

PhD-FSTM-2022-140  
The Faculty of Science, Technology and Medicine

## DISSERTATION

Defence held on 16/12/2022 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG  
EN CHIMIE  
and  
doctor rerum naturalium (Dr. rer. nat.)

by

**Adelene Lai Shuen Lyn**

Born on 30 June 1993 in Kuala Lumpur (Malaysia)

CHEMINFORMATICS AND COMPUTATIONAL  
APPROACHES FOR IDENTIFYING AND MANAGING  
UNKNOWN CHEMICALS IN THE ENVIRONMENT

### Dissertation defence committee

Dr Emma Schymanski, dissertation supervisor

*Associate Professor, Université du Luxembourg/Luxembourg Centre for Systems Biomedicine*

Dr Christoph Steinbeck, dissertation supervisor

*Vice President for Digitalisation, Friedrich-Schiller-Universität Jena*

*Professor, Friedrich-Schiller-Universität Jena*

Dr Reinhard Schneider, Chair

*Professor, Université du Luxembourg/Luxembourg Centre for Systems Biomedicine*

Dr Steffen Neumann

*Group Leader, Leibniz Institute of Plant Biochemistry, IPB Halle*

Dr Michael Stelter

*Professor, Friedrich-Schiller-Universität Jena*

Dr Egon Willighagen

*Assistant Professor, Maastricht University*



# Cheminformatics and Computational Approaches for Identifying and Managing Unknown Chemicals in the Environment

Dissertation  
(KUMULATIV)

zur Erlangung des akademischen Grades  
*doctor rerum naturalium*  
(Dr. rer. nat.)

und

Docteur de l'université du Luxembourg en Chimie

vorgelegt dem Rat der Chemisch-Geowissenschaftlichen Fakultät der  
Friedrich-Schiller-Universität Jena

und

Université du Luxembourg

von

**M. Sc. (Umweltnaturwissenschaften) Adelene Lai Shuen Lyn**  
geboren am 30.06.1993 in Kuala Lumpur, Malaysia





# Table of Contents

<b>Table of Contents</b>	<b>I</b>
<b>Affidavit - University of Luxembourg</b>	<b>III</b>
<b>Zusammenfassung</b>	<b>IV</b>
<b>Summary</b>	<b>VIII</b>
<b>List of Abbreviations</b>	<b>XII</b>
<b>Chapter 1</b>	<b>1</b>
<b>Introduction</b>	<b>1</b>
1.1 Background	1
1.1.1 Chemicals in the Environment	1
1.1.2 Chemicals Management	2
1.1.2.1 Chemicals Assessment	3
1.1.2.1.1 Chemical Identity and Data Representations	5
1.1.2.2 Environmental Monitoring	10
1.1.2.2.1 Compound Identification Workflows for Identifying Environmental Chemicals using LC-HRMS	11
1.1.3 Research Gaps in Identifying and Managing Chemical Unknowns	14
1.2 Aims	16
1.3 Scope of the Dissertation	16
<b>Chapter 2</b>	<b>18</b>
<b>Developing an Open Computational Workflow using Emerging Digital Chemistry Resources for Non-target Analysis</b>	<b>18</b>
Publication A	20
<b>Chapter 3</b>	<b>43</b>
<b>Data Mining Transformation Product Information for Enhanced Suspect Screening</b>	<b>43</b>
Publication B	45
<b>Chapter 4</b>	<b>61</b>
<b>Tackling the Next Frontier of Environmental Unknowns - UVCBs</b>	<b>61</b>
Publication C	62
<b>Chapter 5</b>	<b>83</b>
<b>A Cheminformatics Algorithm for Improved Identification of Homologous Series in Environmental Mixtures</b>	<b>83</b>
Publication D	85
<b>Chapter 6</b>	<b>112</b>

<b>Discussion</b>	<b>112</b>
6.1 Using Environmental Metadata in MetFrag for Unknown Identification	115
6.2 Suspect Screening using Suspect Lists	116
6.3 Different Workflows for Different Studies? Harmonisation of NTA Towards Use in Regulatory Environmental Monitoring	117
6.4 Open Science and FAIR Data Approaches of this Dissertation	118
6.5 Bottleneck in Availability of Data on Environmental Pollutants	120
<b>Chapter 7</b>	<b>122</b>
<b>Conclusion &amp; Perspectives</b>	<b>122</b>
<b>References</b>	<b>127</b>
<b>Erklärungen</b>	<b>XIV</b>
<b>Curriculum Vitae</b>	<b>XV</b>
<b>Acknowledgements</b>	<b>XIX</b>

# Affidavit - University of Luxembourg

I hereby confirm that the PhD dissertation entitled “**Cheminformatics and Computational Approaches for Identifying and Managing Unknown Chemicals in the Environment**” has been written independently and without any other sources than those cited. No clinical samples nor laboratory animals were used in this work.

Luxembourg, \_\_\_\_\_

\_\_\_\_\_  
Adelene Lai Shuen Lyn

# Zusammenfassung

In den meisten Gesellschaften ist die Verwendung chemischer Produkte zu einem Teil des täglichen Lebens geworden. Weltweit sind mehr als 350.000 Chemikalien für den täglichen Gebrauch im Haushalt, in industriellen Prozessen, in der Landwirtschaft usw. registriert. Trotz des Nutzens, den Chemikalien für die Gesellschaft haben können, haben ihre Verwendung, Herstellung und Entsorgung, die schließlich zu ihrer Freisetzung in die Umwelt führt, vielfältige Auswirkungen. Anthropogene Chemikalien wurden in unzähligen Ökosystemen auf der ganzen Welt sowie in den Geweben wild lebender Tiere und des Menschen nachgewiesen. Die potenziellen Folgen einer solchen chemischen Verschmutzung sind noch nicht vollständig geklärt, aber das Auftreten menschlicher Krankheiten und die Bedrohung der Artenvielfalt werden mit dem Vorhandensein von Chemikalien in unserer Umwelt in Verbindung gebracht.

Die Abschwächung der potenziellen negativen Auswirkungen chemischer Stoffe erfordert in der Regel regulatorische Maßnahmen und eine Vielzahl von Akteuren. Ein wichtiger Aspekt dabei ist die Umweltüberwachung, die aus Umweltproben, Messungen, Datenanalyse und Berichterstattung besteht. In den letzten Jahren haben Fortschritte in der gekoppelten Anwendung von Flüssigchromatographie und hochaufgelöster Massenspektrometrie (*Liquid Chromatography-High Resolution Mass Spectrometry*, LC-HRMS), offenen chemischen Datenbanken und Software es der Forschung ermöglicht, sowohl bekannte (z. B. Pestizide) als auch unbekannte Umweltchemikalien zu identifizieren, die gemeinhin als *Suspect* (Verdachtsfall) oder *Non-target* (Nicht-Zielsubstanz) bezeichnet werden. Die Identifizierung unbekannter Chemikalien stellt jedoch eine große Herausforderung dar, da die Analyten nicht von vornherein bekannt sind - alles, was zur Verfügung steht, sind ihre Massenspektrometriesignale. Tatsächlich geht die Zahl der unbekannt Merkmale in einem typischen Massenspektrum einer Umweltprobe in die Tausende bis Zehntausende und erfordert daher eine Priorisierung der Merkmale vor der Identifizierung im Rahmen eines geeigneten Analyseprogramms.

Im Rahmen dieser Doktorarbeit wurde in Zusammenarbeit mit zwei für die Umweltüberwachung zuständigen Behörden versucht, unbekannte Verbindungen in der Umwelt zu identifizieren, insbesondere durch die Entwicklung von

computergestützten Workflows zur Identifizierung unbekannter Verbindungen in LC-HRMS-Daten. Die erste Zusammenarbeit hat **Publikation A** hervorgebracht, die ein gemeinsames Projekt mit dem Zürcher Amt für Wasser, Energie und Luft beinhaltete. Umweltproben, die von Kläranlagen in der Schweiz entnommen wurden, wurden retrospektiv mit Hilfe eines Pre-Screening-Workflows analysiert, der die für die Identifizierung von Nicht-Zielsubstanzen geeigneten Merkmale priorisiert. Zu diesem Zweck wurde ein mehrstufiger Qualitätskontrollalgorithmus entwickelt, der die Qualität der Massenspektraldaten hinsichtlich der Signalintensitäten, der Signalpositionen und des Signal-Rausch-Verhältnisses überprüft und im Rahmen des Pre-Screenings eingesetzt wird. Dieser Algorithmus wurde in das R-Paket Shinyscreen integriert. Merkmale, die durch das Pre-Screening als vorrangig eingestuft wurden, wurden anschließend mit der In-Silico-Fragmentierungssoftware MetFrag identifiziert. Um diese Identifizierungen zu erhalten, wurde MetFrag mit verschiedenen offenen chemischen Informationsquellen wie Spektraldatenbanken wie MassBank Europe und MassBank of North America sowie mit Verdachtslisten aus dem NORMAN Suspect List Exchange und die Datenbank EPA CompTox Chemicals Dashboard gekoppelt. Es wurden eine bestätigte und einundzwanzig vorläufige Identifizierungen von Verbindungen erzielt, die nach einem festgelegten Vertrauenslevel gemeldet wurden. Umfassende Dateninterpretation und detaillierte Kommunikation der MetFrag-Ergebnisse wurden durchgeführt, um evidenzbasierte Empfehlungen zu formulieren, die in zukünftige Umweltüberwachungskampagnen einfließen können.

Aufbauend auf dem in Publikation A entwickelten Vorab-Screening- und Identifizierungs-Workflow ist **Publikation B** das Ergebnis einer Zusammenarbeit mit der luxemburgischen Administration de la gestion de l'eau, die sich zum Ziel gesetzt hat, unbekannte Chemikalien in der luxemburgischen Umwelt zu identifizieren und, wenn möglich, zu quantifizieren. Konkret wurden Oberflächenwasserproben, die im Rahmen einer zweijährigen nationalen Überwachungskampagne entnommen wurden, mit LC-HRMS gemessen und auf pharmazeutische Ausgangsverbindungen und deren Umwandlungsprodukte untersucht. Im Vergleich zu den Informationen über pharmazeutische Verbindungen, die von den lokalen Behörden öffentlich zugänglich sind und für die Verdachtsliste verwendet wurden, sind die Informationen über Umwandlungsprodukte relativ spärlich. Daher wurden in dieser Arbeit neue Ansätze entwickelt, um Daten aus der PubChem-Datenbank sowie aus der Literatur

auszuwerten, um eine Verdachtsliste zu erstellen, die neben den Ausgangsverbindungen auch pharmazeutische Umwandlungsprodukte enthält. Insgesamt wurden 94 Arzneimittel und 14 Umwandlungsprodukte identifiziert, von denen 88 bzw. 2 bestätigt wurden. Das räumlich-zeitliche Auftreten und die Verteilung dieser Verbindungen in der luxemburgischen Umwelt wurden mit Hilfe fortschrittlicher Datenvisualisierungen analysiert, die Muster in bestimmten Regionen und Zeiträumen mit hohem Aufkommen aufzeigten. Diese Ergebnisse können künftige Maßnahmen zum Chemikalienmanagement unterstützen, insbesondere bei der Umweltüberwachung.

Eine weitere Herausforderung beim Umgang mit Chemikalien besteht darin, dass sie meist als komplexe Gemische in der Umwelt und in chemischen Produkten vorkommen. Stoffe mit unbekannter oder variabler Zusammensetzung, komplexe Reaktionsprodukte oder biologische Materialien (*Substances of Unknown or Variable composition, Complex reaction products or Biological materials*, UVCBs) machen 20-40 % der internationalen Chemikalienregister aus und umfassen chlorierte Paraffine, Polymermischungen, Erdölfraktionen und ätherische Öle. Allerdings ist nur wenig über ihre chemische Identität und/oder Zusammensetzung bekannt, was die Bewertung ihres Verbleibs und ihrer Toxizität in der Umwelt erschwert, ganz zu schweigen von ihrer Identifizierung in der Umwelt. **Publikation C** befasst sich mit den Herausforderungen von UVCBs, indem sie einen multidisziplinären Ansatz bei der Literaturarbeit verfolgt, welcher Überlegungen zu ihren chemischen Darstellungen, ihrer Toxizität, ihrem Verbleib in der Umwelt, ihrer Exposition und ihren regulatorischen Ansätzen einbezieht. Verbesserte Anforderungen an die Registrierung von Stoffen, Gruppierungstechniken zur Vereinfachung der Bewertung und die Verwendung von *Mixture InChI* zur Darstellung von UVCBs in einer auffindbaren, zugänglichen, interoperablen und wiederverwendbaren (*findable, accessible, interoperable, and reusable*, FAIR) Weise in Datenbanken gehören zu den wichtigsten Empfehlungen dieser Arbeit.

Eine bestimmte Art von UVCB, Mischungen homologer Verbindungen, werden bisher häufig in Umweltproben nachgewiesen, darunter viele *High Production Volume substances* wie zum Beispiel Tenside. Verbindungen, die homologe Reihen bilden, sind durch ein gemeinsames Kernfragment und eine sich wiederholende chemische Untereinheit verwandt und können durch allgemeine Formeln (z. B.  $C_nF_{2n+1}COOH$ )

und/oder Markush-Strukturen dargestellt werden. Ein erhebliches Problem bei der Identifizierung ist jedoch, ihre charakteristischen analytischen Signale in LC-HRMS-Daten mit Chemikalien in Datenbanken abzugleichen; während kammartige Elutionsmuster und konstante Unterschiede im Masse-Ladungs-Verhältnis auf das Vorhandensein homologer Serien in Proben hinweisen, enthalten die meisten chemischen Datenbanken keine annotierten homologen Serien. Um diese Lücke zu schließen, wird in **Publikation D** ein chemieinformatischer Algorithmus, OngLai, vorgestellt, mit dem homologe Serien in strukturellen Datensätzen erkannt werden können. OngLai, das offen in Python unter Verwendung des RDKit implementiert wurde, erkennt homologe Serien auf der Grundlage von zwei Parametern: einer Liste von Verbindungen und der chemischen Struktur einer sich wiederholenden Untereinheit. OngLai wurde auf drei offene Datensätze aus den Bereichen Umweltchemie, Exposomik und Naturstoffe angewandt, in denen tausende homologe Serien (mit sich wiederholender CH<sub>2</sub>-Untereinheit) entdeckt wurden. Es wird erwartet, dass die Klassifizierung homologer Serien in Umweltdatensätzen deren analytische Erkennung in Umweltproben verbessern wird.

Insgesamt hat die Arbeit in dieser Dissertation dazu beigetragen, die Identifizierung und das Management unbekannter Chemikalien in der Umwelt mit Hilfe der Chemieinformatik und computergestützter Ansätze voranzutreiben. Alle Arbeiten wurden nach den Grundsätzen von Open Science und FAIR data durchgeführt: Alle Software, Datensätze, Analysen und Ergebnisse, einschließlich der endgültigen, im *peer review* Verfahren begutachteten Veröffentlichungen, sind für die Öffentlichkeit zugänglich. Diese Bemühungen sollen weitere Entwicklungen im Bereich der Identifizierung und des Managements unbekannter Chemikalien zum Schutz der Umwelt und der menschlichen Gesundheit ermöglichen.

# Summary

In most societies, using chemical products has become a part of daily life. Worldwide, over 350,000 chemicals have been registered for use in *e.g.*, daily household consumption, industrial processes, agriculture, *etc.* However, despite the benefits chemicals may bring to society, their usage, production, and disposal, which leads to their eventual release into the environment has multiple implications. Anthropogenic chemicals have been detected in myriad ecosystems all over the planet, as well as in the tissues of wildlife and humans. The potential consequences of such chemical pollution are not fully understood, but links to the onset of human disease and threats to biodiversity have been attributed to the presence of chemicals in our environment.

