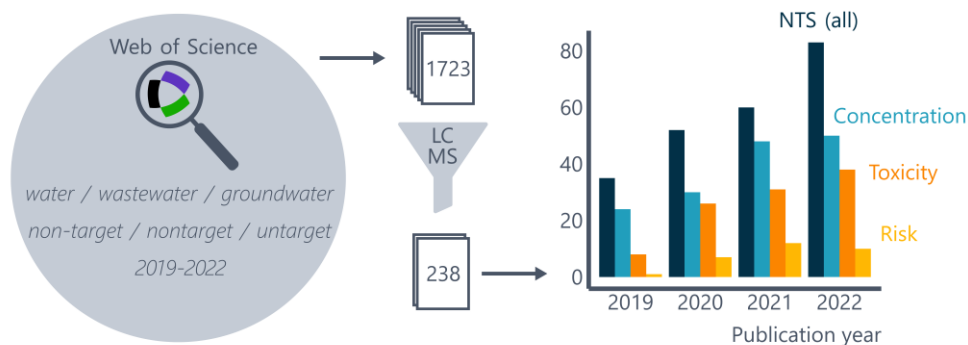


Figure 4. Graphical overview of paper I.

Suspect and non-targeted screening are needed to discover chemicals of emerging concern that are overlooked by targeted methods. Often, new chemicals are detected in the samples, yet the potential risk of these chemicals is unclear from the analysis. Simultaneously, new tools and workflows are emerging that can help to provide such information for detected chemicals. Thus, this paper aimed to review the trends over the years 2019 to 2022 in considering concentration, toxicity, or the risk, i.e., the combination of the two, estimations for the water contaminants detected in NTS.

Firstly, a growing trend in papers applying non-target screening was observed: in total, 35, 52, 60, and 83 papers applied NTS on water samples in 2019, 2020, 2021, and 2022, respectively. Over all 238 relevant papers, 158 (66%) included quantification, and 108 (45%) included toxicity in the analysis of the detected chemicals.

The vast majority of the quantification was performed with an analytical standard, either for predefined target chemicals or by acquiring a standard for the identified feature(s). In 40 papers, alternative standard-free methods were considered for quantification, most commonly, a surrogate standard based on structural similarity

(including parent-TP relation) or retention time. Alternatively, an isotope-labelled standard was used for quantification.

Toxicity estimation in environmental water monitoring gained popularity in this period. In 2019, every fourth paper considered toxicity estimation, while in 2022, it had increased to every second paper. Most commonly, databases and literature searches were used to find toxicity information for detected (identified) chemicals. In addition, approximately every third paper included *in silico* toxicity predictions for the identified features in the analysis. Experimental toxicity, mostly with *in vitro* bioassays, was also performed in every third paper for effect-based monitoring of the whole sample or effect-directed analysis.

The risk was considered in 30 papers out of 238. In 19 papers, risk quotients were calculated for detected features based on the environmental concentration level obtained through target quantification, and literature-derived or *in silico* predicted toxicity values.

While the number of applicable tools for risk estimation has increased, the majority of the research is still either qualitative or highly dependent on the availability of analytical standards for risk estimation. The deeper investigation, validation, and performance comparison of recently developed risk estimation tools could increase their usage and thereby increase the overview of the risk of the detected features. Additionally, as the majority of the tools used in the literature depend on structural assignment, emerging MS<sup>2</sup>-based tools could be helpful for pinpointing high-risk chemicals.

## 4.2. Development of MS<sup>2</sup>-based quantification model (paper II)

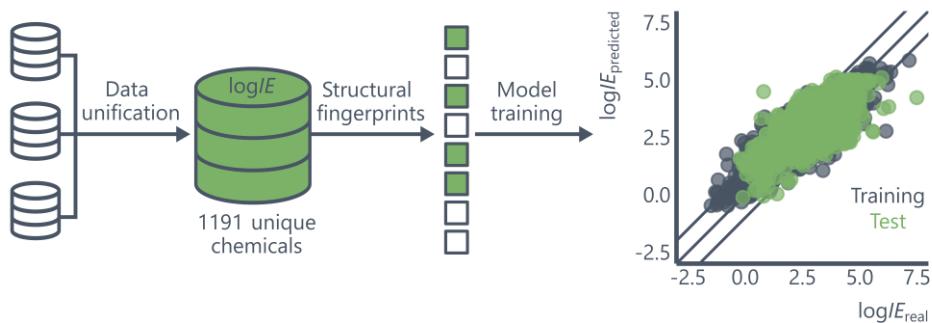


Figure 5. Graphical overview of paper II.

The aim of **paper II** was to develop a machine learning model that can be used to quantify detected features, either based on the structural fingerprints calculated from the assigned structure or predicted from the fragmentation spectrum. The latter can be done by using SIRIUS+CSI:FingerID, which can predict a probabilistic fingerprint of a detected chemical based on its MS<sup>2</sup> spectrum.

### Unification

The calibration graph slopes are specific to the used instrument and experimental conditions; thus, to compile a dataset of chemicals on a relative ionization efficiency scale, the calibration graph slopes from different sources were unified. For this, the dataset from Liigand et al. (hereafter referred to as “dataset 1”) was used as a basis since this dataset contained relative ionization efficiency values for 399 unique chemicals measured under different experimental conditions (Figure 6).

As the slope and *IE* values can differ multiple orders of magnitude, the data were log-transformed. In order to unify a new experimental dataset, chemicals measured under similar experimental conditions were used to construct a linear regression model between dataset 1 and a new dataset. The linear model was then used to transfer all calibration graph slope values from the new dataset to the *IE* scale. In total, 13 datasets were unified, resulting in 6049 datapoints corresponding to 1191 unique chemicals.

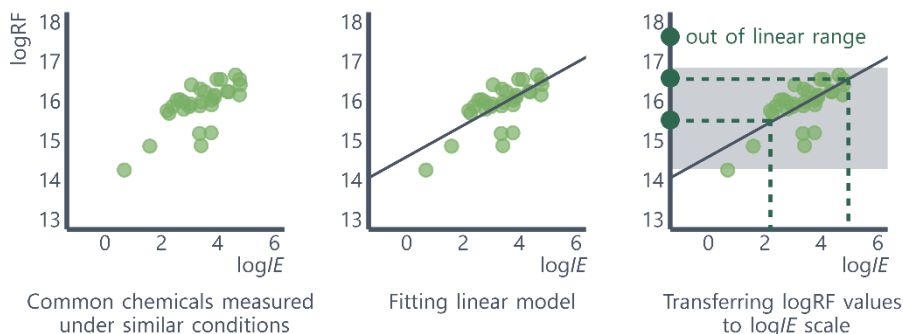


Figure 6. Unification process for obtaining  $\log IE$  data for modelling. First, common chemicals  $\log IE$  dataset and new dataset subject to unification were selected and filtered based on experimental conditions (such as pH and organic modifier percentage during elution). These chemicals were used for fitting a linear regression and transferring the RF values from the dataset subject to unification to the  $\log IE$  scale. Chemicals outside of the linear range were discarded.

### Model training

The unified dataset was divided into a training and test set based on unique chemicals with an 80/20 ratio. Structural fingerprints were calculated based on SMILES notation for both training and test data. The training dataset was cleaned by removing missing values and highly correlated or low-variance features. The remaining features and eluent composition features (surface tension, viscosity, polarity,  $\text{NH}_4^+$  presence) were used for model training. Out of the tested extreme gradient boosting algorithms, `xgbTree` outperformed the other tested algorithms. Complementary structure-dependent models based on different molecular descriptors were trained, with the Pharmaceutical Data Exploration Laboratory (PaDEL) descriptor-based model outperforming others. Thus, the PaDEL-based model was used in the performance comparison with the structural fingerprints-based model, further referred to as MS2Quant.

Based on the MS2Quant predicted  $IE$  values, the root mean squared error (RMSE) of 0.55 (3.5 $\times$ ) and 0.80 (6.3 $\times$ )  $\log$ -units were observed on the training and test set, respectively. In comparison, the RMSE of the  $IE$  values predicted by the PaDEL-based model were 0.56 (3.6 $\times$ ) and 0.81 (6.5 $\times$ )  $\log$ -units for the training and test set, respectively, indicating similar performance of structural fingerprints and molecular descriptors-based models. Here, all predictions were performed based on the structure. For the validation of the MS2Quant model on  $\text{MS}^2$ -based  $IE$  predictions, new

MS2Quant and PaDEL-based models on all available data (training + test set) were trained using the optimized hyperparameters.

### Validation

The samples from the NORMAN interlaboratory comparison of quantification approaches in NTS were used for validation.<sup>104</sup> The concentration was predicted for 39 suspect chemicals spiked into lake water at two concentration levels. In order to convert the predicted *IE* values to experiment-specific RFs, the linear regression between experimental calibration graph slopes and predicted *IE* values of 36 calibrants was used. The *IE* values predicted based on the MS<sup>2</sup> spectra of the suspects were then converted into predicted RFs using the *IE*-RF linear regression model of calibrants. The concentration could then be predicted by dividing the peak areas of the suspects by their corresponding predicted RFs.

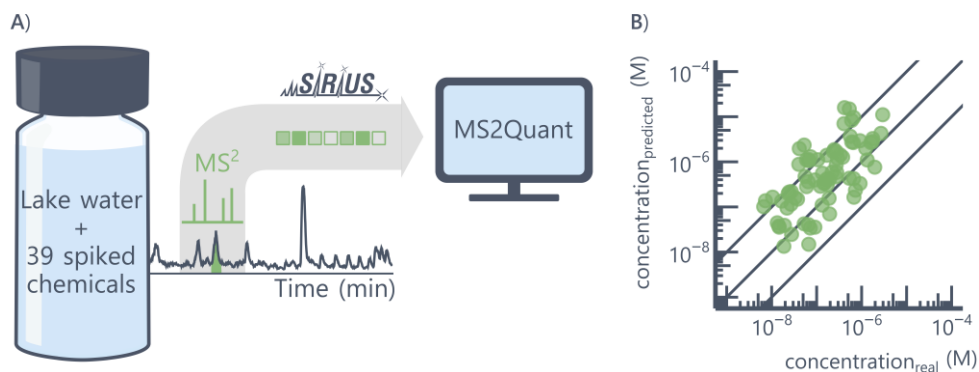


Figure 7. Samples from the NORMAN interlaboratory comparison were used for model validation. A) In total, 39 suspects spiked at two concentration levels were quantified based on experimental MS<sup>2</sup> spectra. For this, structural fingerprints were calculated using SIRIUS+CSI:FingerID software and were used as input to MS2Quant. B) The predicted concentrations showed good correlation with spiked concentrations ( $R^2$  of 0.46, RMSE of 0.77 log-units).

The observed RMSE for the 39 suspect chemicals was 0.77 log-units (5.9 $\times$ ), which was comparable with the error observed on the test set (Figure 7). For the structure-based predictions comparison, highest ranked structure assigned by SIRIUS+CSI:FingerID software was used. The structure-based concentration predictions resulted in RMSE of 0.86 (7.3 $\times$ ) and 0.87 (7.4 $\times$ ) log-units with MS2Quant and PaDEL-based model, respectively. However, the differences in quantification errors with MS<sup>2</sup>-based MS2Quant and PaDEL-based model were statistically insignificant (pair-wise Wilcoxon ranked sum exact test,  $p = 0.13$ ). In the case of the 39 suspects, 34 were

correctly identified as highest ranked structure. The prediction fold error on the five incorrectly identified features was 4.2× and 6.0× for MS<sup>2</sup>-based MS2Quant and PaDEL-based model, respectively, though the differences were statistically insignificant (pair-wise Wilcoxon rank sum exact test,  $p = 0.44$ ). The MS<sup>2</sup>-based predictions were comparable to structure-based predictions, suggesting that MS<sup>2</sup> does not have a large impact on the predictions and most of the uncertainty originates from the model.

The developed concentration prediction model MS2Quant enables high-throughput concentration estimations without the need for structural assignment. The main limitation in concentration prediction is the availability of the probabilistic fingerprints predicted by SIRIUS+CSI:FingerID.

### 4.3. MS<sup>2</sup>-based prioritization of wastewater features (paper III)

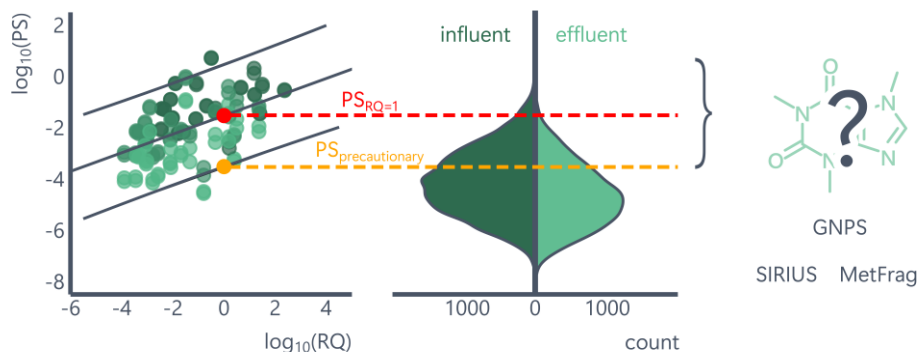


Figure 8. Graphical overview of paper III.

The discovery of emerging contaminants posing a high risk mostly relies on the often inaccurate and low-yield structural identification process. Thus, this paper aimed to investigate the applicability of the two MS<sup>2</sup>-based concentration and toxicity prediction models, MS2Quant and MS2Tox, for risk-based feature prioritization. For this, the MS<sup>2</sup>-based predictions were combined into a priority score (PS):

$$PS = \frac{\text{Area} / 10^{\log IE_{\text{predicted}}}}{10^{\log LC_{50 \text{ predicted}}}} \quad (6)$$

The Area corresponds to the integrated signal area from full scan measurements,  $\log IE_{\text{predicted}}$  corresponds to log-scale  $IE$  predicted with MS2Quant, and  $\log LC_{50 \text{ predicted}}$  corresponds to lethal concentration to 50% of the adult fish (log-mM) predicted with MS2Tox.

The applicability of the MS<sup>2</sup>-based priority score was first validated on a mixture of 159 contaminants spiked in influent and effluent samples of one WWTP, and applied to influent and effluent samples from three WWTPs in Sweden. Additionally, the effect of the data acquisition method used for collecting fragmentation spectra was investigated.

#### *Evaluation of the priority score*

The manual investigation of full scan measurements resulted in 90 mutually detected chemicals that were used in the further investigation. Three different data acquisition

approaches for acquiring MS<sup>2</sup> spectra were tested: (1) top 5 highest intensity features per full scan (Top5); (2) a target inclusion list compiled based on a surface water suspect list (SWS); and (3) data-independent acquisition (DIA). The number of acquired MS<sup>2</sup>, successfully predicted probabilistic structural fingerprints, and the correct formula and ranking of the correct structure were investigated.

Obtaining the predicted fingerprint with SIRIUS+CSI:FingerID primarily depended on the concentration of the spiked contaminants and not on the data acquisition method. However, the correctness of the predicted formula and candidate structure was higher for both DDA methods compared to DIA, suggesting lower spectral quality for the deconvoluted fragmentation spectra from DIA measurements. (Table 1)

*Table 1. The percentage of correctly predicted molecular formulas and structural assignments, calculated based on the number of chemicals with predicted structural fingerprints for the three investigated MS<sup>2</sup> acquisition.*

MS <sup>2</sup> acquisition method	Correct formula (%)	Correct structure among candidates (%)	Correct structure as rank 1 (%)
Top5	51-71%	72-86%	59-77%
SWS	33-77%	56-100%	56-83%
DIA	11-38%	49-65%	21-43%

For all features with predicted probabilistic fingerprint, PS was predicted from the fingerprint and from the SMILES notation of the chemical and compared. Over influent and effluent samples spiked at three concentration levels, the best agreement between MS<sup>2</sup> and SMILES-based PS was observed for Top5 and worst for DIA acquisition method (Figure 9Figure 1). Interestingly, the features with predicted probabilistic fingerprint but without any candidate structure suggestions showed good agreement with the SMILES-based PS ( $R^2$  of 0.18 to 0.72). Approximately 20% of the features with predicted fingerprints yielded no candidate structures, thereby showcasing the advantage of using MS<sup>2</sup>-based methods for risk predictions.

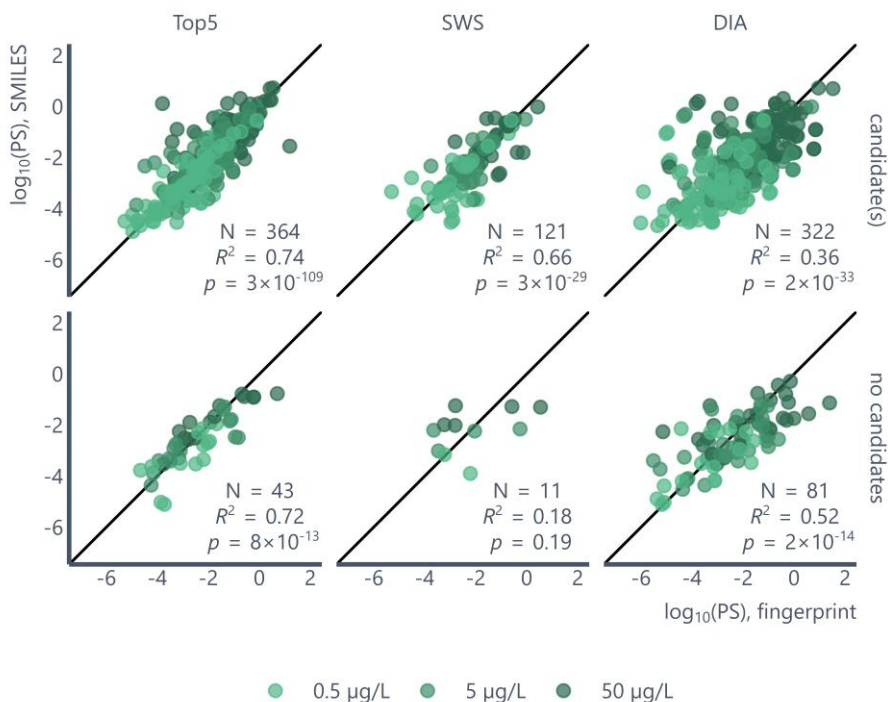


Figure 9. The priority score (PS) predicted based on correct SMILES compared to the PS predicted based on probabilistic fingerprint from MS<sup>2</sup> for three acquisition methods. The data is categorized based on the availability of candidate structures assigned by SIRIUS+CSI:FingerID.