Mitigating the potential negative effects of chemicals typically involves regulatory steps and multiple stakeholders. One key aspect thereof is environmental monitoring, which consists of environmental sampling, measurement, data analysis, and reporting. In recent years, advancements in Liquid Chromatography-High Resolution Mass Spectrometry (LC-HRMS), open chemical databases, and software have enabled researchers to identify known (*e.g.*, pesticides) as well as unknown environmental chemicals, commonly referred to as suspect or non-target compounds. However, identifying unknown chemicals, particularly non-targets, remains extremely challenging because of the lack of *a priori* knowledge on the analytes - all that is available are their mass spectrometry signals. In fact, the number of unknown features in a typical mass spectrum of an environmental sample is in the range of thousands to tens of thousands, and therefore requires feature prioritisation before identification within a suitable workflow.

In this dissertation work, collaborations with two regulatory authorities responsible for environmental monitoring sought to identify relevant unknown compounds in the environment, specifically by developing computational workflows for unknown identification in LC-HRMS data. The first collaboration culminated in **Publication A**, which involved a joint project with the *Zürcher Amt für Wasser, Energie und Luft*. Environmental samples taken from wastewater treatment plant sites in Switzerland were retrospectively analysed using a pre-screening workflow that prioritised features



suitable for non-target identification. For this purpose, a multi-step Quality Control algorithm that checks the quality of mass spectral data in terms of peak intensities, alignment, and signal-to-noise ratio was developed and used within pre-screening. This algorithm was incorporated into the R package *ShinyScreen*. Features that were prioritised by pre-screening then underwent identification using the *in silico* fragmentation tool MetFrag. To obtain these identifications, MetFrag was coupled to various open chemical information resources such as spectral databases like MassBank Europe and MassBank of North America, as well as suspect lists from the NORMAN Suspect List Exchange and the CompTox Chemicals Dashboard database. One confirmed and twenty-one tentative compound identifications were achieved and reported according to an established confidence level scheme. Comprehensive data interpretation and detailed communication of MetFrag's results was performed as a means of formulating evidence-based recommendations that may inform future environmental monitoring campaigns.

Building on the pre-screening and identification workflow developed in Publication A, **Publication B** resulted from a collaboration with the Luxembourgish *Administration de la gestion de l'eau* that sought to identify, and where possible quantify unknown chemicals in Luxembourgish surface waters. More specifically, surface water samples collected as part of a two-year national monitoring campaign were measured using LC-HRMS and screened for pharmaceutical parent compounds and their transformation products. Compared to pharmaceutical compound information, which is publicly available from local authorities (and was used in the suspect list), information on transformation products is relatively scarce. Therefore, new approaches were developed in this work to mine data from the PubChem database as well as from the literature in order to formulate a suspect list containing pharmaceutical transformation products, in addition to their parent compounds. Overall, 94 pharmaceuticals and 14 transformation products were identified, of which 88 and 2 were confirmed identifications respectively. The spatio-temporal occurrence and distribution of these compounds throughout the Luxembourgish environment were analysed using advanced data visualisations that highlighted patterns in certain regions and time periods of high incidence. These findings may support future chemicals management measures, particularly in environmental monitoring.

Another challenging aspect of managing chemicals is that they mostly exist as complex mixtures within the environment as well as chemical products. Substances of Unknown or Variable composition, Complex reaction products or Biological materials (UVCBs) make up 20-40% of international chemical registries and include chlorinated paraffins, polymer mixtures, petroleum fractions, and essential oils. However, little is known about their chemical identities and/or compositions, which poses formidable obstacles to assessing their environmental fate and toxicity, let alone identification in the environment. **Publication C** addresses the challenges of UVCBs by taking an interdisciplinary approach in reviewing the literature that incorporates considerations of their chemical representations, toxicity, environmental fate, exposure, and regulatory approaches. Improved substance registration requirements, grouping techniques to simplify assessment, and the use of Mixture InChI to represent UVCBs in a findable, accessible, interoperable, and reusable (FAIR) way in databases are amongst the key recommendations of this work.

A specific type of UVCB, mixtures of homologous compounds, are commonly detected in environmental samples, including many High Production Volume (HPV) compounds such as surfactants. Compounds forming homologous series are related by a common core fragment and repeating chemical subunit, and can be represented using general formulae (e.g.,  $C_nF_{2n+1}COOH$ ) and/or Markush structures. However, a significant identification bottleneck is the inability to match their characteristic analytical signals in LC-HRMS data with chemicals in databases; while comb-like elution patterns and constant differences in mass-to-charge ratio indicate the presence of homologous series in samples, most chemical databases do not contain annotated homologous series. To address this gap, **Publication D** introduces a cheminformatics algorithm, OngLai, to detect homologous series within compound datasets. OngLai, openly implemented in Python using the RDKit, detects homologous series based on two inputs: a list of compounds and the chemical structure of a repeating unit. OngLai was applied to three open datasets from environmental chemistry, exposomics, and natural products, in which thousands of homologous series with a  $CH_2$  repeating unit were detected. Classification of homologous series in compound datasets is expected to advance their analytical detection in samples.

Overall, the work in this dissertation contributed to the advancement of identifying and managing unknown chemicals in the environment using cheminformatics and computational approaches. All work conducted followed Open Science and FAIR data principles: all code, datasets, analyses, and results generated, including the final peer-reviewed publications, are openly available to the public. These efforts are intended to spur further developments in unknown chemical identification and management towards protecting the environment and human health.

# List of Abbreviations

CASRN	Chemical Abstracts Service Registry Number
CBI	Confidential Business Information
DOI	Digital Object Identifier
ECHA	European Chemicals Agency
EINECS	European Inventory of Existing Chemical Structures
ELINCS	European List of Notified Chemical Substances
EPA	Environmental Protection Agency
GHS	Globally Harmonised System of Classification and Labelling of Chemicals
HPV	High Production Volume
IUCLID	International Uniform Chemical Information Database
LC-HRMS	Liquid Chromatography-High Resolution Mass Spectrometry
MEC	Measured Environmental Concentration
MoNA	MassBank of North America
MS1	Full Scan Mass Spectrum
MS2	Tandem Mass Spectrum (Fragmentation Spectrum)
NTA	Non-Target Analysis
OECD	Organisation for Economic Co-operation and Development
PBT	Persistent, Bioaccumulative, Toxic
PEC	Predicted Environmental Concentration
PFAS	Poly- and perfluorinated Alkyl Substances
PNEC	Predicted No Effect Concentration
POP	Persistent Organic Pollutant
QSAR	Quantitative Structure-Activity Relationship
QSPR	Quantitative Structure-Property Relationship
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
SAICM	Strategic Approach to International Chemicals Management
SLE	(NORMAN-) Suspect List Exchange
SSbD	Safe and Sustainable by Design
$t_R$	Retention time
TSCA	Toxic Substances Control Act
URL	Uniform Resource Locator
UVCB	(Substances) of Unknown or Variable Composition, Complex reaction Materials, or Biological materials

WFD European Union Water Framework Directive

# Chapter 1

## Introduction

### 1.1 Background

In most societies, using chemicals, particularly synthetic compounds, is a part of daily life. From household consumption to industrial production processes, military applications to agriculture, our lives have become dependent on chemicals to the extent that enumerating everything we use has become virtually impossible. For example, direct consumption of drugs and use of cleaning products, pesticides, cosmetics, and fragrances, are just some ways we use chemical products on a daily basis, in addition to indirect passive consumption through contact with everyday objects such as food packaging, high-performance clothing, non-stick cookware, flame-retardant furniture, plastic children's toys, paints and adhesives *etc.* Over 350,000 chemical products are known to be produced and used worldwide<sup>1</sup>, with over 400 million tonnes of industrial chemicals estimated to be produced annually.<sup>2</sup> That chemical products - and thus synthetic *i.e.*, anthropogenic chemicals - pervade daily human life in most industrial societies in the year 2022 is incontrovertible.<sup>3</sup>

#### 1.1.1 Chemicals in the Environment

While anthropogenic chemicals (hereafter, 'chemicals' unless specified otherwise) have brought various benefits to society, their production, use and disposal have resulted in emissions into the environment and subsequent human and environmental exposure. This phenomenon is known as chemical pollution, and is defined as the release and accumulation of chemicals in the environment. Chemical pollution has been identified as one of the nine planetary boundaries delineating a safe operating space for humanity in the form of 'novel entities',<sup>4,5</sup> and a recent study shows that this boundary has already been surpassed.<sup>6</sup> Chemical pollution has also been recognised as a threat to planetary health<sup>3</sup> and a global change factor,<sup>7</sup> to the detriment of ecosystems worldwide. Myriad environmental media in different ecosystems across the planet, including rainforests,<sup>8</sup> deserts,<sup>9</sup> coastal areas,<sup>10</sup> and

even the remote Arctic contain chemicals.<sup>11,12</sup> Chemicals have been found in soils,<sup>13</sup> sediments,<sup>14,15</sup> rivers,<sup>16,17</sup> lakes,<sup>18,19</sup> seas,<sup>20–23</sup> oceans,<sup>24</sup> and even rainwater.<sup>25,26</sup> Wildlife<sup>27–32</sup> and humans<sup>33–44</sup> contain chemicals, sometimes at levels higher than health safety thresholds. Chemicals have also been found in household environments within various matrices, including dust.<sup>45–47</sup> The full impacts of chemical pollution may not be fully known, but multiple studies have shown their connection to the onset of human disease<sup>48–55</sup> and potential role in altering ecology and biodiversity.<sup>56–59</sup> Further complicating the issue is the fact that chemicals in the environment exist as complex mixtures,<sup>60</sup> making it more complicated to track, model, and understand their effects. Multiple chemical compounds are routinely detected within individual environmental samples, which can confound the attribution of specific chemical(s) to their effects on humans and the environment. Evidently, the environmental problem posed by chemical pollution is of multiple dimensions; the problem of chemical pollution is not just of scale and omnipresence, but also of environmental fate - that is, that they end up as complex environmental mixtures.

### 1.1.2 Chemicals Management

Balancing society's dependence on chemicals while protecting human health and environment from the potential negative impacts of chemical pollution is imperative but challenging. There are several aspects to chemicals management, including but not limited to chemicals registration, inventorisation, assessment, authorisation, monitoring, restriction, and remediation, where appropriate. Traditionally, regulators and industry have played a big role in this area (though other stakeholders are beginning to follow suit). These efforts are typically organised at both the national and international levels.

On the national level for example, the Environmental Protection Agency (EPA) was established in the United States (US) in 1970, partly resulting from strong grassroots initiatives spurred on by the impact of Rachel Carson's *Silent Spring*. Landmark legislation in the form of the Toxic Substances Control Act (TSCA) passed after that, enabling the EPA to enforce critical industrial requirements related to chemical safety that continue to this day. In the European Union (EU), the European Chemicals Agency (ECHA) administers the prevailing regulatory mechanism known as REACH (Registration, Evaluation, Authorisation, and Restriction of Chemicals), which entered

into force in 2007. REACH promulgates the “No Data, No Market” paradigm that imposes multiple obligations on chemical manufacturers, importers, and certain downstream users.

Besides operating on a national or regional level, these aforementioned stakeholders often additionally implement and operationalise a number of international policy instruments, including conventions and frameworks that govern issues related to chemicals safety. Multiple conventions, such as the Basel, Rotterdam, Stockholm and Minamata conventions address hazardous waste, global trade of hazardous chemicals, persistent organic pollutants, and mercury, respectively. Additionally, the Globally Harmonized System of Classification and Labelling of Chemicals (GHS) was devised as a system to classify and communicate chemical hazards via labelling and safety data sheets. The Strategic Approach to International Chemicals Management (SAICM) offered what was potentially the most comprehensive global policy framework towards the sound management of chemicals and waste, with an ambitious aim to achieve this goal by 2020. However, implementation has been limited because of its voluntary nature, and SAICM’s mandate expired in 2020. Currently, member states are in negotiations to determine the future of international chemicals and waste management beyond 2020, but the outcomes of this political process are still in progress, with completion anticipated for 2023.

Nevertheless, scientific research contributing to the advancement of different aspects of chemicals management continues to be active, including in the areas of 1) chemicals assessment, and 2) routine environmental monitoring. These two aspects are both fundamental pillars of chemicals management, however in both cases, the lack of knowledge concerning chemical identities, and the inability to identify chemicals in environmental samples respectively, underpin two main obstacles confronting scientists and regulators alike. These challenges are explained in more detail below.

#### 1.1.2.1 Chemicals Assessment

The objective of chemicals assessment, also known as chemical risk assessment in some contexts, is to consider the potential hazards of chemicals on humans and the environment. In its most basic form, a chemical’s risk can be calculated as a function of two factors, hazards and exposure. Hazards encompass a chemical’s inherent potency, that is, its ability to cause harm to humans or the environment through lethal or sublethal



effects like causing disease or hindering growth and development. Within hazard assessment in the context of EU REACH,<sup>61</sup> a screening for Persistence, Bioaccumulation, and Toxicity (PBT) is typically carried out, and dose-response relationships are characterised. In the EU, the results of a hazard assessment may already trigger risk management measures, however in other jurisdictions like the US and Canada, exposure assessments are also required as part of risk assessment.

Environmental exposure refers to how much of a chemical is present within a defined environmental system, and can be calculated considering the following parameters: possible emissions in terms of chemical volumes and routes, environmental fate including transformation pathways and long-range transport potential, as well as possible degradation or distribution in the environment. These factors can be summarised for risk assessment purposes by a single value, within a given uncertainty and variability range and for a specific environmental compartment, called the Predicted Environmental Concentration (PEC) or where available, a Measured Environmental Concentration (MEC). The result of an effect assessment is represented by the Predicted No Effect Concentration (PNEC) value. Taking the ratio of PEC or MEC and PNEC gives the Risk Quotient, which is used in the process of risk characterisation, followed by risk classification according to a set of guidelines. Depending on these results, risk mitigation may be warranted through various measures, including those imposed by the competent authorities.

Throughout chemicals assessment, the availability and transparency of information on chemical identities, specifically knowledge of chemical structure, is crucial. On a fundamental level, knowledge of chemical structure enables the unambiguous identification of a given chemical between and amongst relevant stakeholders. Moreover, full information on chemical structure enables proper registration, inventorisation, and evaluation during risk assessment because a chemical's properties are influenced by its structure. For example, preliminary screening for Persistence (ready biodegradability) and Bioaccumulation (bioconcentration factors) can typically be performed using Quantitative Structure-Activity/Property Relationships (QSARs/QSPRs), which requires input of chemical structure into a model. The most well-known and commonly used models include BIOWIN, KOWWIN, BCFBAF, which are contained within the US EPA's EPI Suite<sup>62</sup> and are used for predicting the aerobic/anaerobic biodegradability, octanol-water partitioning coefficient  $K_{ow}$ ,<sup>63</sup> and fish bioconcentration factor of a chemical, respectively. Other, more recent QSAR/QSPR models include OPERA, for predicting environmental fate and physicochemical properties.<sup>64</sup> Large-scale screenings of entire

databases typically rely on these models and the availability of chemical structure for every compound.<sup>65</sup> Applying QSARs to model toxicity endpoints is also common, for example, via ECOSAR contained within EPI Suite. Besides predicting structure-activity or structure-property relationships, regulatory prioritisation and restriction exercises may be performed based on filtering of chemical lists for specific chemical substructures, for example the presence of certain functional groups, as in the case of the recent restriction of bisphenols by ECHA,<sup>66</sup> which would be impossible to achieve without knowledge of chemical structure.