### Risk quotient equivalent priority scores

The developed PS was compared to the risk quotient (RQ) to investigate its applicability in the environmental context. The experimental LC<sub>50</sub> values from three fish species (bluegill, fathead minnow, rainbow trout) were available for 23 spiked chemicals out of 90; thus, the RQs were calculated by dividing the spiked concentration by the experimental LC<sub>50</sub> value for the influent and effluent samples spiked at three concentration levels. The PS corresponding to RQ=1, based on the linear fit (PS<sub>RQ=1</sub>) and PS corresponding to the lower limit of the 95% prediction interval at RQ=1 (PS<sub>precautionary</sub>), were defined using SMILES-based PS comparison with RQs (Figure 10A).

In order to validate the prioritization approach, binary classification was used on the features with predicted probabilistic fingerprints. Chemicals with RQ>1 were classified as “ground truth” high-risk chemicals, and the chemicals that exceeded PS<sub>RQ=1</sub>

were classified as predicted high-risk chemicals. Similarly, features exceeding  $PS_{\text{precautionary}}$  were predicted as precautionary risk features.

More than half of the high-risk features ( $\geq PS_{RQ=1}$ ) were prioritized by Top5 and DIA, resulting in respective recall values of 0.55-0.68 and 0.55-0.74, over different fish species. Using the  $PS_{\text{precautionary}}$ , the recall values were 0.97-1.00 and 0.94-1.00 for Top5 and DIA, respectively (Figure 10B). The correlations between PS and RQ were statistically insignificant for all SWS measurements and for all correlations with  $LC_{50}$  derived from fathead minnow fish species. Additionally, as the SWS approach did not acquire data for contaminants exceeding  $RQ=1$ , the calculated binary evaluation metrics were uninformative. The recall values from Top5 and DIA measurements from realistic exposure scenarios indicate that  $PS_{RQ=1}$  and  $PS_{\text{precautionary}}$  capture the majority of the high-risk features and can substantially reduce the feature set.

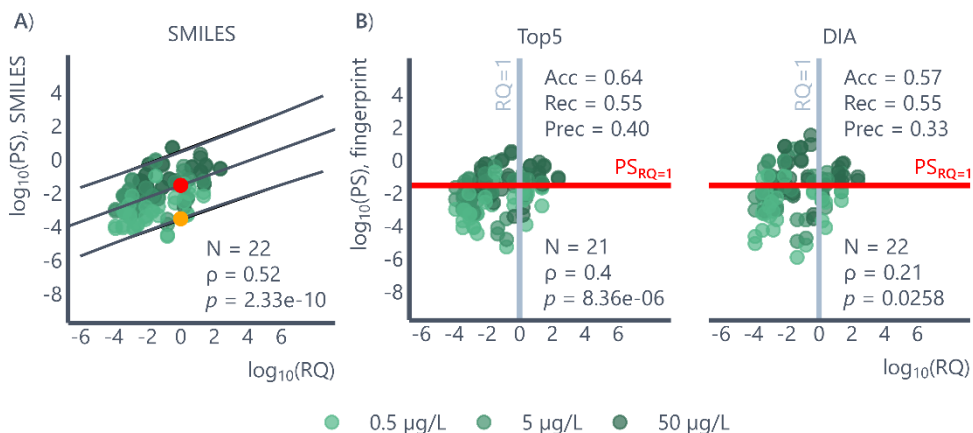


Figure 10. A) The  $PS_{RQ=1}$  and  $PS_{\text{precautionary}}$  are defined based on the SMILES-based priority score (PS) and risk quotient (RQ) correlation. B) Binary classification using  $PS_{RQ=1}$  for Top5 and DIA acquisition methods, as no features surpassed  $RQ=1$  for the SWS method. Here,  $LC_{50}$  values from one fish species (bluegill) are shown.

### Applying the priority score approach to wastewater samples

The  $PS_{RQ=1}$  and  $PS_{\text{precautionary}}$  were applied to the influent and effluent samples from three WWTPs in Sweden. In total, 9,751 to 15,923 features were detected per WWTP. Features detected in full scan were joined with the respective  $MS^2$  information from the three data acquisition methods based on retention time ( $< 0.2$  min) and  $m/z$  ( $< 10$  ppm) agreement. For 45% to 51% of the features per WWTP, at least one  $MS^2$  spectrum was assigned from the applied data acquisition methods. The probabilistic

fingerprints were successfully predicted for 81% to 86% of the features with MS<sup>2</sup> data, enabling further risk predictions for 38% to 43% of all detected WWTP features.

For influent samples, 32% to 35% of the features with predicted risk exceeded PS<sub>precautionary</sub>, and 2% to 3% exceeded PS<sub>RQ=1</sub>. Similarly, 20% to 27% of the effluent features exceeded PS<sub>precautionary</sub>, and approximately 1% exceeded PS<sub>RQ=1</sub>. Thus, depending on the prioritization approach, the feature set could be reduced by 65% to 99%.

All features in influent and effluent samples over the three WWTPs exceeding PS<sub>precautionary</sub> were subject to library matching with GNPS. A library match was found for 557 (11%) features out of 4,963. As the effluents are released to the environment, the focus further shifted to effluent features exceeding PS<sub>RQ=1</sub> and PS<sub>precautionary</sub>. On the example of WWTP2, a library match was found for 67 features out of 775 exceeding PS<sub>precautionary</sub>. Comparing the SMILES-based PS to the MS<sup>2</sup>-based PS for the library matched features, the MS<sup>2</sup>-based risk was generally slightly higher, which is suitable for the precautionary approach, and the majority of the predictions (81%) were within a 10× difference (RMSE of 0.79, R<sup>2</sup> of 0.28, Figure 11A). In total, 29 features detected in the effluent sample of WWTP2 exceeded PS<sub>RQ=1</sub> and were subject to *in silico* matching with MetFrag and SIRIUS+CSI:FingerID. While plausible candidates were suggested within the top 10 candidate structures, the two *in silico* tools rarely agreed. Additionally, the PSs predicted for the candidate structures varied widely for both methods, and the evidence was insufficient for confident annotation (level 2 or 1) of any feature with the *in silico* tools (Figure 11B).

Based on the analysis of the spiked contaminants and library-matched features, the MS<sup>2</sup>-based PS approach can be useful for prioritizing high risk features. However, assigning one probable candidate structure for the high-risk features prioritized in the effluent samples remains challenging. The ambiguity of structural annotation using *in silico* tools, which yielded candidates with a wide range of structure-based predicted priority scores, further highlights the importance of using experimental data in prioritization.

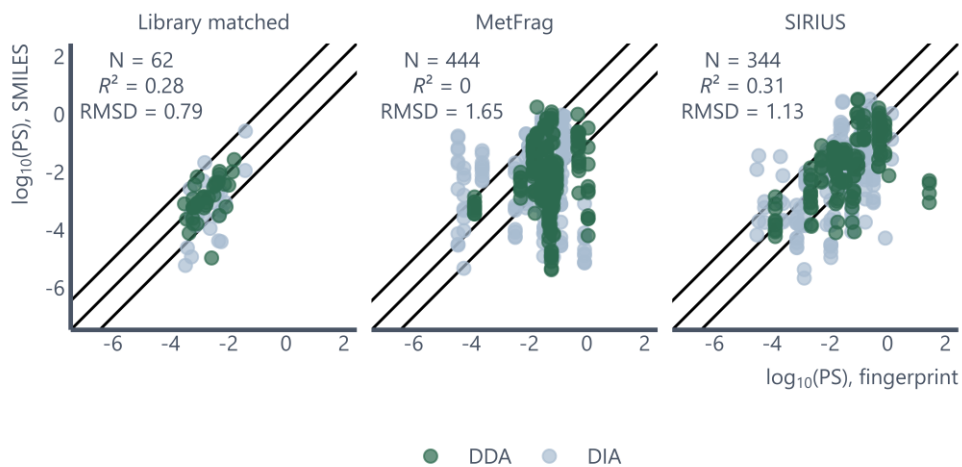


Figure 11. Correlation between  $MS^2$ -based probabilistic fingerprint priority score (PS) compared to the PS of A) library matched structures, as well as from B) MetFrag and C) SIRIUS+CSI:FingerID in silico assigned candidates.

#### 4.4. Experimental validation of MS<sup>2</sup>-based predictions (paper IV)

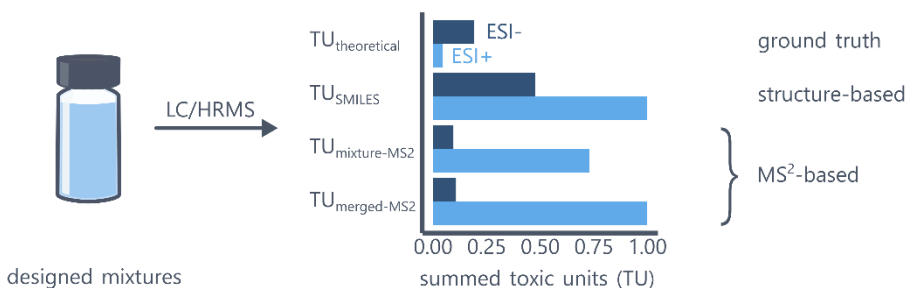


Figure 12. Graphical overview of paper IV.

The developed MS<sup>2</sup>-based *IE* and LC<sub>50</sub> prediction models can be applied for prioritization as demonstrated in paper III. In addition, these models provide estimates that can be combined into toxic units (TU), representing the contribution of individual chemicals to mixture toxicity. TUs are calculated similarly to RQs as a ratio of concentration to a toxicity value (Eq. 1); however, TUs are mainly used in the context of mixture toxicity. To evaluate the accuracy of predicted TUs and the cumulative mixture toxicity, designed mixtures were prepared and analyzed with LC/HRMS. The MS2Quant model was applied for predicting concentration. Furthermore, to enable experimental validation of predicted toxicity values, a complementary fish embryo acute toxicity (FET) model was trained and applied to the LC/HRMS data of the mixtures.

##### *Designed mixtures*

In total, 23 chemicals were used in designed mixtures. These chemicals were selected to cover a wide range of log<sub>K<sub>ow</sub></sub> and LC<sub>50</sub> values, and be detectable in positive and/or negative mode electrospray ionization. The mixtures were designed to represent different mixture scenarios, including mixture toxicity driven by one chemical as well as an equipotent mixture.

Out of 23 chemicals, 16 were detected in the positive ESI mode and 12 in the negative ESI mode. Seven chemicals were detected in both modes, while two chemicals were not detected in either mode. In addition to the MS<sup>2</sup> spectra acquired for each mixture separately (further referred to as “mixture-specific spectra”), the MS<sup>2</sup> spectra were acquired over a range of collision energy values (CE 10 to 90 eV with a step of 10 eV). These spectra were merged, resulting in one spectrum per chemical and ESI mode and

are further referred to as “merged spectra”. The spectral quality was investigated based on the agreement of the correct molecular formula suggested by SIRIUS+CSI:FingerID. The merged MS<sup>2</sup> spectra improved correct formula prediction compared to the mixture-specific MS<sup>2</sup> spectra.

### Concentration and toxicity predictions

The concentrations and toxicity were predicted based on SMILES notation, mixture-specific MS<sup>2</sup> spectra, and merged MS<sup>2</sup> spectra. As expected, the highest agreement with the spiked concentrations was observed for the concentration predictions from SMILES (RMSE of 0.54 and 0.69, for ESI+ and ESI-, respectively; Figure 13); however, the differences between structure and MS<sup>2</sup>-based predictions were small, and over 80% of the predicted concentrations were within 10× error, indicating reliable concentration predictions.

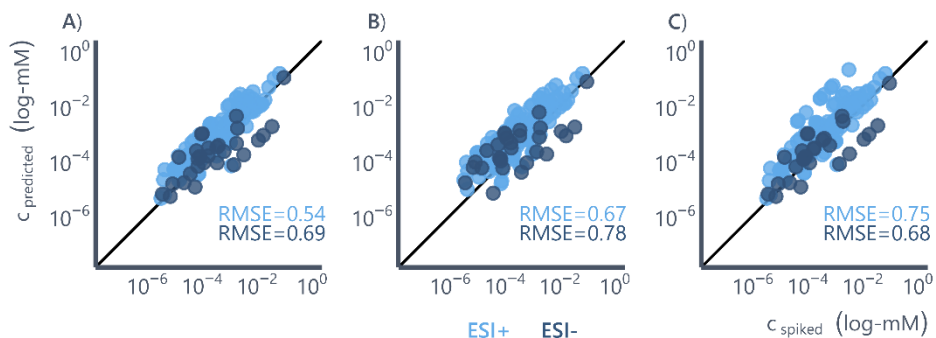


Figure 13. The spiked concentration comparison to the concentration predicted in ESI+ and ESI- mode based on A) SMILES notation of the chemical, B) mixture-specific MS<sup>2</sup> spectrum, and C) merged MS<sup>2</sup> spectrum.

The LC<sub>50</sub> values of the chemicals in the designed mixtures were obtained under different experimental conditions. Thus, similar to the training data, data unification was needed to transfer the LC<sub>50</sub> values to experimental conditions corresponding to the test guidelines. Additionally, the FET LC<sub>50</sub> was experimentally determined for eight chemicals in this work and used as ground truth values instead of the dataset values. In total, LC<sub>50</sub> values were unified for 20 chemicals out of the 23. The three chemicals with LC<sub>50</sub> values measured under different experimental conditions that did not allow unification were excluded from all statistical analyses.

The mixture-specific MS<sup>2</sup> yielded multiple predictions per chemical and mode, which were averaged prior to calculating  $R^2$  and RMSE. Compared to concentration

predictions, higher prediction errors were observed for predicted  $LC_{50}$  values, as the RMSE above 0.9 log-mM for both structure and  $MS^2$ -based methods in both electrospray modes (Figure 14) was observed. The lowest prediction errors were observed for merged spectra (RMSE of 0.90 and 1.07 log-mM, for ESI+ and ESI-, respectively) and the  $LC_{50}$  predictions for positive ESI mode always outperformed negative ESI mode. This was also observed for SMILES-based predictions, which are unaffected by the ESI mode, although the differences may be driven by the small chemical set size.

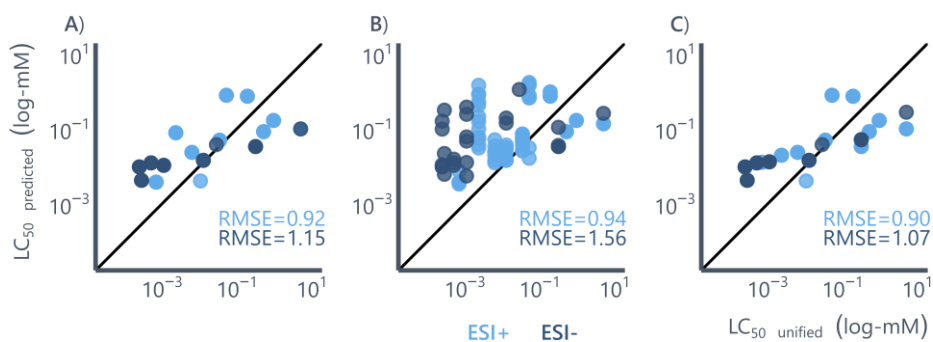


Figure 14. FET  $LC_{50}$  values compared to the  $LC_{50}$  values predicted in ESI+ and ESI- mode based on A) SMILES notation of the chemical, B) mixture-specific  $MS^2$  spectrum, and C) merged  $MS^2$  spectrum. While SMILES-based and merged spectra-based approaches yield one prediction per chemical and mode, the mixture-specific  $MS^2$ -based approach includes multiple spectra from different measurements and therefore results in more predictions.

### Toxic unit predictions

Toxic units were predicted to evaluate how well the developed models capture the contribution of individual chemicals to mixture toxicity based on SMILES and  $MS^2$  data. The prediction fold errors for chemicals were estimated using the ratio of predicted TU over theoretical TU. Considering that uncertainty factors of 10 $\times$  or 100 $\times$  are often used in risk assessment, predictions within a 10 $\times$  error were considered acceptable. The prediction error was within 10 $\times$  in at least one ESI mode for 60-70% of the chemicals over the three approaches (Figure 15). The geometric mean of fold errors ranged from 6 $\times$  to 8 $\times$  across all modes and approaches, except for mixture-specific  $MS^2$  in negative ESI mode, where a geometric mean of 21 $\times$  was observed. Generally, lower prediction errors were observed for merged  $MS^2$ , confirming that the spectral quality affects the prediction accuracy. Still, both structure and  $MS^2$ -based methods resulted in a similar magnitude of prediction errors, indicating that there is room for improvement in the mixture toxicity prediction models.

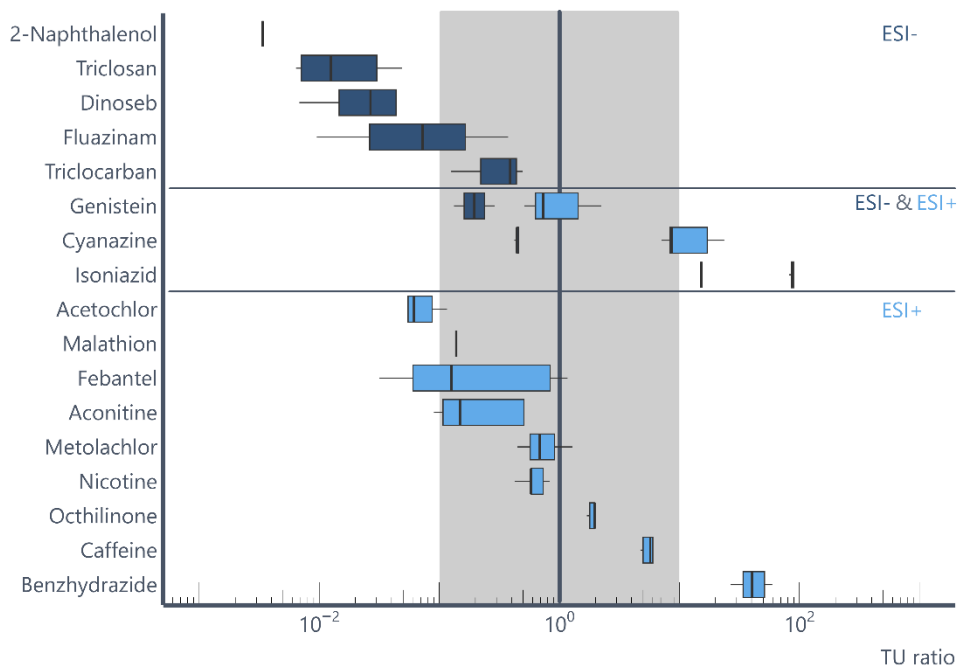


Figure 15. Theoretical toxic units (TU) compared to the predicted toxic units predicted based on mixture-specific  $MS^2$  spectra. The perfect alignment (dark grey vertical line) and 10 $\times$  error (light grey box) are plotted and chemicals detected in ESI-, ESI+, or both modes are divided by horizontal lines.