However, the availability of chemical structure information within the risk assessment process is not a given. From a compound registration perspective, there is often limited transparency in chemical structure information because manufacturers may be exempt from such disclosures due to Confidential Business Information (CBI) clauses that were claimed to protect the intellectual property of chemical companies as a form of competitive advantage. A recent study of about 20 national/regional chemical inventories found that over 50,000 chemicals or mixtures have their identities protected by CBI clauses.<sup>1</sup> Alternatively, chemical products that are challenging to characterise may also be exempted from non-ambiguous structure disclosure requirements in a *de facto* manner; mixtures of polymers typically fall into this category, as do substances of Unknown or Variable composition, Complex reaction products, or Biological materials (UVCBs),<sup>67</sup> which include a wide range of chemical products of both natural and synthetic origin, including essential oils, petroleum fractions and chemical reaction intermediates. This lack of non-ambiguous structural information has negative implications for their risk assessment: without available structure information, traditional QSARs cannot be performed to predict the PBT properties of these substances. Thus, these substances are typically excluded from these *in silico* assessments, or would otherwise have to undergo empirical testing, which is a time-consuming, resource-intensive process.

In the absence of chemical structure availability, alternative identification methods must be used to represent chemical identities. These are described in detail below.

#### 1.1.2.1.1 Chemical Identity and Data Representations

Generally, there exist several ways of representing chemicals, irrespective of whether structural information is available. In environmental chemistry, particularly with regards to chemicals assessment, nomenclature in the form of chemical identifiers are commonly used to represent compounds and typically consist of alphanumeric

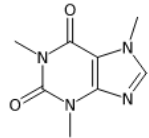
strings of characters (Table 1). These chemical identifiers are one-dimensional, and inherently do not convey any meaning, let alone structural information. Therefore, these identifiers can be generated in the absence of structural information and may instead represent administrative, non-chemistry-related information such as the sequence or status of chemical registration. For example, European Community Numbers (EC No.) beginning with '2' or '3' represent substances that were recorded as commercially available between 1971 and 1981 in the European Inventory of Existing Chemical Structures (EINECS), while those beginning with '4' represent substances that were 'new' and registered under the European List of Notified Chemical Substances (ELINCS) from 1981 onwards.<sup>68</sup> Importantly, such identifiers tend to require look-up in a specific reference database that may be associated with a regulatory framework or industrial sector, for example Chemical Abstract Service Registry Numbers (CASRN) in relation to the US EPA Toxic Substances Control Act (TSCA) Inventory,<sup>69</sup> or the COSIng database for the cosmetics industry respectively.<sup>70</sup> As there is no one universal identifier for all chemicals, it is common for chemical dossiers to contain a list of multiple identifiers for the same chemical to ensure its identifiability across e.g., different databases and toxicological studies.

Notably, chemical identifiers are not the only ways to represent chemical identity. Line notations like IUPAC name, trivial name, Simplified Molecular Input Line Entry (SMILES) strings, International Chemical Identifiers (InChI), and bit vectors representing chemical fingerprints are typical two-dimensional representations used to encode structures. These representations are independent of database origin. Depending on the type of representation, different levels of detail regarding chemical structure can be encoded, for example stereochemistry, charge, conformer state *etc.* Table 1 summarises these representations. In the context of risk assessment, SMILES and InChIs are commonly found in chemical dossiers submitted for evaluation.

Fundamentally, chemical representations facilitate the ability to exchange chemical information. However, it is important to note that converting a physical entity - chemicals - to such representations inevitably results in information loss; in practical terms, this means no chemical representation is perfect. Nevertheless, as a chemical representation is essentially the currency used to exchange, convert, and communicate chemical information, the type of representation chosen is integral to

the performance and efficiency of such exchanges, and in turn, the tasks at hand. For example, an identifier that is short and unique may be preferred for indexing or querying chemical databases, such as a numeric identification number or code like EC No.. In contrast, performing tasks that require computing, for instance substructure searches or building toxicity models based on QSARs, requires structures, thus making representations that contain structural information such as SMILES or InChIs more suitable.

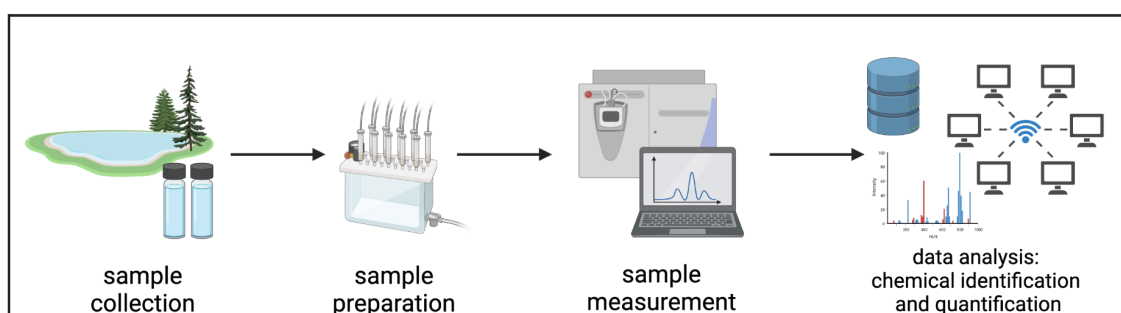
**Table 1. Common types of representations of chemicals used in environmental cheminformatics.** All formats are open unless specified as proprietary. This is a non-exhaustive list.

Common Name	Origin	Characteristics/Properties	Suitable for Querying Databases	Contains Information on Chemical Structure Without Requiring Lookup	Example for Caffeine
Skeletal Structure (Bond-line)	convention	Image	no	yes	
Simplified Molecular Input Line Entry System (SMILES)	Weininger <sup>71</sup>	Alphanumeric; line notation; widely parseable; some flavours proprietary	yes (if canonicalised)	yes	<chem>CN1C=NC2=C1C(=O)N(C(=O)N2C)C</chem>
International Chemical Identifier (InChI)	InChI Trust <sup>72</sup>	Alphanumeric; line notation	no	yes	InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
International Chemical Identifier Key (InChIKey)	InChI Trust <sup>72</sup>	Alphanumeric; line notation; hashed version of InChI; identifier	yes	no	RYYVLZVUVIJVGH-UHFFFAOYSA-N
Chemical Abstracts Service Registry No. (CASRN)	Chemical Abstracts Service	Alphanumeric; identifier; proprietary	yes	no	58-08-2 (deprecated:

<b>Common Name</b>	<b>Origin</b>	<b>Characteristics/Properties</b>	<b>Suitable for Querying Databases</b>	<b>Contains Information on Chemical Structure Without Requiring Lookup</b>	<b>Example for Caffeine</b>
					71701-02-5, 95789-13-2)
European Community No. (EC No.)	European Union	Numeric; identifier	yes	no	200-362-1
CompTox DTXSID	EPA CompTox Chemicals Dashboard	Alphanumeric; identifier	yes	no	DTXSID0020232
PubChem Compound Identification Number (CID)	PubChem	Numeric; identifier	yes	no	2519
Bit vectors (cheminformatics fingerprints)	NA	Multiple implementations in various cheminformatics toolkits	yes	yes	-

### 1.1.2.2 Environmental Monitoring

Another area of chemicals management that continues to be the subject of active research is environmental monitoring, specifically, measuring and detecting chemical compounds in environmental samples. The environment consists of multiple compartments, also known as environmental media or matrices, for example soil, sediments, sludge, air, and aquatic systems, aqueous and marine. Each environmental matrix has its own distinctive physico-chemical properties and thus requirements with respect to performing chemical analysis. Nevertheless irrespective of the matrix, environmental analytical chemists typically undertake the following sequence of steps, (Fig. 1): sample collection, preparation, measurement, followed by data analysis to obtain results.



**Figure 1. Typical workflow for environmental monitoring of chemicals.** Created with BioRender.com.

In recent years, advances in laboratory instrumentation, particularly in liquid chromatography high-resolution mass spectrometry (LC-HRMS), have enabled environmental analytical chemists to detect and measure organic compounds at concentrations as low as picograms per litre of aqueous solution. LC-HRMS systems typically comprise two parts: the first is liquid chromatography, which involves the separation of chemical compounds according to differences in their physicochemical properties, like polarity, within a chromatographic column. This column is usually filled with porous packing material such as silica or resin that forms the stationary phase, which comes into contact with the analyte dissolved in a liquid mobile phase that is forced through the column at high pressure. In normal phase chromatography, the stationary phase is polar and the mobile phase is non-polar, meaning that non-polar analytes are the first to elute from the column. Conversely, polar analytes are the first to elute in reverse phase chromatography that consists of a non-polar stationary

phase, and a polar mobile phase. The order of elution determines the retention times,  $t_R$ , of the analytes, which is a property specific to a given chromatographic system and method.

Once eluted, the analytes enter the mass spectrometer where they become ionised, accelerated, and separated according to their mass-to-charge ratios ( $m/z$ ). There are multiple types of mass analysers that perform this task, including but not limited to low resolution quadrupoles and ion traps, to high resolution Orbitraps. A mass detector then records their masses, forming a mass spectrum (MS1) of signals of varying intensity; depending on the instrument, the mass of the charged molecule can be accurately measured up to four decimal places. In tandem mass spectrometry, which is commonly used for structure elucidation, the parent ion whose mass was recorded in the MS1, also referred to as a precursor ion in LC-HRMS, is then fragmented once more by a collision source, and the masses of these ionised molecular fragments, also known as fragment ions, are recorded in a fragment mass spectrum (MS2). These four pieces of information recorded using LC-HRMS ( $t_R$ ,  $m/z$ , MS1 spectrum, and MS2 spectrum) are then used as inputs for compound identification.

#### 1.1.2.2.1 Compound Identification Workflows for Identifying Environmental Chemicals using LC-HRMS

Thousands of analytical signals are routinely detected in environmental samples, which reflects the fact that virtually all these samples comprise complex chemical mixtures. Compound identification workflows typically follow one of three approaches: target, suspect, and non-target. Krauss and colleagues were amongst the first to describe these workflows in their seminal paper,<sup>73</sup> which are discussed in further detail below.

Target compounds are those that the analyst is certain are present in the sample and have previously been well characterised using reference standards. As such, library spectra and retention time values are available, making these compounds relatively easy to identify through matching the properties of the detected analytes with these references. Well-documented target compounds in environmental samples include common ubiquitous pesticides and certain pharmaceuticals. These compounds



typically make up a small fraction of all the signals detected, unlike the majority of signals measured in a given sample. In fact, the majority of signals represent Unknowns, which manifest in three forms: Unknown Knowns, Known Unknowns, and Unknown Unknowns. This paradigm is summarised by the so-called Rumsfeld's Matrix (Table 2), and has been used in multiple studies to conceptualise the respective compound identification approaches used.<sup>74-77</sup>

**Table 2.** Different types of environmental unknowns, classified according to the Rumsfeld's Matrix paradigm.

	<b>Knowns</b>	<b>Unknowns</b>
<b>Known</b>	Known <b>Knowns</b> <i>Target compounds</i>	Known <b>Unknowns</b> <i>Non-target compounds</i>
<b>Unknown</b>	Unknown <b>Knowns</b> <i>Suspect compounds</i>	Unknown <b>Unknowns</b> <i>Non-target compounds</i>

Suspect compounds, also referred to as Unknown Knowns, are the next most challenging chemicals to identify in environmental samples after targets. These compounds are 'known' because they are expected or likely to be present in samples based on domain knowledge, and are screened within environmental samples using  $m/z$ , usually in the absence of reference standards. For example, environmental samples obtained near industrial areas likely contain certain compounds specifically used in those industries that are then emitted into the surrounding environment. Alternatively, the presence of a compound in a chemical registry indicating local production and/or use is another example of probable suspect. To conduct suspect screening,<sup>78</sup> researchers typically use suspect lists - lists of chemicals that have been curated by researchers with specific interests in particular groups of chemicals. These lists tend to be thematically related to the researchers' project, their domain knowledge of a particular environment and the compounds expected to be present there, or are regional in scope and may be derived from local chemical registries or other regulatory sources of information. If a match for a particular  $m/z$  is found, the experimental MS2 spectrum can then be compared with an existing MS2 spectrum in a spectral database like MassBank Europe,<sup>79</sup> MassBank of North America (MoNA),<sup>80</sup> NIST Mass Spectral libraries,<sup>81</sup> or an in-house library. If not, compound databases

such as PubChem, ChemSpider, or the US EPA CompTox Chemicals Dashboard are consulted. However these databases are relatively large - containing millions of compounds - and general in scope, and thus may introduce greater risk of false positives during screening, because the chances of obtaining a matching mass are relatively high the bigger the database used for screening becomes. The use of so-called suspect lists instead of entire compound databases therefore offers a “screen smart, not screen big” approach.<sup>82</sup>

Notably, over time and through intensive chemical data curation efforts, many suspect lists have become shared, open resources and *de facto* environmental chemistry databases as part of the the Suspect List Exchange (SLE), an initiative of the NORMAN Network that comprises over 80 reference laboratories and research centres around the world.<sup>82,83</sup> The NORMAN-SLE has since become a widely-used resource for open environmental chemical information, and currently hosts these suspect lists based on the FAIR data principles - findability, accessibility, interoperability, and reusability. At time of writing, there exist 99 suspect lists covering a range of themes, for example, transformation products, national pesticide lists, algal toxins, chemicals associated with plastic packaging *etc.*

The most challenging of environmental unknowns are still the non-target compounds, which tend to form the majority of signals in a given environmental sample. These unknown compounds remain unidentified but may be so well recognised through repeated detection of their analytical signals that they are “Known Unknowns”. For example, HPV surfactants forming homologous series of compounds are frequently detected in the environment, particularly in wastewater samples, but remain difficult to identify because of current technological limitations in the ability to match their analytical signals to compounds in databases. Alternatively, researchers may not even be aware of their lack of knowledge of these compounds, and are hence also referred to as “Unknown Unknowns”. No *a priori* information is available concerning the chemical identities of these unknown compounds, which makes their identification extremely difficult considering the vastness of chemical space.

Numerous open software packages and tools for environmental non-target analysis have been developed in recent years, reflecting diverse approaches to non-target identification. For example, MetFrag is an *in silico* fragmenter that predicts the

fragmentation spectrum of chemicals in a provided database based on bond dissociation energies, and scores how well the predicted fragment spectrum matches an experimental MS2 spectrum of interest.<sup>84</sup> The presence or absence of a candidate on a suspect list or database can determine MetFrag's final score and thus the candidates suggested for a given unknown.<sup>85</sup> SIRIUS takes a different approach by predicting the molecular formula of a compound based on its isotope patterns.<sup>86</sup> Fragmentation trees are then generated for each molecular formula, followed by conversion to molecular fingerprints that are used to search compound databases as part of structure elucidation. Another approach is that of CFM-ID's, where all theoretically-possible fragments are generated for a given molecular structure and assigned probabilities based on model outputs that were trained on known molecules and their MS spectra.<sup>87</sup>

Irrespective of the approach used, identification of chemical compounds is typically reported using an established level scheme.<sup>88</sup> According to the scheme, identification confidence levels range from Level 1 (confirmed structure by reference standard), to Level 5 (exact mass of interest). For chemical identification to be conclusive *i.e.*, for an identification to be considered 'confirmed' at Level 1, a minimum of three orthogonal pieces of information and/or confirmation using a measured internal standard are required. Level 2 identifications are considered probable structures, either because there is an unequivocal match with library or literature spectra (Level 2a), or by diagnosis *i.e.*, the experimental information available does not reasonably fit any other structure, but the identification cannot be confirmed for lack of literature or standard references (Level 2b). Tentative candidates are described by Level 3 confidence, where multiple plausible structures could be represented by the available evidence, whereas Levels 4 and 5 correspond to having just the unequivocal molecular formula or exact mass respectively.