Additionally, predicted cumulative mixture toxicity was investigated based on cumulative TU ratios for predicted and theoretical cumulative TUs. The predicted cumulative mixture toxicity values agreed well with theoretical values over all approaches, as the geometric mean of fold error ranged from 2 $\times$  to 3 $\times$ , and 77% to 97% of the predicted cumulative TUs were within 10 $\times$  error, depending on the method. These results indicate that the predicted toxic units capture general risk trends and can be useful for prioritization.

Larger prediction errors were observed for the FET  $LC_{50}$  prediction model, which may be attributed to the intrinsic biological variability in  $LC_{50}$  measurements and the complexity of the underlying mechanism, a slightly smaller training dataset compared to the *IE* dataset, or the limitations of the structural fingerprints in representing the structural features relevant to toxicity. While caution should be taken when interpreting the cumulative mixture toxicity predictions, the first experimental validation shows promising results and provides a basis for further development and validation in future studies.

## 5. Conclusions and future perspectives

Non-targeted screening (NTS) detects thousands of emerging environmental contaminants. In order to focus on the chemicals posing the highest risk, feature prioritization plays a crucial role in NTS. This thesis aimed to develop and validate prioritization models that rely on mass spectrometric data instead of structural assignment and thus, increase the subset of detected features considered in risk-based prioritization.

As seen in the NTS trend analysis in **paper I**, even when some emerging contaminants are identified in NTS, the subsequent risk estimation of these features is often lacking. At the same time, many structure-based prediction tools for identified features have emerged that do not require matching reference standards, enabling high-throughput estimation of concentration and toxicity. More recently, MS<sup>2</sup>-based prediction models have emerged, mostly for predicting toxicity of the unidentified features. Systematic evaluation of model predictions against benchmark datasets can help to reveal the strengths and limitations of the developed tools and support their application.

The findings of this thesis show that mass spectrometric data can be used to predict exposure levels of unidentified chemicals. While concentration prediction models that do not rely on reference standards have been developed and used before, **paper II** showcases for the first time that the concentration can also be predicted based on MS<sup>2</sup> spectra without structural assignment. Furthermore, similar accuracy to structure-based predictions can be achieved. For this purpose, a model named MS2Quant was developed to predict ionization efficiency of a chemical and convert the peak area to concentration based on this prediction.

Combining MS2Quant with the complementary toxicity prediction model MS2Tox was useful for feature prioritization. On the example of wastewater influent and effluent samples in **paper III**, the structure-based priority score correlated well with the MS<sup>2</sup>-based priority score, especially for spectra measured in data-dependent mode. Furthermore, the MS<sup>2</sup>-based methods covered up to 20% more of the detected LC/HRMS features. In the effort to identify features exceeding priority score equivalent to RQ=1, *in silico* tools SIRIUS+CSI:FingerID and MetFrag were used. However, the suggested candidate structures rarely agreed, and the structure-based priority scores had high variation for both methods, highlighting the advantage of MS<sup>2</sup>-based prioritization methods.

The MS<sup>2</sup>-based concentration and toxicity predictions can also be combined into toxic units, which show the chemicals' contribution to the overall cumulative toxicity. The cumulative toxic units can be used to estimate the overall mixture toxicity, as the concentration addition model is considered a suitable approximation for complex mixtures of low concentration contaminants. To enable experimental validation of the predicted toxic units, a complementary FET LC<sub>50</sub> model was developed in **paper IV**. The error in toxic unit predictions was primarily driven by the toxicity predictions, with root mean square errors reaching 9× in positive and 36× in negative electrospray ionization mode. Although the models need further improvement to accurately predict toxic units for chemicals, the predicted toxic units show potential for prioritization.

The observed model errors highlight the difficulty of the toxicity modelling. There could be many explanations for this; for example, the LC<sub>50</sub> values have an intrinsic experimental variability due to differences between organisms. Also, the training data gathered for chemicals over multiple decades might not sufficiently represent emerging contaminants. Additionally, in real-life applications, the single chemical toxicity data might not translate to the same behaviour in a complex mixture. Further investigations are needed to understand how and to what extent the toxicity models can be improved.

In this thesis, predicted lethal concentrations for adult fish and fish embryos were applied. These endpoints were investigated as proof of concept and were considered applicable in prioritization as they provide a continuous value that enables comparison of two chemicals based on their toxicity. The further development of this work could expand to additional endpoints, including sublethal effects, which often occur at lower exposure levels. Regardless of endpoint, including more high-quality data and the mechanistic understanding of toxic effects could improve the toxicity predictions.

Many chemicals remain undetected in NTS depending on the used instrumentation. The subset of detected chemicals highly depends on the experimental conditions and instrumental limitations. For example, the concentrations that cause an effect in living organisms may be lower than the limit of detection of the analytical instruments. Additionally, some of the relevant structural information cannot be acquired with MS. For example, MS cannot capture information about, e.g., chirality, which, due to differences in metabolic, pharmacokinetic, and toxicological profiles, can change the toxicity of a chemical.

The analysis of detected features can be complicated by the overlapping features, in-source fragments, and adducts. As the models used in this thesis all rely on the SIRIUS+CSI:FingerID for extracting structural information from MS<sup>2</sup>, the uncertainties in structural fingerprint prediction further contribute to the model prediction error. In addition to input data quality, a further limitation in model performance is related to training data, which can have high variation or insufficient chemical coverage. Furthermore, the selected modelling approach, including the selection of features such as molecular descriptors or structural fingerprints, affects the performance if the training data is insufficiently described for the task.

For the models to be practically useful, the uncertainties in predictions need to be evaluated. Therefore, the experimental validation of the developed models is essential to both confirm the performance and provide an overview of the expected prediction errors in real-life applications.

## Acknowledgements

Embarking on the journey of a PhD, I knew I would learn much from working with you, Anneli. In hindsight, I am grateful for so much more than just the knowledge you have shared. You have been a supportive mentor and a true role model, you have cheered me on through the good times, supported me through the lows, and always encouraged me to take that extra step forward. Mul on nii hea meel, et meie teed ristusid ning ma olen väga tänulik, et olen saanud sinuga koos töötada ja teadlasena kasvada.

I am also sincerely grateful to my co-supervisors Matthew, Jonathan M., and Magnus. Thank you for your enthusiasm, guidance, and support throughout my studies.

During my studies, I have had the opportunity to work together with many kind and brilliant people. I am thankful to all my co-authors across different projects. Special thanks to Mélanie, Jonathan B. and Malte. Mélanie, thank you for being both a great mentor and a dear friend. Thank you, Frederic and Lisa, for hosting and supporting me during my research exchange in Amsterdam, and for your constant encouragement and valuable feedback. I am also grateful to Erica and Ulrika for reviewing my thesis and providing thoughtful feedback.

I have been incredibly fortunate to be surrounded by such wonderful people in the fellowship of Kruvelab. To the postdocs I have looked up to: Drew, Eliise, Frank, Henrik, Ida, Pilleriin – thank you for your support, kindness, and for being such inspiring role models. I am grateful to my fellow PhD students who have walked the same path alongside me (even if we reach Mordor at different times): Amina, Harry, Iris, Louise, Lucas, Ruchun, Yvonne, Gordian, and our visiting PhD students Lisa and Nienke. I am also thankful to all Bachelor's and Master's students who have been working in our group. Thank you for the fun afterwork activities, inspiring discussions, ice cream breaks and fikas, as well as for making even the longest days in the lab enjoyable. Most of all, thank you for never running out of supportive and encouraging words.

I am also thankful for my wonderful colleagues at the Analytical Chemistry and Contaminant Chemistry units. Thank you, Anna, Adam, Awat, Alex, Claudia, Eleni, Ena, Georgios, Ian, Ioannis, Josefine, Leopold, Lorena, Merle, Michael, Nicole, Nicolina, Ulrika, Pablo, Pablo, Romain, Tim, and Vendela. Ulrika, thank you for your

kindness and support throughout my studies. Thank you, Claudia, for your support both inside and outside the lab. Special thanks to Elin and Karin, for helping me navigate all the paperwork.

Finally, I cannot thank enough the most important pillars in my life – my friends and family. Mu kallid ema, isa, õeraas Ly, ja kogu laiendatud pere – te olete mu suurim õnn. Aitäh, et saan teie peale alati loota ning et olete mulle alati olemas. Tack, Anton, för att du alltid finns där för mig, och tack till din familj för att ni har tagit in mig som en del av er. To my dear friends, especially Eliise, Anu, and Miriam – thank you for your constant support. And of course my little pup Gimli – thank you for always being excited to see me, even on days when it felt like Reviewer 2 strongly disagreed, for making sure I take enough walk breaks with you, and for reminding me that there is more to life than deadlines and perfect presentations.