### 1.1.3 Research Gaps in Identifying and Managing Chemical Unknowns

Overall, chemical unknowns are rife within the fields of analytical environmental chemistry as well as regulatory chemicals management. More specifically, information on the structures of these chemical unknowns is lacking, which generally

manifests in two ways. Firstly, their structures are ambiguous or simply unavailable in the public domain. These phenomena present multiple obstacles from a chemicals assessment point of view in terms of screening and hazard assessment, which in turn potentially weakens our capacity to safely manage chemicals, including through restriction or possible mitigation measures. Secondly, bottlenecks in identifying chemical unknowns in environmental samples persist; the number of chemical unknowns detected in measurements of a typical environmental sample far outnumber those that are known or can be identified with reasonable confidence. As a result, our ability to identify chemicals in environmental samples remains far from comprehensive, which compromises the effectiveness of environmental monitoring.

This pernicious knowledge gap of chemical unknowns bears consequences for both humans and the environment. In terms of the onset of human diseases, many of which are posited to stem from both genetic and environmental factors, our understanding is severely hampered, in part because of our incomplete knowledge of the chemical exposome,<sup>89,90</sup> defined as the totality of human exposures to chemicals over time and space. Furthermore, without knowledge of the structures of chemicals in our environment, mechanistic understanding of their impacts on the environment, ecological systems, and stressors remain limited. For example, exposure to toxic chemicals released by a dumping event in the River Oder is suspected to be the reason behind the recent death of thousands of fish, but the identities of these chemicals, and the nature of their toxic effects is still unknown.<sup>91</sup>

These aforementioned knowledge gaps represent urgent areas for research. With the scale, diversity, and number of open chemical information resources rapidly growing, the possibility to leverage them in workflows and approaches to tackle the problems of identifying environmental unknowns is highly warranted. On both conceptual (chemical representation) and practical (different data formats, accessibilities, application programming interfaces *etc.*) levels, the challenges of integrating myriad chemical information resources is not trivial, but are certainly tractable. Furthermore, the increased sharing of environmental chemistry data has opened up new avenues for the development of cheminformatics approaches, such as data analysis algorithms that may eventually support and inform further workflow development.

## 1.2 Aims

In light of the above, this dissertation addresses three main aims in identifying and managing unknown chemical pollutants in the environment:

1. Use Open chemical resources (software, databases, and tools) within computational workflows to enhance the identification of unknown compounds in the environment.
2. Enumerate and review the specific regulatory assessment challenges, including those posed by limited structural information, concerning substances of Unknown or Variable composition, Complex reaction products, or Biological materials (UVCBs).
3. Develop cheminformatics methods to support the identification of UVCBs in environmental samples.

## 1.3 Scope of the Dissertation

This work addresses the three aims mentioned above as a cumulative dissertation comprising seven chapters, four of which are articles that have been published or submitted for publication in peer-reviewed journals:

**Chapter 2** - Developing an open computational workflow using emerging digital chemistry resources for non-target analysis

Related publication: *Lai, A., Singh, R. R., Kovalova, L., Jaeggi, O., Kondić, T. & Schymanski, E. L. Retrospective non-target analysis to support regulatory water monitoring: from masses of interest to recommendations via in silico workflows. Environ Sci Eur 2021, 33, 43, 1-21. DOI: 10.1186/s12302-021-00475-1*

**Chapter 3** - Data mining transformation product information for enhanced suspect screening

Related publication: *Singh, R. R., Lai, A., Krier, J., Kondić, T., Diderich, P. & Schymanski, E. L. Occurrence and Distribution of Pharmaceuticals and Their Transformation Products in Luxembourgish Surface Waters. ACS Environ Au 2021, 1, 1, 58–70. DOI: 10.1021/acsenvironau.1c00008*

**Chapter 4** - Tackling the next frontier of environmental unknowns - UVCBs

Related publication: *Lai, A., Clark, A.M., Escher, B. I., Fernandez, M., McEwen, L. R., Tian, Z., Wang, Z., & Schymanski, E. L. The Next Frontier of Environmental Unknowns: Substances of Unknown or Variable Composition, Complex Reaction Products, or Biological Materials (UVCBs). Environ Sci Technol 2022, 56, 12, 7448–7466. DOI: 10.1021/acs.est.2c00321*

**Chapter 5** - A cheminformatics algorithm for improved identification of homologous series in environmental mixtures

Related publication: *Lai, A., Schaub, J., Steinbeck, C., & Schymanski, E. L. An Algorithm to Classify Homologous Series within Compound Datasets. J Cheminform 2022, 14, 85. DOI: 10.1186/s13321-022-00663-y*

**Chapter 6** - Discussion of the strengths and limitations of the results achieved in this dissertation in relation to the aims

**Chapter 7** - Conclusions and perspectives, including outlining future avenues for research

## Chapter 2

# Developing an Open Computational Workflow using Emerging Digital Chemistry Resources for Non-target Analysis

Analysis of non-target compounds in environmental samples is extremely challenging because of the lack of information on possible chemical identities - all that is available are the analytical signals measured in the environmental sample. Despite this challenge, regulators may be interested in identifying these unknowns beyond their routine environmental monitoring of targets, so as to enhance their environmental screening activities and proactively monitor substances of potential concern to human health and environment. However, besides the already daunting challenge of non-target analysis, regulators working on a regional scale must typically deal with large amounts of data that were collected from the multiple sites within their geographical mandate, which adds to the urgent requirement for feature prioritisation in identification workflows.

In this work, data measured from Swiss environmental samples collected near wastewater sites by collaborators from the Zurich Office of Waste, Water, Energy, and Air (*Zürcher Amt für Abfall, Wasser, Energie und Luft*) were retrospectively analysed with the ultimate goal of non-target identification. To address this challenge, an open computational workflow was developed comprising two main novel aspects: (1) a pre-screening and quality control step for prioritising suitable non-target masses for identification, and (2) an identification pipeline that exploited emerging environmental chemistry resources using the tool MetFrag. The former features an algorithm that performs 6 automatic Quality Control steps of tandem mass spectrometry data (now integrated into the R package *ShinyScreen*), including checks for peak intensity and alignment, to ensure their suitability for non-target identification; these steps represent the typical logic that environmental analytical chemists apply when manually inspecting data prior to non-target analysis. As mass spectral features that did not meet the quality criteria were discarded from further consideration, this

pre-screening is effectively a form of prioritisation that may represent first steps towards non-target analysis in routine environmental monitoring. The latter leverages MetFrag, an *in silico* fragmentation tool as a platform for integrating 'environmental metadata' into non-target identification workflows based on chemical information newly provided by Swiss, Swedish, EU-wide, and American regulators.

Overall, 21 compounds were tentatively identified with Level 3 confidence, and one with Level 1. The tentative identifications were communicated using transparent breakdowns, analysis, and interpretation as justification of MetFrag's results, with the intention of guiding regulators in their next steps, for example in devising future sampling campaigns.



## Publication A

### Retrospective non-target analysis to support regulatory water monitoring: from masses of interest to recommendations via in silico workflows

Lai, A.<sup>1</sup>, Singh, R. R.<sup>2</sup>, Kovalova, L.<sup>3</sup>, Jaeggi, O.<sup>4</sup>, Kondić, T.<sup>5</sup> & Schymanski, E. L.<sup>6</sup>

*Environ Sci Eur* 2021, 33, 43, 1-21

DOI: 10.1186/s12302-021-00475-1

Reproduced with permission from Springer Nature.  
Article is open-access, distributed under CC-BY licence.





<b>Author Contributions</b> (Underlined numbers refer to PhD students)						
<b>Author No.</b>	<b><u>1</u></b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>
Conceptual Research Design	x	x	x	x		x
Planning of Research Activities	x	x	x	x		x
Reviewing the Tools	x					x
Data Collection		x	x	x		
Data Analysis & Interpretation	x	x			x	x
Manuscript Writing	x	x	x	x		x
Suggested Publication Equivalence Value	1.0					

RESEARCH

Open Access



# Retrospective non-target analysis to support regulatory water monitoring: from masses of interest to recommendations via in silico workflows

Adelene Lai<sup>1,2\*</sup> , Randolph R. Singh<sup>1,3</sup> , Lubomira Kovalova<sup>4</sup>, Oliver Jaeggi<sup>4</sup>, Todor Kondić<sup>1</sup>  and Emma L. Schymanski<sup>1\*</sup> 

## Abstract

**Background:** Applying non-target analysis (NTA) in regulatory environmental monitoring remains challenging—instead of having exploratory questions, regulators usually already have specific questions related to environmental protection aims. Additionally, data analysis can seem overwhelming because of the large data volumes and many steps required. This work aimed to establish an open in silico workflow to identify environmental chemical unknowns via retrospective NTA within the scope of a pre-existing Swiss environmental monitoring campaign focusing on industrial chemicals. The research question addressed immediate regulatory priorities: identify pollutants with industrial point sources occurring at the highest intensities over two time points. Samples from 22 wastewater treatment plants obtained in 2018 and measured using liquid chromatography–high resolution mass spectrometry were retrospectively analysed by (i) performing peak-picking to identify masses of interest; (ii) prescreening and quality-controlling spectra, and (iii) tentatively identifying priority “known unknown” pollutants by leveraging environmentally relevant chemical information provided by Swiss, Swedish, EU-wide, and American regulators. This regulator-supplied information was incorporated into MetFrag, an in silico identification tool replete with “post-relaunch” features used here. This study’s unique regulatory context posed challenges in data quality and volume that were directly addressed with the prescreening, quality control, and identification workflow developed.

**Results:** One confirmed and 21 tentative identifications were achieved, suggesting the presence of compounds as diverse as manufacturing reagents, adhesives, pesticides, and pharmaceuticals in the samples. More importantly, an in-depth interpretation of the results in the context of environmental regulation and actionable next steps are discussed. The prescreening and quality control workflow is openly accessible within the R package ShinyScreen, and adaptable to any (retrospective) analysis requiring automated quality control of mass spectra and non-target identification, with potential applications in environmental and metabolomics analyses.

**Conclusions:** NTA in regulatory monitoring is critical for environmental protection, but bottlenecks in data analysis and results interpretation remain. The prescreening and quality control workflow, and interpretation work performed here are crucial steps towards scaling up NTA for environmental monitoring.

\*Correspondence: adelene.lai@uni.lu; emma.schymanski@uni.lu

<sup>1</sup> Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 Avenue du Swing, 4367 Belvaux, Luxembourg  
Full list of author information is available at the end of the article

**Keywords:** Non-target analysis, Suspect screening, Retrospective, Wastewater, Micropollutants, Cheminformatics, Identification, Monitoring, Regulation

## Background

Organic pollutants are well-documented in aquatic environments [59]. Traditionally, target strategies that look for chemicals known in advance have been used to identify these compounds [27]. In contrast, non-target analysis (NTA) helps discover previously undetected, unexpected and/or unknown substances. NTA has been under intense development in recent years, aided by advances in instrumentation and computational approaches [17, 27]. Considering the vast chemical space of possible environmental pollutants [65], the need for NTA is becoming more pressing in order to tackle the growing challenge of identifying chemical unknowns in samples. Yet, data analysis in NTA remains a formidable challenge. To ease the “identification burden” in NTA, simplifying approaches like Suspect Screening, where chemicals on discrete lists suspected to be present in the sample are screened, are being taken in the interim [17].

Various successful examples of NTA [1, 4, 5, 19, 28, 50, 53, 60] have inevitably encouraged interest in its potential role to monitor and manage chemical pollutants in the environment [17]. As the field matures, there is some consensus that NTA is “Ready to Go”, with calls for it to be applied more widely within the regulatory frameworks of local, regional, and national authorities [17, 18]. Data-mining routines like *enviMass* have contributed to such initiatives [34]; *enviMass* facilitates NTA by peak-picking and prioritising unknown features of interest worthy of further identification efforts. It does so by connecting mass spectral features based on criteria such as having signals of sufficient intensity, grouping together isotopologues and adducts of the same component, and detecting temporal trends, ultimately giving as output a list of *m/z*-retention time pairs, plus accompanying information for further identification efforts.

However, challenges for regulators to perform NTA persist, particularly with respect to high-throughput data analysis and identification following the mass prioritisation and peak-picking steps described above. For example, regulators may lack specific NTA expertise and/or resources to apply the potentially many and complicated computational workflows [15, 33] available for analysing the copious amounts of data. In addition to the time-consuming and complex nature of data interpretation, issues related to standardisation and reproducibility exist, as there is currently no ‘one size

fits all’ approach to identifying compounds using NTA [16]. As a result, NTA is currently often considered by regulators as “too much effort for too little sound evidence”.

Another more systemic obstacle to applying NTA in a regulatory context relates to the divergent interests of scientists in academia, who are (currently) responsible for driving most NTA developments, and scientists in regulatory practice, who would implement these developments towards regulatory compliance and environmental protection. While the former aim often to develop and publish novel work, the primary mandate of the latter is regulatory compliance towards environmental protection. One possible consequence of this reality is that academic research outcomes resulting from NTA may not be directly relevant or in a form that is readily usable for regulators. In other words, researchers’ questions may not be regulators’ questions—what is possibly *scientifically* interesting may not be of priority or directly useful to regulators.

Despite these aforementioned challenges, it is possible (and important) to navigate both research and regulatory needs in NTA. The present work is an example of academic research driven primarily by regulatory priorities. In this “top-down” approach, pre-existing data were used to generate results of direct environmental relevance and with immediate implications for environmental management.

Three practical challenges characteristic of applying NTA in a regulatory environmental monitoring context arose in this study: (i) the study was framed by superlative questions that required a large volume of data to be analysed, i.e. identify unknown compounds occurring at the *highest intensities* and *highest temporal frequency* with *point sources* across all the samples of the sampling campaign; (ii) there was a strict and limited timeframe allowed for the study following project management procedures of the regulatory body, and (iii) the data originally collected had been repurposed for this NTA study as there was no capacity nor further resources available within the scope of the project to do additional measurements. The latter point was all the more critical as preliminary manual inspection of the available data revealed that not all measurements were fully suitable for the intended non-target identification. These challenges called for a high-throughput approach capable of processing large volumes of data of variable quality in a fast and reproducible way that would be compatible with

identification approaches downstream. Additionally, unlike the seemingly increasing complexity of existing workflows [33], an uncomplicated and ‘minimal, bare-bones’ but fully functional approach that is transparent and easily explainable is critical given the regulatory context.