## References

1. Wang, Z., Walker, G. W., Muir, D. C. G. & Nagatani-Yoshida, K. Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. *Environ. Sci. Technol.* **54**, 2575–2584 (2020).
2. Richardson, S. D. & Kimura, S. Y. Emerging environmental contaminants: Challenges facing our next generation and potential engineering solutions. *Environ. Technol. Innov.* **8**, 40–56 (2017).
3. Naidu, R. *et al.* Chemical pollution: A growing peril and potential catastrophic risk to humanity. *Environ. Int.* **156**, 106616 (2021).
4. Richardson, S. D. & Manasfi, T. Water Analysis: Emerging Contaminants and Current Issues. *Anal. Chem.* **96**, 8184–8219 (2024).
5. Hu, D., Henderson, K. & Coats, J. Fate of Transformation Products of Synthetic Chemicals. in *Transformation Products of Synthetic Chemicals in the Environment* (ed. Boxall, A. B. A.) vol. 2P 103–120 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2009).
6. Wall, J. T. *et al.* Development of the toxicity values database, ToxValDB: A curated resource for experimental and derived human health-relevant toxicity data. *Comput. Toxicol.* **35**, 100365 (2025).
7. Rockström, J. *et al.* A safe operating space for humanity. *Nature* **461**, 472–475 (2009).
8. MacLeod, M. *et al.* Identifying Chemicals That Are Planetary Boundary Threats. *Environ. Sci. Technol.* **48**, 11057–11063 (2014).
9. Persson, L. *et al.* Outside the Safe Operating Space of the Planetary Boundary for Novel Entities. *Environ. Sci. Technol.* **56**, 1510–1521 (2022).
10. Cousins, I. T., Johansson, J. H., Salter, M. E., Sha, B. & Scheringer, M. Outside the Safe Operating Space of a New Planetary Boundary for Per- and Polyfluoroalkyl Substances (PFAS). *Environ. Sci. Technol.* **56**, 11172–11179 (2022).
11. Landrigan, P. J. *et al.* The Lancet Commission on pollution and health. *The Lancet* **391**, 462–512 (2018).
12. Hou, L. *et al.* Pesticide Pollution Reduces the Functional Diversity of Macroinvertebrates in Urban Aquatic Ecosystems. *Environ. Sci. Technol.* **59**, 8568–8577 (2025).
13. Sigmund, G. *et al.* Addressing chemical pollution in biodiversity research. *Glob. Change Biol.* **29**, 3240–3255 (2023).

14. Groh, K., vom Berg, C., Schirmer, K. & Tlili, A. Anthropogenic Chemicals As Underestimated Drivers of Biodiversity Loss: Scientific and Societal Implications. *Environ. Sci. Technol.* **56**, 707–710 (2022).
15. Grandjean, P. & Landrigan, P. J. Neurobehavioural effects of developmental toxicity. *Lancet Neurol.* **13**, 330–338 (2014).
16. Costa, L., G. Neurotoxicity of pesticides: a brief review. *Front. Biosci.* **13**, 1240 (2008).
17. Catapano, P. *et al.* What is the relationship between exposure to environmental pollutants and severe mental disorders? A systematic review on shared biological pathways. *Brain Behav. Immun. - Health* **43**, 100922 (2025).
18. Hassen, H. Y. *et al.* Association of environmental pollutants with asthma and allergy, and the mediating role of oxidative stress and immune markers in adolescents. *Environ. Res.* **265**, 120445 (2025).
19. Skakkebaek, N. E. *et al.* Male Reproductive Disorders and Fertility Trends: Influences of Environment and Genetic Susceptibility. *Physiol. Rev.* **96**, 55–97 (2016).
20. Karwacka, A., Zamkowska, D., Radwan, M. & Jurewicz, J. Exposure to modern, widespread environmental endocrine disrupting chemicals and their effect on the reproductive potential of women: an overview of current epidemiological evidence. *Hum. Fertil.* **22**, 2–25 (2019).
21. Nian, M., Braun, G., Escher, B. I. & Fang, M. Toxicological Study of Human Exposure to Mixtures of Chemicals: Challenges and Approaches. *Environ. Sci. Technol. Lett.* **11**, 773–782 (2024).
22. Cheng, Y., Ding, J., Arenas, C. E. D., Brinkmann, M. & Ji, X. A brief review on the assessment of potential joint effects of complex mixtures of contaminants in the environment. *Environ. Sci. Adv.* **3**, 661–675 (2024).
23. Tisler, S. *et al.* From data to reliable conclusions: Identification and comparison of persistent micropollutants and transformation products in 37 wastewater samples by non-target screening prioritization. *Water Res.* **219**, 118599 (2022).
24. Meng, D. *et al.* Development of an integral strategy for non-target and target analysis of site-specific potential contaminants in surface water: A case study of Dianshan Lake, China. *Chemosphere* **243**, 125367 (2020).
25. Choi, Y. *et al.* Identification, quantification, and prioritization of new emerging pollutants in domestic and industrial effluents, Korea: Application of LC-HRMS based suspect and non-target screening. *J. Hazard. Mater.* **402**, 123706 (2021).

26. Zang, L. *et al.* Characterization of non-volatile organic contaminants in coking wastewater using non-target screening: Dominance of nitrogen, sulfur, and oxygen-containing compounds in biological effluents. *Sci. Total Environ.* **837**, 155768 (2022).
27. Paszkiewicz, M. *et al.* Advances in suspect screening and non-target analysis of polar emerging contaminants in the environmental monitoring. *TrAC Trends Anal. Chem.* **154**, 116671 (2022).
28. Schwarzenbach, R. P. *et al.* The Challenge of Micropollutants in Aquatic Systems. *Science* **313**, 1072–1077 (2006).
29. Mitra, S., Saran, R. K., Srivastava, S. & Rensing, C. Pesticides in the environment: Degradation routes, pesticide transformation products and ecotoxicological considerations. *Sci. Total Environ.* **935**, 173026 (2024).
30. Postigo, C. & Barceló, D. Synthetic organic compounds and their transformation products in groundwater: Occurrence, fate and mitigation. *Sci. Total Environ.* **503–504**, 32–47 (2015).
31. Maculewicz, J. *et al.* Transformation products of pharmaceuticals in the environment: Their fate, (eco)toxicity and bioaccumulation potential. *Sci. Total Environ.* **802**, 149916 (2022).
32. Li, J. & Richardson, S. D. Fifty years' study of drinking water disinfection by-products: Identification, formation, occurrence, and state-of-the-art analytical methods. *TrAC Trends Anal. Chem.* **198**, 118668 (2026).
33. Meyer, C. *et al.* Ozonation of Pharmaceuticals and Their Human Metabolites in Wastewater—Insights from Laboratory Experiments and Field Data. *Environ. Sci. Technol.* **59**, 20842–20859 (2025).
34. Li, J. *et al.* Occurrence, Removal, and Risk Assessment of Antimicrobials and Their Transformation Products in Effluent from Australian Wastewater Treatment Plants. *Environ. Sci. Technol.* **59**, 6825–6838 (2025).
35. Xia, Y. *et al.* Structural Molecular Network for Nontargeted Screening and Prioritization of New Pollutants in Urban Wastewater. *Environ. Sci. Technol.* **59**, 17846–17856 (2025).
36. Hu, J., Lyu, Y., Chen, H., Li, S. & Sun, W. Suspect and Nontarget Screening Reveal the Underestimated Risks of Antibiotic Transformation Products in Wastewater Treatment Plant Effluents. *Environ. Sci. Technol.* **57**, 17439–17451 (2023).
37. Finckh, S. *et al.* Mapping chemical footprints of organic micropollutants in European streams. *Environ. Int.* **183**, 108371 (2024).

38. Malaj, E. *et al.* Organic chemicals jeopardize the health of freshwater ecosystems on the continental scale. *Proc. Natl. Acad. Sci.* **111**, 9549–9554 (2014).
39. Itzel, F. *et al.* Evaluation of a biological post-treatment after full-scale ozonation at a municipal wastewater treatment plant. *Water Res.* **170**, 115316 (2020).
40. Houtman, C. J. *et al.* High resolution effect-directed analysis of steroid hormone (ant)agonists in surface and wastewater quality monitoring. *Environ. Toxicol. Pharmacol.* **80**, 103460 (2020).
41. Zwart, N. *et al.* Identification of mutagenic and endocrine disrupting compounds in surface water and wastewater treatment plant effluents using high-resolution effect-directed analysis. *Water Res.* **168**, 115204 (2020).
42. Finckh, S. *et al.* Endocrine disrupting chemicals entering European rivers: Occurrence and adverse mixture effects in treated wastewater. *Environ. Int.* **170**, 107608 (2022).
43. Houtman, C. J. *et al.* Characterisation of (anti-)progestogenic and (anti-)androgenic activities in surface and wastewater using high resolution effectdirected analysis. *Environ. Int.* **153**, 106536 (2021).
44. Escher, B. I., van Daele, C., Dutt, M., Tang, J. Y. M. & Altenburger, R. Most Oxidative Stress Response In Water Samples Comes From Unknown Chemicals: The Need For Effect-Based Water Quality Trigger Values. *Environ. Sci. Technol.* **47**, 7002–7011 (2013).
45. Acheampong, P. & Thomas, S. H. L. Determinants of hepatotoxicity after repeated suprathreshold paracetamol ingestion: systematic review of reported cases. *Br. J. Clin. Pharmacol.* **82**, 923–931 (2016).
46. Allen, A. L. The diagnosis of acetaminophen toxicosis in a cat. *Can. Vet. J.* **44**, 509–510 (2003).
47. Sánchez-Bayo, F., van den Brink, P. J. & Mann, R. M. *Ecological Impacts of Toxic Chemicals*. (Bentham Science Publishers, 2011).
48. Casarett, L. J., Doull, J. & Klaassen, C. D. *Casarett and Doull's Toxicology: The Basic Science of Poisons*. (McGraw-Hill, New York, 2008).
49. Escher, B., Neale, P. & Leusch, F. *Bioanalytical Tools in Water Quality Assessment*. (IWA Publishing, London, 2021).
50. Su, T. *et al.* The feasibility of the zebrafish embryo as a promising alternative for acute toxicity test using various fish species: A critical review. *Sci. Total Environ.* **787**, 147705 (2021).
51. Escher, B. I., Stapleton, H. M. & Schymanski, E. L. Tracking complex mixtures of chemicals in our changing environment. *Science* **367**, 388–392 (2020).

52. Kortenkamp, A. Low dose mixture effects of endocrine disrupters and their implications for regulatory thresholds in chemical risk assessment. *Curr. Opin. Pharmacol.* **19**, 105–111 (2014).
53. Kim, S. *et al.* PubChem 2025 update. *Nucleic Acids Res.* **53**, D1516–D1525 (2025).
54. Bolton, E., Schymanski, E., Kondić, T., Thiessen, P. & Zhang, J. PubChemLite for Exposomics. <https://doi.org/10.5281/zenodo.14560968> (2024).
55. Schymanski, E. L. *et al.* Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag. *J. Cheminformatics* **13**, 19 (2021).
56. Network, N. *et al.* S0 | SUSDAT | Merged NORMAN Suspect List: SusDat. Zenodo <https://doi.org/10.5281/zenodo.15741861> (2025).
57. Krauss, M., Singer, H. & Hollender, J. LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns. *Anal. Bioanal. Chem.* **397**, 943–951 (2010).
58. Hollender, J., Schymanski, E. L., Singer, H. P. & Ferguson, P. L. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? *Environ. Sci. Technol.* **51**, 11505–11512 (2017).
59. Hollender, J. *et al.* NORMAN guidance on suspect and non-target screening in environmental monitoring. *Environ. Sci. Eur.* **35**, 75 (2023).
60. Kebarle, P. & Tang, L. From ions in solution to ions in the gas phase - the mechanism of electrospray mass spectrometry. *Anal. Chem.* **65**, 972A–986A (1993).
61. Cech, N. B. & Enke, C. G. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrom. Rev.* **20**, 362–387 (2001).
62. Szabo, D. *et al.* Online and Offline Prioritization of Chemicals of Interest in Suspect Screening and Non-targeted Screening with High-Resolution Mass Spectrometry. *Anal. Chem.* **96**, 3707–3716 (2024).
63. Zweigle, J. *et al.* Prioritization strategies for non-target screening in environmental samples by chromatography – High-resolution mass spectrometry: A tutorial. *J. Chromatogr. A* **1751**, 465944 (2025).
64. Guo, J. & Huan, T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics. *Anal. Chem.* **92**, 8072–8080 (2020).
65. Peets, P., Rian, M. B., Martin, J. W. & Krueve, A. Evaluation of Nontargeted Mass Spectral Data Acquisition Strategies for Water Analysis and Toxicity-Based Feature Prioritization by MS2Tox. *Environ. Sci. Technol.* **58**, 17406–17418 (2024).

66. Szabo, D., Fischer, S., Mathew, A. P. & Krueve, A. Prioritization, Identification, and Quantification of Emerging Contaminants in Recycled Textiles Using Non-Targeted and Suspect Screening Workflows by LC-ESI-HRMS. *Anal. Chem.* **96**, 14150–14159 (2024).
67. Gago-Ferrero, P., Krettek, A., Fischer, S., Wiberg, K. & Ahrens, L. Suspect Screening and Regulatory Databases: A Powerful Combination To Identify Emerging Micropollutants. *Environ. Sci. Technol.* **52**, 6881–6894 (2018).
68. Tue, N. M. *et al.* Nontarget Screening of Organohalogen Compounds in the Liver of Wild Birds from Osaka, Japan: Specific Accumulation of Highly Chlorinated POP Homologues in Raptors. *Environ. Sci. Technol.* **55**, 8691–8699 (2021).
69. Zweigle, J., Bugsel, B. & Zwiener, C. Efficient PFAS prioritization in non-target HRMS data: systematic evaluation of the novel MD/C-m/C approach. *Anal. Bioanal. Chem.* **415**, 1791–1801 (2023).
70. Trier, X., Granby, K. & Christensen, J. H. Polyfluorinated surfactants (PFS) in paper and board coatings for food packaging. *Environ. Sci. Pollut. Res.* **18**, 1108–1120 (2011).
71. Meekel, N., Krueve, A., Lamoree, M. H. & Been, F. M. Machine Learning-based Classification for the Prioritization of Potentially Hazardous Chemicals with Structural Alerts in Nontarget Screening. *Environ. Sci. Technol.* **59**, 5056–5065 (2025).
72. Haalck, I., Krauss, M., Brack, W. & Huber, C. Exploring Domestic Discharge Patterns in Wastewater through LC-HRMS Screening and Temporal Clustering. *Environ. Sci. Technol.* **59**, 15375–15384 (2025).
73. Liscio, C., Abdul-Sada, A., Al-Salhi, R., Ramsey, M. H. & Hill, E. M. Methodology for profiling anti-androgen mixtures in river water using multiple passive samplers and bioassay-directed analyses. *Water Res.* **57**, 258–269 (2014).
74. Cha, J. *et al.* Novel polar AhR-active chemicals detected in sediments of an industrial area using effect-directed analysis based on in vitro bioassays with full-scan high resolution mass spectrometric screening. *Sci. Total Environ.* **779**, 146566 (2021).
75. Gwak, J. *et al.* Effect-directed identification of novel aryl hydrocarbon receptor-active aromatic compounds in coastal sediments collected from a highly industrialized area. *Sci. Total Environ.* **803**, 149969 (2022).

76. Jonkers, T. J. H. *et al.* High-Performance Data Processing Workflow Incorporating Effect-Directed Analysis for Feature Prioritization in Suspect and Nontarget Screening. *Environ. Sci. Technol.* **56**, 1639–1651 (2022).
77. Hupatz, H. *et al.* Critical review on in silico methods for structural annotation of chemicals detected with LC/HRMS non-targeted screening. *Anal. Bioanal. Chem.* **417**, 473–493 (2025).
78. Nguyen, D. H., Nguyen, C. H. & Mamitsuka, H. Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Brief. Bioinform.* **20**, 2028–2043 (2019).
79. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
80. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
81. Xue, J., Guijas, C., Benton, H. P., Warth, B. & Siuzdak, G. METLIN MS2 molecular standards database: a broad chemical and biological resource. *Nat. Methods* **17**, 953–954 (2020).
82. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminformatics* **8**, 3 (2016).
83. Allen, F., Greiner, R. & Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **11**, 98–110 (2015).
84. Goldman, S., Li, J. & Coley, C. W. Generating Molecular Fragmentation Graphs with Autoregressive Neural Networks. *Anal. Chem.* **96**, 3419–3428 (2024).
85. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
86. Goldman, S. *et al.* Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nat. Mach. Intell.* **5**, 965–979 (2023).
87. Stravs, M. A., Dührkop, K., Böcker, S. & Zamboni, N. MSNovelist: de novo structure generation from mass spectra. *Nat. Methods* **19**, 865–870 (2022).
88. Litsa, E. E., Chenthamarakshan, V., Das, P. & Kavrakı, L. E. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Commun. Chem.* **6**, 132 (2023).
89. Bittremieux, W. *et al.* MassSpecGym: A benchmark for the discovery and identification of molecules. in *Advances in Neural Information Processing Systems*