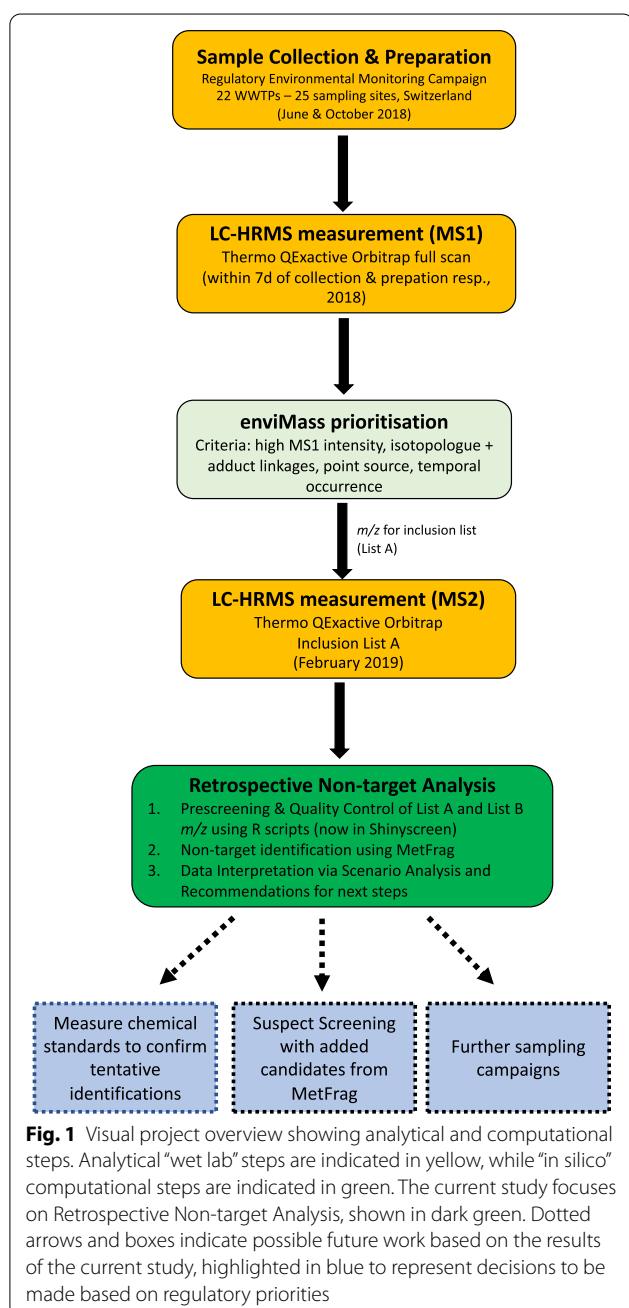
MetFrag, used in this work to support identification efforts, is an example of an open in silico identification approach which satisfies the aforementioned criteria. Released in 2010 [68], it first retrieves potential candidates with matching mass from compound databases such as PubChem [23] (111 million chemical structures, August 2020), ChemSpider [7, 48] (103 million chemical structures, February 2021), or smaller biological databases like the Human Metabolome Database [67], 20 (114,304 metabolites, February 2021). These candidates are then scored according to how well the experimental spectrum matches the in silico fragments generated per candidate using a bond dissociation approach [68], and subsequently ranked according to this FragmenterScore (sometimes referred to as the Fragmentation Score or FragScore, or simply the MetFrag Score when it is the only component thereof). For the identification of environmental “known unknowns”, using fragmentation information alone in this way can give mediocre results (e.g., ~22 and 6% of 473 environmentally relevant standards ranked first with ChemSpider and PubChem, respectively [51]). This outcome may have various causes: (i) the search databases used are too large and/or do not contain only environmentally relevant compounds, therefore resulting in too many candidates that are not meaningful, and/or (ii) there is simply not enough information to distinguish candidates when considering their fragmentation alone.

To address these limitations, MetFrag was ‘relaunched’ in 2016 to incorporate further identification strategies beyond fragmentation, such as retention time information, substructure in/exclusion, availability of literature and patent information, presence/absence in suspect lists, and user-defined scoring terms [51]. Over time, spectral similarity comparison with spectra from the MassBank of North America (MoNA) (Fiehn [12] with and without a MetFusion approach [14] was also integrated into MetFrag. Since then, two further open-science/environmental chemistry developments have contributed significantly to MetFrag’s extended capabilities for identifying environmental unknowns. Firstly, the release and integration of the United States Environmental Protection Agency’s CompTox Chemicals Dashboard [66] (hereafter, “CompTox”) into MetFrag provides a search database of >850,000 compounds of environmental and toxicological relevance [54], while allowing users to leverage the “MS-Ready” concept [37] and various forms of chemical

metadata availability in CompTox as user-defined scoring terms. Secondly, critical information from international regulatory bodies can now be exploited through MetFrag towards identifying environmental chemicals. Beyond (i) the US EPA’s Chemicals and Products database (CPDat) ([62, [10] and other CompTox-related metadata terms that are already integrated via CompTox, MetFrag’s user-defined scoring terms can also be configured to incorporate information such as (ii) hazard and exposure from the Swedish Chemicals Agency KEMI [13], (iii) European chemicals registration, *i.e.* REACH [2], and (iv) the NORMAN Network’s merged suspect list of chemicals of emerging concern known as SusDat (NORMAN [43] representing knowledge gathered from NORMAN members, which include >70 regulatory and academic reference laboratories throughout the world, as well as external contributions. Used in this way, MetFrag connects disparate resources from various regulatory agencies and academic researchers towards identifying environmental unknowns, practically ‘helping researchers and regulators help each other’ by providing an interconnected information platform with identification functionality.

Since MetFrag’s relaunch in 2016, work on the identification of environmental unknowns has used MetFrag’s post-relaunch functionality to varying extents. Some research simply uses MetFrag purely for its in silico fragmentation capabilities, *i.e.* not paired with any compound database [9, 40, 49]. Many examples use only the FragmenterScore to rank candidates retrieved from ChemSpider alone [3, 31, 35], PubChem alone [29, 61, 64], or a combination of either or both with other databases [8, 25, 45, 47] like KEGG [22], FOR-IDENT [30] and MassBank [36]. Several studies have begun to use one or more of MetFrag’s post-relaunch capabilities such as data source, patent, and/or reference counts for the respective compound database used [4, 5, 11, 39, 41, 42, 63], spectral library similarity [4, 5, 11, 21, 63], and presence in suspect lists [5, 28, 41]. Albergamo and colleagues [1] were amongst the first to use MetFrag’s post-relaunch capabilities heavily, in particular those provided via CompTox and by international regulators and scientists.

The present work aimed to exploit “post-relaunch” MetFrag and Open Science developments towards retrospectively identifying non-target environmental pollutants in a regulatory context, as summarised in Fig. 1. Here, pollutants determined to be of regulatory concern by regulators originating from industrial activities found in Swiss wastewater treatment plant (WWTP) effluents were the main subjects of this study, which focused on developing the open in silico workflow to identify them. A prescreening and quality control workflow for high-throughput automated data processing was developed



to analyse a provided list of unknown *m/z* prioritised by enviMass. The use of MetFrag in this work leverages the state-of-the-art open resources mentioned above, chief among them, regulatory information from multiple international sources, in addition to exploiting many of MetFrag’s post-relaunch capabilities. The identifications provided by MetFrag were analysed with respect to the specific environmental regulatory context of this study and communicated using an established system of confidence levels, discussed in detail in the next section.

## Methods

Daily water samples were collected from 25 sites based at 22 WWTPs distributed across Switzerland within sampling campaigns focusing on point sources of industrial chemicals. Of these 25 sampling sites, 19 correspond to WWTP effluents (*i.e.*, 1 site per WWTP), while 6 constitute paired influent and effluent sampling sites of 3 WWTPs (*i.e.*, 2 sites per WWTP) which employ ozonation. The effluent from these 3 WWTPs employing ozonation came from secondary clarifiers. Five sites were sampled twice each (in June and October 2018, respectively), while 20 were sampled only once (June 2018), giving a total of 30 samples.

During each sampling campaign, 2 L of the 24-h flow-proportional composite samples were collected daily at each sampling site over seven consecutive days. The sample was filled into two 1-L glass bottles and kept closed at 4 °C until the last day of the respective sampling campaign. That day, all samples were transported cooled to an analytical laboratory and were filtered, flow-proportionally mixed, and sent cooled for MS-analysis. The final samples used for measurement were flow-proportional 7-day composites.

## Sample measurement

Prior to analysis, samples were filtered through a glass fibre filter and isotopically labelled internal standards were added (26 for positive and 7 for negative ionisation mode, respectively). Samples were analysed without enrichment by direct injection of 100 µl into the chromatographic system. Chromatographic separation of the analytes was performed using a Waters Atlantis T3 column (150 × 3 mm, 3 µm particle size) connected to a Thermo Scientific Accela liquid chromatography system equipped with a 1250 pump, open autosampler, and Thermo Scientific Column Oven 300. The mobile phase eluent A consisted of ultrapure water (ELGA LabWater Purelab Ultra from Labtec Services AG, 5 mM ammonium formate), while eluent B consisted of LC-MS grade methanol (Scharlau Chemie S.A, 5 mM ammonium formate). The gradient programme started with 10% B, which was kept for 1 min before a linear ramp to 95% B for 12 min. This condition was kept for 5 min before returning to starting mobile phase conditions at 18.5 min. The column was re-equilibrated for 4.5 min giving a total run time of 23 min with a flow rate of 300 µl/min.

A full-scan single MS measurement was performed using a Thermo Scientific QExactive Orbitrap LC/MS system with resolving power of 70,000 (at *m/z* = 200) within 7 days of sample collection and preparation. A scan range of 100 to 1000 was used in both positive and negative electrospray ionisation modes. A heated



electrospray ionisation (HESI) source with a vapouriser temperature of 350 °C, sheath gas flow of 35 arbitrary units (au), auxiliary gas flow of 10 au, spray voltage of 3400 V (positive) and 3000 V (negative), S-lens level of 50, and capillary temperature of 270 °C was used. The samples were then stored at 4 °C.

Following the prioritisation of non-target masses (described in Part 1 of the prescreening workflow of the next section), the resulting list of non-target masses formed the inclusion list for MS2 measurements of the same samples in data-dependent acquisition mode in February 2019. Normalised collision energy of 35 was used. The same measurement protocol as described above was applied with resolving power of 17,500 (at  $m/z=200$ ).

## Computational methods

### Part 1—enviMass prioritisation of masses of interest

enviMass (v.3.5, [34]) was used to prioritise non-target masses of interest based on the following criteria: high-intensity MS1 peaks (used as a proxy for high concentration), presumed point source (occurring at one or only a few sampling sites), multiple temporal occurrences across the sampling campaign, *i.e.* high-frequency occurrences, and existing isotopologue and adduct linkages. Initially, a list of 300 non-target masses of interest was identified and used as an inclusion list for MS2 acquisition in the second round of measurements in February 2019 using the same samples that had been stored at 4 °C as described above. Of these 300 masses, 125 masses with associated  $[M+H]^+$  and  $[M-H]^-$  information from enviMass (117 and 8, respectively) were considered for further processing in the next step and constituted “List A”. A further 60 masses with associated  $[M+H]^+$  and  $[M-H]^-$  information (28 and 32, respectively) were also considered for the next step (“List B”), but had not been measured as part of the inclusion list. The enviMass parameters used to derive Lists A and B are detailed in

the SI. These lists were the starting point for the workflows described here.

### Part 2—prescreening and quality control workflow

Data files in .RAW format were first converted to .mzML format using MSConvert from Proteowizard (v.3.0.19182-51f676f6be, [6]), with full settings available in the SI (Additional file 1: Figure S1). The data were preliminarily inspected manually using XCalibur Qual Browser (v.4.2.28.14, Thermo Fisher Scientific, Waltham MA, USA). Then, a workflow to extract, prescreen, and quality control the spectra of the precursor masses in Lists A and B was developed and performed prior to further identification efforts.

The prescreening workflow first extracts all MS1 and MS2 ion chromatograms of each  $m/z$  from each *mzML* file supplied to it as input. No post-processing of mass spectral features such as peak removal, filtering, or scaling is performed whatsoever during the extraction of spectra. Extracted MS1 precursors whose retention times are within 2 min of the mean retention time given by enviMass were deemed as matching the original list entries, considering possible drifts caused by wastewater matrix effects and normal variations in the LC analytical set-up, unless specified otherwise.

A ‘case’ was defined as a measurement whose chromatograms and corresponding spectra have the same  $m/z$ , retention time, and file source (essentially, a single unique measurement). As part of the prescreening, each case was subject to quality control: the MS1 and MS2 ion chromatograms were checked *automatically* by an algorithm within the workflow in a stepwise fashion as per checks and thresholds 1–5 listed in Table 1. Failure to meet any of the criteria in the checks caused the case to be rejected from further identification efforts.

Cases that passed quality control checks 1–6 were manually inspected for peak shape and width (check 7, Table 1). Only cases that passed all quality control checks

**Table 1** Quality control checks within the prescreening workflow applied to the MS1 and MS2 spectral data for each case

Quality control check	Description	Positive mode threshold	Negative mode threshold
1	Availability of MS1 precursor	Presence/absence	
2	Minimum MS1 intensity	$1 \times 10^5$	$1 \times 10^4$
3	Maximum MS1 noise level	3x (average baseline intensity)	
4	Availability of MS2 corresponding to MS1 precursor	Presence/absence	
5	MS1–MS2 alignment window	0.3 min ( <i>i.e.</i> $\pm 0.15$ min)	
6	Deduplication of cases	Highest MS1 intensity	
7	Minimum peak width and overall shape (manual QC)	0.1 min	

Thresholds apply to data measured using an Orbitrap instrument. Checks 1–5 are part of the automated prescreening workflow, while checks 6–7 were performed manually

1–7 were used as input for MetFrag identification in the next part of the workflow.

This prescreening workflow developed and used as part of this work has been embedded into the openly available R package Shinyscreen (v.0.1.1-paper, [24]).

### Part 3—identification using MetFrag

Tentative identification was performed using MetFrag (command line v.2.4.5, [51, 68]). CompTox was used as the candidate database in the form of a local.csv file [54]. R scripts, building on the code bases of ReSOLUTION (v.0.1.8, [55]) and RChemMass (v.0.1.27, [56]), were written to accomplish the following steps.

First, the neutral monoisotopic mass corresponding to the  $[M + H]^+$  or  $[M - H]^-$  adducts indicated by enviMass in positive and negative mode, respectively, was calculated. Then, candidates of matching mass with a relative deviation of 5 ppm (selected to reflect the analytical mass error, also known as “Search ppm”) were retrieved from CompTox. Subsequently, candidates were fragmented in silico using the following fragmentation settings: Absolute Fragment Peak Match Deviation 0.001 Da (“Mzabs”), Relative Fragment Peak Match Deviation 5 ppm (“Mzppm”), and Maximum Tree Depth 2. Then, candidates were ranked according to the MetFrag Score, calculated as the sum of ten weighted scoring terms summarised in Table 2 and explained in detail below. These terms are either already built-in, or can easily be configured within MetFrag since its relaunch [51]. Candidates with identical first block InChIKeys (*i.e.*, stereoisomers, with the same structural skeleton) were grouped together.

Three scoring terms within the MetFrag Score reflect the contribution of the fragmentation spectra to the proposed identification: the FragmenterScore (in silico fragments explaining measured peaks, a function of peak count and bond dissociation energy), OfflineMetFusion (spectral similarity to entries in MassBank of North America (MoNA) using a MetFusion approach [14], and OfflineIndivMoNA (maximum spectral similarity with MoNA entries having exact InChIKey match). Four scoring terms relate to the availability of the chemical’s metadata: CPDAT\_COUNT [66] (number of entries within US EPA’s Chemicals and Products database), DATA\_SOURCES [66] (number of data sources underlying CompTox, which performs similarly to the reference count), KEMIMARKET\_HAZ (v.S17.0.1.3, [13]) (scaled and normalised hazard score calculated by the Swedish Chemicals Agency), and KEMIMARKET\_EXPO (v.S17.0.1.3, [13]) (scaled and normalised exposure score calculated by the Swedish Chemicals Agency KEMI). The remaining three terms account for the candidate’s presence or absence in suspect lists, another form of metadata availability: INDACT (Industrial Activity chemicals

**Table 2** MetFrag scoring terms and weights used in tentative identification

MetFrag scoring terms	Weights
Spectral terms	
FragmenterScore	1.0
OfflineMetFusion	1.0
OfflineIndivMoNA	1.0
Total contribution to MetFrag Score:	3.0
Metadata terms	
CPDAT_COUNT	1.0
DATA_SOURCES	1.0
KEMIMARKET_EXPO	1.0
KEMIMARKET_HAZ	1.0
NORMANSUSDAT	0.5*
REACH2017	0.5*
INDACT	1.0
Total contribution to MetFrag Score:	6.0
Maximum MetFrag Score	
Total	9.0

An asterisk (\*) indicates these terms were given lower weights to avoid overweighting due to possible redundancy across the databases

known to be used near the sampling sites, supplied by the regulator), REACH2017 (v.S32.0.1.3, [2]) (chemicals registered under the European legislation framework REACH), and NORMANSUSDAT (v.S0.0.2.0, NORMAN [43] (chemicals in the merged NORMAN Suspect List Exchange). All metadata scoring terms were weighted 1 except for REACH2017 and NORMANSUSDAT, which were both weighted 0.5 due to the high redundancy across the two databases.