- 37 110010–110027 (Neural Information Processing Systems Foundation, Inc. (NeurIPS), Vancouver, BC, Canada, 2024).
90. Schymanski, E. L. *et al.* Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).
  91. Williams, A. J. *et al.* The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J. Cheminformatics* **9**, 61 (2017).
  92. Fonger, G. C., Hakkinen, P., Jordan, S. & Publicker, S. The National Library of Medicine's (NLM) Hazardous Substances Data Bank (HSDB): Background, recent enhancements and future plans. *Toxicology* **325**, 209–216 (2014).
  93. Guo, W. *et al.* Review of machine learning and deep learning models for toxicity prediction. *Exp. Biol. Med.* **248**, 1952–1973 (2023).
  94. Ecological Structure Activity Relationships (ECOSAR) Predictive Model. U.S. Environmental Protection Agency (2022).
  95. Liigand, P., Liigand, J., Kaupmees, K. & Kruve, A. 30 Years of research on ESI/MS response: Trends, contradictions and applications. *Anal. Chim. Acta* **1152**, 238117 (2021).
  96. Liigand, J. *et al.* Quantification for non-targeted LC/MS screening without standard substances. *Sci. Rep.* **10**, 5808 (2020).
  97. Oss, M., Kruve, A., Herodes, K. & Leito, I. Electrospray Ionization Efficiency Scale of Organic Compounds. *Anal. Chem.* **82**, 2865–2872 (2010).
  98. Oss, M. *et al.* Quantitative electrospray ionization efficiency scale: 10 years after. *Rapid Commun. Mass Spectrom.* **35**, (2021).
  99. Cech, N. B. & Enke, C. G. Relating Electrospray Ionization Response to Nonpolar Character of Small Peptides. *Anal. Chem.* **72**, 2717–2723 (2000).
  100. Amad, M. H., Cech, N. B., Jackson, G. S. & Enke, C. G. Importance of gas-phase proton affinities in determining the electrospray ionization response for analytes and solvents. *J. Mass Spectrom.* **35**, 784–789 (2000).
  101. Liigand, J., Laaniste, A. & Kruve, A. pH Effects on Electrospray Ionization Efficiency. *J. Am. Soc. Mass Spectrom.* **28**, 461–469 (2017).
  102. Liigand, J., Kruve, A., Leito, I., Girod, M. & Antoine, R. Effect of Mobile Phase on Electrospray Ionization Efficiency. *J. Am. Soc. Mass Spectrom.* **25**, 1853–1861 (2014).
  103. Ojakivi, M., Liigand, J. & Kruve, A. Modifying the Acidity of Charged Droplets. *ChemistrySelect* **3**, 335–338 (2018).

104. Malm, L. *et al.* Quantification Approaches in Non-Target LC/ESI/HRMS Analysis: An Interlaboratory Comparison. *Anal. Chem.* **96**, 16215–16226 (2024).
105. Lauria, M. Z. *et al.* Closing the Organofluorine Mass Balance in Marine Mammals Using Suspect Screening and Machine Learning-Based Quantification. *Environ. Sci. Technol.* **58**, 2458–2467 (2024).
106. Wu, L. *et al.* Quantitative structure–ion intensity relationship strategy to the prediction of absolute levels without authentic standards. *Anal. Chim. Acta* **794**, 67–75 (2013).
107. Kruve, A. & Kaupmees, K. Predicting ESI/MS Signal Change for Anions in Different Solvents. *Anal. Chem.* **89**, 5079–5086 (2017).
108. Liigand, P., Liigand, J., Cuyckens, F., Vreeken, R. J. & Kruve, A. Ionisation efficiencies can be predicted in complicated biological matrices: A proof of concept. *Anal. Chim. Acta* **1032**, 68–74 (2018).
109. Aalizadeh, R., Panara, A. & Thomaidis, N. S. Development and Application of a Novel Semi-quantification Approach in LC-QToF-MS Analysis of Natural Products. *J. Am. Soc. Mass Spectrom.* **32**, 1412–1423 (2021).
110. Kruve, A., Kaupmees, K., Liigand, J. & Leito, I. Negative Electrospray Ionization via Deprotonation: Predicting the Ionization Efficiency. *Anal. Chem.* **86**, 4822–4830 (2014).
111. Palm, E. & Kruve, A. Machine Learning for Absolute Quantification of Unidentified Compounds in Non-Targeted LC/HRMS. *Molecules* **27**, 1013 (2022).
112. Mayhew, A. W., Topping, D. O. & Hamilton, J. F. New Approach Combining Molecular Fingerprints and Machine Learning to Estimate Relative Ionization Efficiency in Electrospray Ionization. *ACS Omega* **5**, 9510–9516 (2020).
113. Chalcraft, K. R., Lee, R., Mills, C. & Britz-McKibbin, P. Virtual Quantification of Metabolites by Capillary Electrophoresis-Electrospray Ionization-Mass Spectrometry: Predicting Ionization Efficiency Without Chemical Standards. *Anal. Chem.* **81**, 2506–2515 (2009).
114. Khabazbashi, S., Engelhardt, J., Möckel, C., Weiss, J. & Kruve, A. Estimation of the concentrations of hydroxylated polychlorinated biphenyls in human serum using ionization efficiency prediction for electrospray. *Anal. Bioanal. Chem.* **414**, 7451–7460 (2022).
115. Grouw, A. V. *et al.* Toward Machine Learning Electrospray Ionization Sensitivity Prediction for Semiquantitative Lipidomics in Stem Cells. *J. Chem. Inf. Model.* **65**, 1826–1836 (2025).

116. Panagopoulos Abrahamsson, D., Park, J.-S., Singh, R. R., Sirota, M. & Woodruff, T. J. Applications of Machine Learning to In Silico Quantification of Chemicals without Analytical Standards. *J. Chem. Inf. Model.* **60**, 2718–2727 (2020).
117. Aalizadeh, R., Nikolopoulou, V., Alygizakis, N., Slobodnik, J. & Thomaidis, N. S. A novel workflow for semi-quantification of emerging contaminants in environmental samples analyzed by LC-HRMS. *Anal. Bioanal. Chem.* **414**, 7435–7450 (2022).
118. Wang, W.-C., Amini, N., Huber, C., Kull, M. & Krueve, A. Active Learning Improves Ionization Efficiency Predictions and Quantification in Nontargeted LC/HRMS. *Anal. Chem.* **97**, 13131–13139 (2025).
119. Kaskow, J. A. *et al.* Predicting Peptide Ionization Efficiencies for Electrospray Ionization Mass Spectrometry Using Machine Learning. *J. Am. Soc. Mass Spectrom.* **35**, 2297–2307 (2024).
120. Nikolopoulos, A. *et al.* Ionization efficiency prediction of electrospray ionization mass spectrometry analytes based on molecular fingerprints and cumulative neutral losses. *J. Cheminformatics* **17**, 183 (2025).
121. Dührkop, K. Deep kernel learning improves molecular fingerprint prediction from tandem mass spectra. *Bioinformatics* **38**, i342–i349 (2022).
122. Heinonen, M., Shen, H., Zamboni, N. & Rousu, J. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics* **28**, 2333–2341 (2012).
123. Bushuiev, R. *et al.* Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. *Nat. Biotechnol.* 1–11 (2025).
124. Huber, F., van der Burg, S., van der Hooft, J. J. J. & Ridder, L. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J. Cheminformatics* **13**, 84 (2021).
125. Voronov, G. *et al.* MS2Prop: A machine learning model that directly generates de novo predictions of drug-likeness of natural products from unannotated MS/MS spectra. Preprint at <https://doi.org/10.1101/2022.10.09.511482> (2024).
126. Peets, P. *et al.* MS2Tox Machine Learning Tool for Predicting the Ecotoxicity of Unidentified Chemicals in Water by Nontarget LC-HRMS. *Environ. Sci. Technol.* **56**, 15508–15517 (2022).
127. Rahu, I., Kull, M. & Krueve, A. Predicting the Activity of Unidentified Chemicals in Complementary Bioassays from the HRMS Data to Pinpoint Potential Endocrine Disruptors. *J. Chem. Inf. Model.* **64**, 3093–3104 (2024).

128. Arturi, K. *et al.* MLin vitroTox reloaded for high-throughput hazard-based prioritization of high-resolution mass spectrometry data. *J. Cheminformatics* **17**, 14 (2025).
129. Fan, F. *et al.* FT-GNN Tool for Bridging HRMS Features and Bioactivity: Uncovering Unidentified Estrogen Receptor Agonists in Sewage. *Environ. Sci. Technol.* **59**, 7736–7746 (2025).
130. Meekel, N., Krueve, A., Lamoree, M. H. & Been, F. M. Machine Learning-based Classification for the Prioritization of Potentially Hazardous Chemicals with Structural Alerts in Nontarget Screening. *Environ. Sci. Technol.* **59**, 5056–5065 (2025).
131. Alvarez-Mora, I. *et al.* High-Throughput Effect-Directed Analysis of Androgenic Compounds in Hospital Wastewater: Identifying Effect Drivers through Non-Target Screening Supported by Toxicity Prediction. *Environ. Sci. Technol.* **59**, 24513–24525 (2025).
132. Scholz, S., Klüver, N. & Kühne, R. *Analysis of the Relevance and Adequateness of Using Fish Embryo Acute Toxicity (FET) Test Guidance (OECD 236) to Fulfil the Information Requirements and Addressing Concerns under REACH.* 104 [https://echa.europa.eu/documents/10162/17231/fet\\_report\\_en.pdf/b6036bdb-9041-41c8-a390-d9b66b244a4b?t=1466155378221](https://echa.europa.eu/documents/10162/17231/fet_report_en.pdf/b6036bdb-9041-41c8-a390-d9b66b244a4b?t=1466155378221) (2016).
133. Tsugawa, H. *et al.* MS-DIAL 4: accelerating lipidomics using an MS/MS, CCS, and retention time atlas. Preprint at <https://doi.org/10.1101/2020.02.11.944900> (2020).
134. Takeda, H. *et al.* MS-DIAL 5 multimodal mass spectrometry data mining unveils lipidome complexities. *Nat. Commun.* **15**, 9903 (2024).