To calculate the maximum possible MetFrag Score, all the scoring terms except NORMANSUSDAT, REACH2017, INDACT, and OfflineIndivMoNA are first normalised to their respective largest values among the candidate set and scaled between 0–1. These normalised and scaled values are then summed together with the presence/absence scores of NORMANSUSDAT, REACH2017, and INDACT (0.5, 0.5, 1.0 if present, 0, 0, 0, if absent, respectively), and the similarity score from OfflineIndivMoNA (which is not scaled as it is already defined between 0 and 1).

Tentative identifications by MetFrag were communicated using an established system of levels [57], reiterated here with study-specific context for clarity: as MetFrag is an in silico method, it generally gives identifications of Level 3 confidence based on evidence for possible chemical structure using MS1, MS2 and experimental data/context. These identifications are tentative and require further validation before achieving higher confidence levels, as do Level 2a identifications of probable structure

(See figure on next page.)

**Fig. 2** Examples of cases which pass and fail quality control within the prescreening workflow. Quality control helped isolate measurements which were suitable for non-target identification and discarded those which are not. Panel A shows ShinyScreen's graphical user interface and an example of a case whose MS1–MS2 measurement is suitable for non-target identification—its extracted ion chromatogram shows a MS1 peak of sufficiently high intensity, a corresponding MS2 event that is temporally well-aligned, and its MS2 spectrum. The remaining panels show examples of cases that were eliminated from further identification efforts by the workflow as they were deemed unsuitable due to an excessively noisy MS1 spectrum (B; check 3 in Table 1), the absence of an MS2 event, (C; check 4) misaligned MS1 and MS2 events (D; check 5), and poor MS1 peak shape and width (E; check 7)

based on a library spectrum match, corresponding to a high MoNA individual similarity score ( $>0.9$ ) in the present work. Level 1 identifications require confirmation of the structure using a reference standard and includes target compounds.

## Results

### Prescreening and quality control

Preliminary manual inspection of the data using XCalibur Qual Browser (v.4.2.28.14, Thermo Fisher Scientific, Waltham MA, USA) indicated that not all measurements of each individual  $m/z$  were suitable for non-target identification because, *e.g.*, MS1 precursors were often at low intensity, some MS2 spectra were absent, and spikes and/or noise were observed in the MS1 extracted ion chromatogram instead of actual peaks. Therefore, the prescreening workflow consisting of 7 quality control checks (Table 1) was implemented to isolate measurements that were suitable for non-target identification. Figure 2 provides examples of measurements visualised using ShinyScreen which passed all quality control checks (Panel A) and failed either one or more checks (Panels B–E), respectively. The latter were automatically eliminated from further consideration by the workflow because they were deemed unsuitable for use in non-target identification.

For identification, a total of 185 non-target  $m/z$  from both List A and List B were prescreened in each of the 30 *mzML* files, resulting in 5,550 cases possible for identification. For List A containing 117  $m/z$  measured in positive mode, the prescreening workflow runtime was approximately 8 h on a laptop machine with 8 GB RAM and 2 physical cores over all 30 *mzML* files. Runtime was estimated based on timestamps from results file generation.

Of the 5,550 cases, 899 cases satisfied checks 1–5 listed in Table 1. Duplicate cases by  $m/z$  (*e.g.*, if it was detected at more than one site) were eliminated by prioritising those with the highest MS1 intensity (check 6), leaving 157 cases (approximately 0.03% of total cases) to be manually inspected for peak width and shape (check 7, Fig. 2e). Of these 157 cases, only 22 passed manual inspection and qualified for further identification efforts using MetFrag (listed in full in Additional file 1: Table S2). Figure 3 summarises this data reduction

outcome as a result of quality control within the prescreening workflow.

### Tentative identification using MetFrag

Tentative identifications for the 22  $m/z$  that passed quality control checks were obtained using MetFrag. Candidates for each  $m/z$  were proposed as ranked lists according to their respective MetFrag Scores comprising the ten scoring terms described in Table 2 (full MetFrag results with lists of ranked candidates available in Massive). Figure 4 shows the distribution of MetFrag Scores classified into tertiles for the top-ranked candidate for each of the 22  $m/z$ .

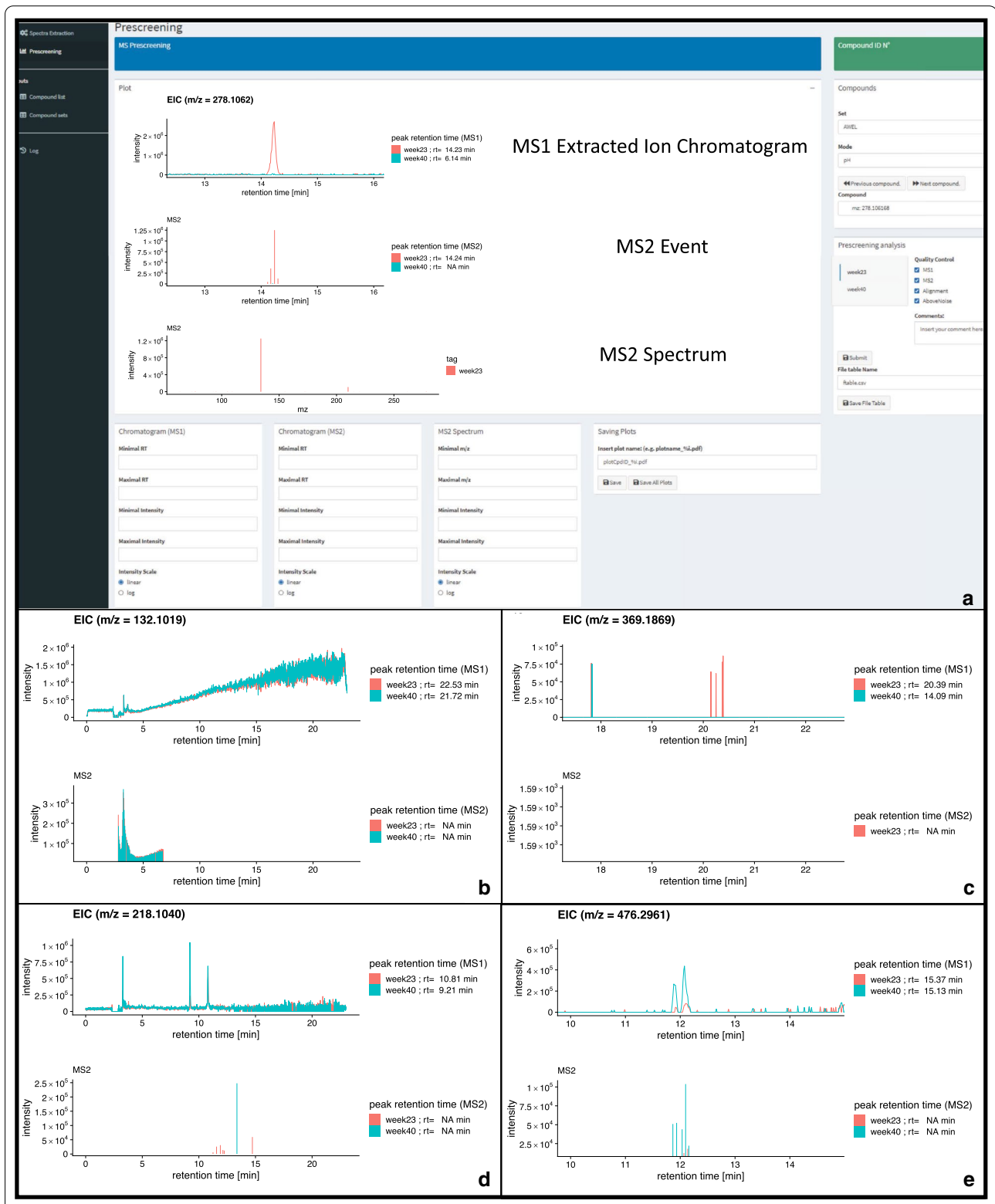
### Interpretation of MetFrag results

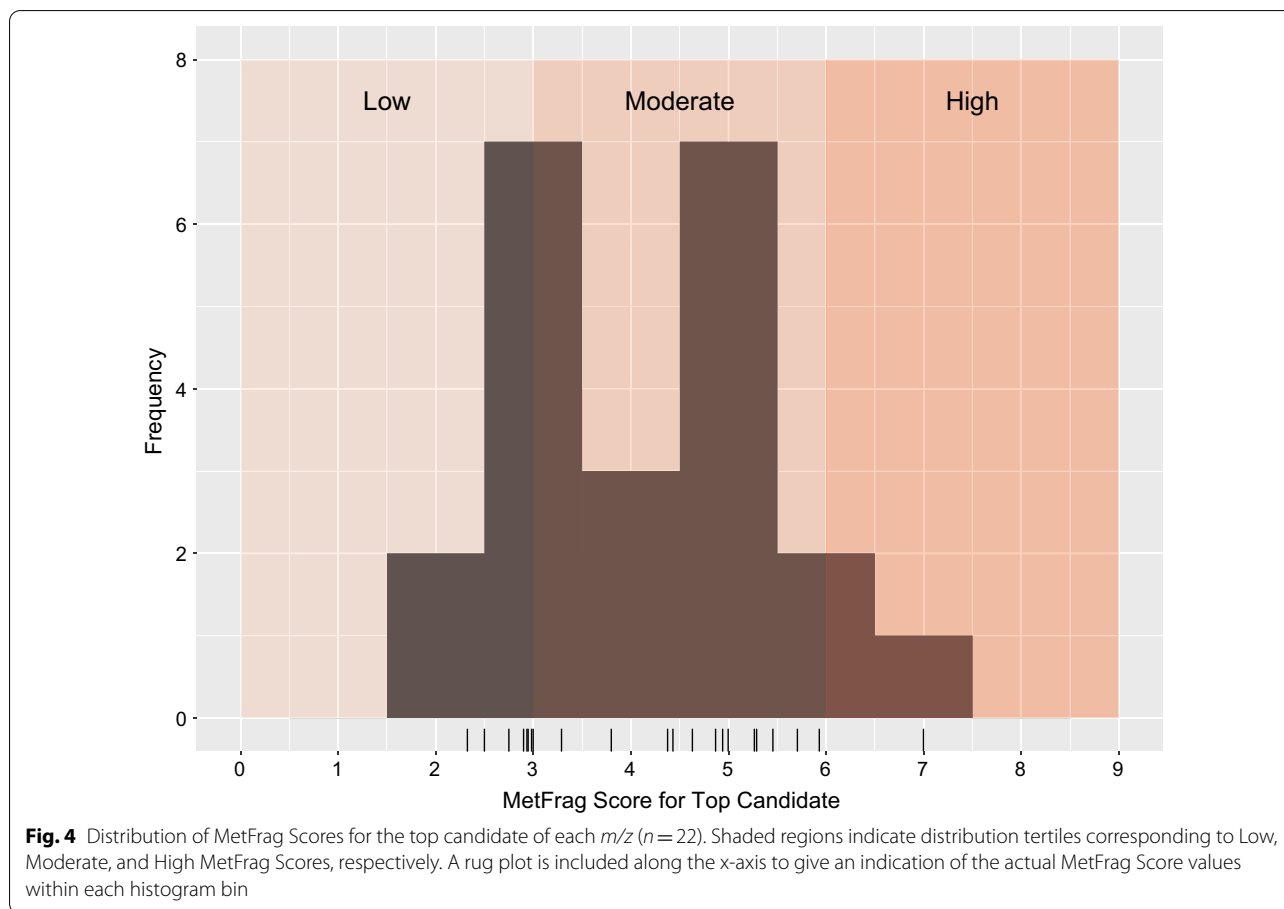
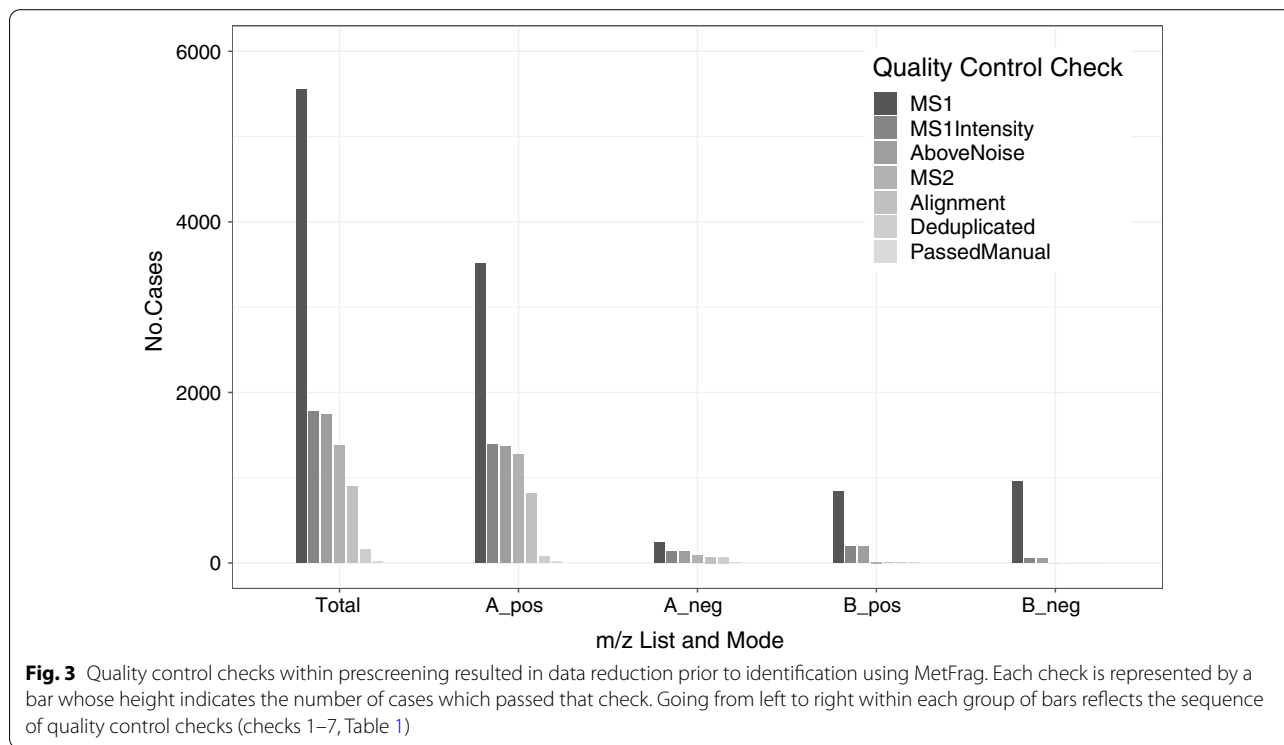
Given the background and context of this work (*i.e.* NTA in environmental monitoring to identify high-priority unknowns), the MetFrag results described above do not represent a satisfactory end-point/end-product of this study. In other words, it does not suffice to present MetFrag's outputs (lists of ranked candidates, one list per  $m/z$ ) alone, as these results alone do not provide sufficient direction for the next regulatory steps. Rather, it is crucial that these scientific outcomes are translated into transparent and actionable information for regulatory scientists to aid their future decision-making with respect to the following questions:

1. What does the distribution of MetFrag Scores mean and what are the implications?
2. How can this information guide evidence-based decision-making regarding further identification efforts? (*e.g.*, by adding candidates to suspect lists for future Suspect Screenings, purchasing reference standards for confirmation, etc.)

The following section addresses these two questions through in-depth interpretation of MetFrag's results at two levels: at a global level across all 22  $m/z$  studied, and at a candidate level per  $m/z$ , respectively. The aim of these interpretations is to deliver information based on scientific premises that is actionable from a regulatory point of view and in doing so, present 'complex' MetFrag results in an interpretable way using Scenario Analysis.







**Table 3** Four different scenarios corresponding to the four possible combinations of Spectral and Metadata scores

	High Metadata score	Low Metadata score
High Spectral score	Scenario 1: high MetFrag Score (> 6)	Scenario 3: moderate MetFrag Score (3–6)
Low Spectral score	Scenario 2: moderate MetFrag Score (3–6)	Scenario 4: low MetFrag Score (< 3)

Spectral and Metadata scores are components of the final MetFrag Score (Table 2). Scores falling into the different tertiles of the MetFrag Score distribution are classified as low, moderate, and high, respectively, as indicated in Fig. 4

**Table 4** MetFrag Score breakdown for the top candidates of four  $m/z$ 

	MetFrag Score (weighted)			
	7.00	4.63	2.95	2.50
MetFrag Score distribution classification	High (>6)	Moderate (3–6)	Moderate (~ 3–6; borderline)	Low (< 3)
Scenario	Scenario 1— <i>high</i> Spectral and Metadata scores	2— <i>low</i> Spectral and <i>High</i> Metadata scores	3— <i>high</i> Spectral and <i>Low</i> Metadata scores	4— <i>low</i> Spectral and Metadata scores
$m/z$	278.1062	187.0938	152.0198	199.1050
MetFrag Score breakdown (top candidate only)				
Spectral terms (raw scores)				
FragmenterScore	95.30	7.88	217.84	19.48
OfflineMetFusion	4.64	0.88	2.06	2.81
OfflineIndivMoNA	1.00	0	0	0
Metadata terms (raw scores)				
CPDAT_COUNT	0	0	0	0
DATA_SOURCES	47	42	1	1
KEMIMARKET_EXPO	16	11	0	0
KEMIMARKET_HAZ	9	2	0	0
NORMANSUSDAT	1	1	0	0
REACH2017	1	1	0	0
INDACT	0	0	0	0

Each MetFrag Score here represents one of the four scenarios in Table 3

Regarding the MetFrag Scores of the top candidates for each  $m/z$  (Fig. 4), this distribution arises as a result of four possible combinations of Spectral and Metadata Score components contributing toward the final MetFrag Score (Table 3). The distribution is split into tertiles based on the range of MetFrag Scores possible (0–9), and each tertile is assigned an associated scenario, as explained below.

Scenario 1 features both strong spectral and metadata evidence supporting a given candidate, resulting in a High MetFrag Score. Moderate MetFrag Scores result when one of these two scoring components, Spectral or Metadata, is low and the other is high, leading to Scenarios 2 and 3. Finally, Scenario 4 describes situations where both Spectral and Metadata scores are low, resulting in Low MetFrag Scores. Table 4 shows the breakdown of the MetFrag Score into its component Spectral and Metadata terms for four illustrative examples, one for each scenario. These representative examples were selected from

the distribution (Fig. 4) and are the respective top-ranked candidates for 4  $m/z$ .

The implications of this distribution (Fig. 4) can guide future actions depending on whether depth or breadth of the NTA study is more important. For example, if the ultimate goal is to fully identify one or two high-priority non-target unknowns to Level 1 confidence, pursuing candidates with High MetFrag Scores (3<sup>rd</sup> tertile, dark red region in Fig. 4, Scenario 1 in Table 3) is recommended. Alternatively, if gaining a wide survey of the possibly relevant but as yet unknown environmental pollutants throughout the sampling campaign is preferred (akin to a ‘first-approximation’ of the situation), then even candidates with moderate and/or low scores can also be considered further depending on the relevance of the scoring terms to the context. Additionally, further decisions on future actions can be made based on possible limitations of the study which may be known from the outset (see Discussion).

**Table 5** MetFrag Score breakdown by scoring term for the top 4 candidates for  $m/z$  278.1062 (ultimately identified as metazachlor with Level 1 confidence)

MetFrag Scoring terms	Candidate 1 DTXSID4058156	Candidate 2 DTXSID90916646	Candidate 3 DTXSID40736053	Candidate 4 DTXSID30150421
Spectral terms (raw scores)				
FragmenterScore	<b>95.30</b>	18.00	61.52	47.52
OfflineMetFusion	<b>4.64</b>	3.65	3.25	2.99
OfflineIndivMoNA	<b>1.00</b>	0	0	0
Metadata terms (raw scores)				
CPDAT_COUNT	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
DATA_SOURCES	<b>47</b>	2	1	7
KEMIMARKET_EXPO	<b>16</b>	0	0	0
KEMIMARKET_HAZ	<b>9</b>	0	0	0
NORMANSUSDAT	<b>1</b>	0	0	0
REACH2017	<b>1</b>	<b>1</b>	0	0
INDACT	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
MetFrag Score (weighted)				
Total	7.00	1.52	1.37	1.29

Raw scores are given for interpretability; the maximum raw score over all candidates (used to normalise for the ranking) is indicated in bold. The final MetFrag Score is a sum of the normalised and weighted scoring terms as described in the Methods. Here, Candidate 1 has the highest overall MetFrag Score, supported by both spectral and metadata scoring terms. Full details on the candidates are available in MassIVE

Close inspection of the MetFrag Score, namely its component spectral and metadata scoring terms, enables results interpretation on the individual candidate level for each  $m/z$ . Irrespective of whether a breadth or depth strategy is chosen, the lists of ranked candidates should always be scrutinised for plausibility because although each identification has a top candidate ranked first by MetFrag, the top candidate may not be the only candidate worth considering (if at all) given the context of the study. Below, an in-depth analysis and results interpretation of the top 4 candidates for selected  $m/z$  is presented in the following tables as examples of each of the scenarios (Table 3). Distributed Structure-Searchable Toxicity Substance Identifiers from CompTox, known as DTXSIDs are given as identifiers. The choice to use DTXSID as candidate identifiers and not their compound names is addressed in the Discussion.

### $m/z$ 278.1062

#### Scenario 1: high Spectral and Metadata scores (high MetFrag Score; > 6)

Thirty-three compounds with matching mass were retrieved from CompTox and scored by MetFrag using the ten scoring terms (Table 2). The top-ranked candidate, DTXSID4058156, has the highest total MetFrag Score out of all the candidates proposed (Table 5). In terms of spectral information, it has the highest FragmenterScore and OfflineMetFusion score of all the candidates, as well as a MoNA library match of 0.998, while all other candidates had a MoNA library match of 0.

In terms of metadata and presence in suspect lists, DTXSID4058156 has abundant metadata, is present on many suspect lists compiled by the NORMAN Network (REACH2017, SusDat and KEMIMARKET), and has 47 underlying data sources in CompTox. Based on this aforementioned evidence, this identification has confidence level 2a.

Overall, both the spectral and metadata evidence strongly support Candidate 1 over the others, as seen in the large difference between the candidates' MetFrag Scores.

**Candidate recommendation:** Candidate 1 should be strongly considered for further identification efforts.

A reference standard of DTXSID4058156 (metazachlor) provided a retention time match within 0.03 min, thereby confirming the identification of this unknown as metazachlor with Level 1 confidence.

### $m/z$ 187.0938

#### Scenario 2: low Spectral but high Metadata scores (moderate MetFrag Score; 3–6)

For  $m/z$  187.0938, identified as a  $[M+H]^+$  adduct by enviMass, the top candidate scored poorly in the Spectral terms compared to subsequent candidates. However, its strong scoring in the metadata terms ultimately drove its high MetFrag Score (Table 6).

The distribution of MetFrag Scores in Table 6 indicates that the top 3 (or even 4) candidates have relatively similar scores. Although the spectral data rather support Candidates 2 or 3 as better matching the experimental

**Table 6** MetFrag Score breakdown by scoring term for the top 4 candidates for  $m/z$  187.0938

MetFrag Scoring terms	Candidate 1 DTXSID5020526	Candidate 2 DTXSID70198185	Candidate 3 DTXSID10185791	Candidate 4 DTXSID70382365
Spectral terms (raw scores)				
FragmenterScore	7.88	<b>65.03</b>	50.21	40.46
OfflineMetFusion	0.88	<b>1.04</b>	1.01	0.86
OfflineIndivMoNA	0	0	0	0
Metadata terms (raw scores)				
CPDAT_COUNT	0	0	0	0
DATA_SOURCES	<b>42</b>	7	5	7
KEMIMARKET_EXPO	<b>11</b>	2	2	6
KEMIMARKET_HAZ	2	<b>3</b>	<b>3</b>	<b>3</b>
NORMANSUSDAT	<b>1</b>	<b>1</b>	<b>1</b>	<b>1</b>
REACH2017	<b>1</b>	<b>1</b>	<b>1</b>	0
INDACT	0	0	0	0
MetFrag Score (weighted)				
Total	4.63	4.34	4.03	3.65

Raw scores are given for interpretability; the maximum raw score over all candidates (used to normalise for the ranking) is indicated in bold. The final MetFrag Score is a sum of the normalised and weighted scoring terms as described in the Methods. Here, Candidate 1 has the highest overall MetFrag Score despite low Spectral term scores due to its high scoring Metadata. Full details on the candidates are available in MassIVE

data, the high KEMIMARKET\_EXPO score for Candidate 1 indicates that it may be of greater concern in a regulatory context due to the potentially large exposure volumes, and could be considered for further confirmation efforts to eliminate this from consideration in future campaigns.

**Candidate recommendation:** All top four candidates should be considered for further identification efforts due to high exposure and hazard scores.

### **$m/z$ 249.0728**

#### **Additional example for Scenario 2: low Spectral but high Metadata scores (moderate MetFrag Score; 3–6)**

The information provided by high Metadata scores can serve as the discriminating factor between candidates when their Spectral scores yield little/poor information which in turn gives little indication of how to rank the candidates if only spectral evidence had been considered. In this sense, Metadata scoring terms contribute an extra layer of information beyond spectral evidence towards identifying potentially relevant unknowns.

For example, the top four candidates of  $m/z$  249.0728 (Table 7) have comparably poor Spectral scores meaning there is overall little spectral evidence supporting these identifications. However, Candidate 1 distinguishes itself significantly from the other candidates because of its relatively high Metadata scores, in particular its KEMIMARKET\_EXPO, KEMIMARKET\_HAZ, and presence in REACH2017. Therefore, it has higher environmental relevance than subsequent candidates, which explains its top ranking.

**Candidate recommendation:** Candidate 1 should be considered for further identification efforts given the moderate KEMI exposure and hazard scores, indicating potential environmental relevance in Europe.

### **$m/z$ 142.0975**

#### **Additional example for Scenario 2: low Spectral but high Metadata scores (moderate MetFrag Score; 3–6)**

Similar to the previous example, candidates for  $m/z$  142.0975 have comparable performance in the Spectral scores and would be practically indistinguishable from each other if not for the large difference in their Metadata scores (Table 8). Candidate 1 differs strongly from subsequent candidates because of its relatively high KEMIMARKET\_EXPO, KEMIMARKET\_HAZ and REACH2017 scores that support its top ranking.

**Candidate Recommendation:** Candidate 1 should be considered for further identification efforts given high Europe-relevant Metadata scores.

### **$m/z$ 152.0198**

#### **Scenario 3: high Spectral scores but low Metadata scores (moderate MetFrag Score; 3–6)**

For the top candidates of  $m/z$  152.0198, practically no metadata exists except for DATA\_SOURCES—each candidate has 1, indicating that these are not particularly well-known chemicals (or, potentially newly discovered and not well documented in public databases yet). However, the FragmenterScores of the candidates differed sufficiently to discriminate between them and indicate that Candidate 1 may be the best match in this case (Table 9).

**Table 7** MetFrag Score breakdown by scoring term for the top 4 candidates for  $m/z$  249.0728

MetFrag scoring terms	Candidate 1 DTXSID50885566	Candidate 2 DTXSID60154230	Candidate 3 DTXSID70233803	Candidate 4 DTXSID80278866
Spectral terms (raw scores)				
FragmenterScore	0	0	0	0
OfflineMetFusion	0.67	0.64	0.63	0.70
OfflineIndivMoNA	0	0	0	0
Metadata terms (raw scores)				
CPDAT_COUNT	0	0	0	0
DATA_SOURCES	<b>6</b>	3	3	2
KEMIMARKET_EXPO	<b>2</b>	0	0	0
KEMIMARKET_HAZ	<b>3</b>	0	0	0
NORMANSUSDAT	0	0	0	0
REACH2017	<b>1</b>	0	0	0
INDACT	0	0	0	0
MetFrag Score (weighted)				
Total	4.43	1.39	1.38	1.30

Raw scores are given for interpretability; the maximum raw score over all candidates (used to normalise for the ranking) is indicated in bold. The final MetFrag Score is a sum of the normalised and weighted scoring terms as described in the Methods. Here, differences in candidates' Metadata scores allowed them to be differentiated from each other despite equally poor Spectral scores. Full details on the candidates are available in MassIVE

**Table 8** MetFrag Score breakdown by scoring term for the top 4 candidates for  $m/z$  142.0975

MetFrag Scoring terms	Candidate 1 DTXSID40200921	Candidate 2 DTXSID50863460	Candidate 3 DTXSID40233077	Candidate 4 DTXSID90380247
Spectral terms (raw scores)				
FragmenterScore	<b>200.29</b>	156.23	143.16	229.32
OfflineMetFusion	3.44	3.64	<b>3.96</b>	3.52
OfflineIndivMoNA	0	<b>0.01</b>	0	0
Metadata terms (raw scores)				
CPDAT_COUNT	0	0	0	0
DATA_SOURCES	6	<b>11</b>	7	2
KEMIMARKET_EXPO	<b>2</b>	0	0	0
KEMIMARKET_HAZ	<b>3</b>	0	0	0
NORMANSUSDAT	<b>1</b>	<b>1</b>	0	0
REACH2017	<b>1</b>	0	0	0
INDACT	0	0	0	0
MetFrag Score (weighted)				
Total	5.29	3.11	2.26	2.07

Raw scores are given for interpretability; the maximum raw score over all candidates (used to normalise for the ranking) is indicated in bold. The final MetFrag Score is a sum of the normalised and weighted scoring terms as described in the Methods. Here, differences in candidates' Metadata scores allowed them to be differentiated from each other despite equally good Spectral scores. Full details on the candidates are available in MassIVE

**Candidate recommendation:** Candidate 1 may be considered for further identification efforts, but candidates for other masses are more promising in the regulatory context (Table 10).

### $m/z$ 199.1050

#### Scenario 4: low Spectral scores, low Metadata scores (low MetFrag Score; < 3)

Candidates proposed for  $m/z$  199.1050 had neither particularly strong spectral nor metadata information, resulting in low overall MetFrag Scores. In this case, there is no strong evidence that any of the candidates available in CompTox are of particular interest in the context of the investigation.

**Table 9** MetFrag Score breakdown by scoring term for the top 4 candidates for  $m/z$  152.0198

MetFrag Scoring terms	Candidate 1 DTXSID30534106	Candidate 2 DTXSID30540904	Candidate 3 DTXSID90610112	Candidate 4 DTXSID40849677
Spectral terms (raw scores)				
FragmenterScore	<b>217.84</b>	158.82	144.54	142.75
OfflineMetFusion	2.06	2.08	<b>2.17</b>	2.02
OfflineIndivMoNA	0	0	0	0
Metadata terms (raw scores)				
CPDAT_COUNT	0	0	0	0
DATA_SOURCES	1	1	1	1
KEMIMARKET_EXPO	0	0	0	0
KEMIMARKET_HAZ	0	0	0	0
NORMANSUSDAT	0	0	0	0
REACH2017	0	0	0	0
INDACT	0	0	0	0
MetFrag Score (weighted)				
Total	2.95	2.69	2.66	2.60

Raw scores are given for interpretability; the maximum raw score over all candidates (used to normalise for the ranking) is indicated in bold. The final MetFrag Score is a sum of the normalised and weighted scoring terms as described in the Methods. Here, the Spectral scores provided the means for MetFrag to differentiate the candidates despite their equally poor Metadata scores. Full details on the candidates are available in MassIVE

**Table 10** MetFrag Score breakdown by scoring term for the top 4 candidates for  $m/z$  199.1050

MetFrag Scoring terms	Candidate 1 DTXSID40514171	Candidate 2 DTXSID00556299	Candidate 3 DTXSID20776997	Candidate 4 DTXSID50511555
Spectral terms (raw scores)				
FragmenterScore	<b>19.48</b>	2.43	8.12	6.00
OfflineMetFusion	2.808	2.809	2.800	<b>2.810</b>
OfflineIndivMoNA	0	0	0	0
Metadata terms (raw scores)				
CPDAT_COUNT	0	0	0	0
DATA_SOURCES	1	2	1	1
KEMIMARKET_EXPO	0	0	0	0
KEMIMARKET_HAZ	0	0	0	0
NORMANSUSDAT	0	0	0	0
REACH2017	0	0	0	0
INDACT	0	0	0	0
MetFrag Score (weighted)				
Total	2.50	2.12	1.91	1.81

Raw scores are given for interpretability; the maximum raw score over all candidates (used to normalise for the ranking) is indicated in bold. The final MetFrag Score is a sum of the normalised and weighted scoring terms as described in the Methods. Full details on the candidates are available in MassIVE

**Candidate recommendation:** Candidate 1 may be considered for further identification efforts, but candidates for other masses are more promising.

#### Information for regulatory decision-making on further identification efforts/next steps

Table 11 summarises the candidate recommendations presented above, where 7–9 candidates are

recommended for further identification efforts for the 6  $m/z$  presented here.

The top four candidates for each of the remaining 16  $m/z$  were analysed in the same way as discussed above, and candidates were evaluated based on the same criteria as described: prioritisation according to tertile, scenario, and Spectral and Metadata scores, including potential exposure and hazards (Additional file 1: Tables S3–S18). For these 16  $m/z$ , a total of 25–49 candidates (out of

**Table 11** Candidates for six  $m/z$  meriting further identification efforts based on individual evaluations

$m/z$	MetFrag results scenario	Candidates for further consideration	Justification for candidate recommendation
278.1062	Scenario 1	1	High MetFrag Score overall (high Spectral and Metadata scores); subsequent candidates very poor in comparison
187.0938	Scenario 2	4	Moderate MetFrag Score overall (low Spectral but high Metadata scores); MetFrag Scores very similar across candidates, therefore all worth consideration
249.0728	Scenario 2 (additional example)	1	Moderate MetFrag Score overall (low Spectral but high Metadata Scores); non-zero KEMIMARKET_EXPO and KEMIMARKET_HAZ, and presence in REACH2017 suspect list unlike subsequent candidates
142.0975	Scenario 2 (additional example)	1	Moderate MetFrag Score overall (low Spectral but high Metadata Scores); non-zero KEMIMARKET_EXPO and KEMIMARKET_HAZ, and presence in REACH2017 suspect list unlike subsequent candidates
152.0198	Scenario 3	0–1	Moderate MetFrag Score overall (high Spectral but low Metadata scores); borderline low MetFrag Score, only worth (weakly) considering Candidate 1
199.1050	Scenario 4	0–1	Low MetFrag Score overall (low Spectral and Metadata scores); only worth (weakly) considering Candidate 1

Candidates were evaluated on an individual level for 6  $m/z$  (selected out of 22  $m/z$  as representative examples). Full details on further candidates are available in *MassIVE*

possible 16 times  $4=64$ ) are recommended for further identification efforts (Additional file 1: Table S19). Thus, for all the 22  $m/z$  which underwent MetFrag identification in this study, an overall total of 32–58 candidates (out of possible 22 times  $4=88$ ) are recommended for further identification efforts. These candidate numbers are provided as ranges to allow for flexibility in project management and future steps, which may depend on available resources (see [Discussion](#)).

## Discussion

In this study, non-target analysis was performed retrospectively on samples from Swiss WWTP effluents that had been collected as part of an existing regulatory environmental monitoring campaign. Instead of an exploratory approach that is still common amongst NTA studies, the research questions that directed this study were derived from regulatory priorities, thereby ensuring outcomes of direct and immediate relevance for environmental monitoring and protection.

Unknowns of regulatory interest were defined as those with the *highest intensities* and *highest temporal frequency* with *point sources* across all the samples of the sampling campaign. These criteria had been predefined by the regulatory coauthors of this study, and resulted in a list of  $m/z$  of interest that were manually selected after filtering and sorting the masses using *enviMass*. In the current work, the mass spectra of the  $m/z$  of interest from the given list were subjected to pre-screening and quality control (Fig. 2) to ensure their suitability for use in non-target identification. Quality control isolated measurements worthy of further identification efforts and eliminated those of poor standard, effectively resulting in

data reduction (Fig. 3). The prescreening workflow was written in R and is now openly available within the package *ShinyScreen* [24].

Then, MetFrag [51, 68] was employed to provide tentative identifications for these unknowns, leveraging its extensive metadata capabilities “post-relaunch”, as well as several open resources/information sources, including chemical information from regulators around the world. MetFrag analysis was performed via the command line using scripts based on *ReSOLUTION* [55] and *RChem-Mass* [56].

Tentative identifications for 22  $m/z$  were obtained using MetFrag (21 at Level 3, 1 at Level 2a, whose identity was eventually confirmed to Level 1). These identifications were evaluated in terms of (i) a score distribution for the top candidates (Fig. 4) and (ii) Scenario Analysis (Table 3) according to the regulatory context and research questions underlying this work. Final candidate recommendations were given based on MetFrag Score breakdowns, thereby providing in-depth and transparent analyses of the spectral and metadata evidence for proposed candidates. For the 22  $m/z$  analysed, 32–58 candidates were recommended for further identification efforts.

Regarding the analytical method, direct injection without enrichment was used here, as non-target compounds of *high intensity* were of primary interest and enrichment was not considered necessary. Additionally, Mechelke et al. recently found that direct injection is comparatively better suited to capturing a broader range of compounds, including highly polar compounds that would otherwise experience poor recovery during enrichment [38]. The spectral data were recorded using data-dependent acquisition mode with an inclusion list in this study. While



future NTA work could explore the use of data-independent acquisition (DIA), omitting the necessity for an inclusion list, this adds other complexities, as lower intensity precursors may not yield fragments of sufficient intensity and data interpretation inevitably becomes more complicated, especially if complex matrices like wastewater with many co-eluting compounds are being studied.

Quality control was a critical element in the prescreening workflow, as preliminary manual inspection of the data using XCalibur revealed variable data quality. In fact, most data (>80% cases) were not fully suitable for the intended non-target identification. R scripts (now embedded within ShinyScreen package) were written to automate most of the quality control checks (Table 1, checks 1–5). Automated quality control allowed for quick and reproducible processing of the large quantity of data needed to answer the superlative research questions guiding this work. The variable quality of the data had several likely causes: (i) List B masses were not in the inclusion list; (ii) MS2 were not measured immediately after MS1, therefore sample degradation over long storage time between MS1 and MS2 measurements could have occurred, and (iii) possibly over-restrictive enviMass prioritisation criteria. Thus, the small number of cases (~0.03% of total) passing all quality control checks and qualifying for MetFrag identification was not unexpected.

MetFrag was configured to comprise both Spectral and Metadata scoring terms, including chemical suspect lists and scoring terms from international regulators within the latter such as KEMIMARKET\_EXPO, KEMIMARKET\_HAZ, REACH2017, NORMANSUSDAT, and CPDAT\_COUNT. Paired with CompTox as its candidate database, MetFrag was thus specifically customised to perform non-target identification of environmental unknowns in WWTP samples within a regulatory context in this work. Beyond using fragmentation information alone, using metadata to inform MetFrag's identifications proved to be especially important in certain situations, *e.g.*, when Spectral scores based on fragmentation were not informative enough to distinguish candidates from each other (Tables 7 and 8). Crucially, the information provided by metadata can serve as guidance for future regulatory actions in the context of the environmental protection aims of this study. For example, although certain candidate(s) may not be top-ranked or have strong spectral evidence (Table 6), potentially concerning hazard and exposure scores may qualify a certain candidate for serious consideration in future work in the spirit of applying the Precautionary Principle.

Regarding the components of the MetFrag Score, a total of ten scoring terms, three Spectral and seven Metadata, were used to score candidates. Compared to most

previous studies which used MetFrag as mentioned in the Introduction, this number may seem large. However, adding extra scoring terms does not appear to compromise MetFrag's identification capabilities. In fact, the additional scoring terms were beneficial because further bases for differentiating between candidates became available. In other words, using more scoring terms can provide more granularity when distinguishing candidates, which is important for candidate evaluation and recommendation. Further scoring terms based on physical–chemical properties could be integrated in the future such as correlation of the partitioning coefficient  $\log K_{ow}$  (or  $\log P$ ) with retention time as already available in MetFrag [51]. While such scoring criteria would help filter out any unrealistic candidates based on objective criteria like ionisability and polarity, insufficient information was available to perform retention time correlation via MetFrag in this study.

With respect to the individual terms, CPDAT\_COUNT, INDACT, and OfflineIndividualMoNA proved to be relatively uninformative in this particular study, evidenced by their frequent zero-value scores. As a database containing consumer chemical products ranging from those used in home maintenance (paints, sealants, lubricants, cleaners, etc.) to personal care products (hair gel, nail polish, face cream, makeup, etc.), CPDAT's limited applicability in wastewater studies such as the present one is unsurprising, and it instead may be more suitable for exposomics studies involving, *e.g.*, household dust. INDACT, the list of industrial activity chemicals known to be used in the vicinity of the WWTPs as disclosed to the regulator, had the strongest potential to improve the identification results. However, not a single candidate across all the MetFrag results was present on this suspect list, which could suggest that the chemical disclosures made by the industries were either incomplete, unsuitable for identification purposes (*e.g.*, parent compounds were disclosed but possibly only transformation products are present in the environment/are detectable, UVCBs with unspecific chemical identities, etc.), and/or inherently do not end up in wastewater if the compounds themselves are used in closed circuits, are recycled, or partition into sludge if they are very non-polar. Lastly, while mass spectral libraries are inherently incomplete [44], a low OfflineIndividualMoNA score does not necessarily indicate poor spectral library matches. Rather, low OfflineIndividualMoNA scores could also signify that the candidate is not present within MoNA to begin with, or result from noisy experimental spectra even if the match would otherwise be good. Therefore, evaluating candidates on this scoring term alone must be done with these factors in mind, and improvements to its design to avoid possible faulty interpretations could constitute future work. Other future

work on MetFrag itself could involve the addition of new Spectral scoring terms which do not require scaling via normalisation of the maximum value, as this maximum value is highly dependent on the candidate database chosen. For instance, a simple spectral similarity metric such as cosine similarity would evaluate how well the *in silico* and experimental fragmentation spectra align, independent of those of other candidates.

CompTox, the candidate database chosen here, remains one of the most environmentally-focused open databases of chemical compounds as it exclusively contains chemicals of environmental and toxicological relevance. Compared to other open databases like PubChem (111 million chemical structures, August 2020), CompTox is also smaller in size (883,000 chemicals, February 2021). Therefore, MetFrag paired with CompTox is likely to suggest smaller lists of candidates which are *de facto* environmentally-meaningful, making workflow runtimes shorter and candidate evaluation relatively easier. However, using CompTox has drawbacks, principally stemming from its lack of comprehensiveness when compared to PubChem. In some cases, there may be a lack of candidates matching the identification criteria when using CompTox with MetFrag simply because they may not exist within CompTox itself to begin with due to its limited size and scope. PubChemLite [55, 56, 58] represents one complementary alternative to these issues, as it is by design essentially a subset of environmentally relevant compounds based on compound classifications. Overall, the ability to subset databases based on usage and classification information of chemicals can be beneficial, as different regulatory bodies may have different mandates, and studies can be designed to align with those mandates accordingly, *e.g.*, focus only on chemicals with (i) known usage in industrial manufacturing, or (ii) agricultural chemicals, or (iii) pharmaceuticals, etc.

Using scenarios as a framework to interpret MetFrag's results was critical considering the specific regulatory aims of this work: tentatively identify pollutants of high priority (with minimum Level 3 confidence) to guide further monitoring and identification efforts.

Scenario Analysis revealed in detail whether Spectral, Metadata, or both contributed to a given MetFrag Score and in turn provided the rationale behind proposed candidates. As our evaluation has shown, multiple candidates are worth considering especially if they have very similar scores (*e.g.*, Table 6), or have more compelling evidence represented by individual scoring terms as described above. In this way, Scenario Analysis as used here is highly suitable for transparently communicating scientific results in a regulatory context. On a larger scale, such analyses address a key weakness common to NTA studies: the current lack of ability to perform detailed

data interpretation – especially in a high-throughput, automatable and reproducible manner.

Furthermore, Scenario Analysis as used here can inform decision-making regarding the next steps. Besides addressing study priorities based on “depth vs. breadth” as discussed in the Results, the scenarios can be used to devise a prioritisation scheme for future work. For example, if authentic standards can only be purchased/analysed for 10 compounds due to resource limitations, those compounds should be the recommended candidates with MetFrag Scores from Scenario 1 > Scenarios 2/3 > > Scenario 4. Alternatively, if it is known from the outset that spectral data may be poor quality, Scenario 2 candidates may take precedence over Scenario 3 candidates, as the former rely on high Metadata scores and not high Spectral scores for their high MetFrag Scores. Additionally, applying the precautionary principle may motivate prioritising identity confirmations of candidates with concerning metadata like high toxicity and/or exposure (corresponding to KEMIMARKET\_HAZ and KEMIMARKET\_EXPO scores), even if those candidates are not necessarily ranked highly by MetFrag.

Practically speaking, next steps in environmental monitoring based on the results here (besides identity confirmation using authentic standards) could include expanding suspect lists using the recommended candidates to improve future suspect screening activities. These new suspects could in turn be added to the inclusion lists of future measurements, thereby already gaining an analytical ‘upper-hand’ for future NTA studies. Expanding suspect and inclusion lists in this way, possibly in combination with using a rarity score [26] that prioritises high intensity, infrequently occurring peaks, represents an evidence-based approach towards more meaningful environmental monitoring in the long-run, as these candidate compounds were tentatively ‘observed’ and are therefore *site-specific*. Otherwise, suspect lists are typically expanded based on information from national or international chemical registration lists, whose applicability may be limited depending on the actual usage/exposure in the region of concern. Therefore, an additional outcome of this study is a means to bridge target and non-target analysis by supplying meaningful candidates for suspect screening.

This work is one contribution to a much larger discussion surrounding (i) how NTA can support regulatory environmental monitoring, and (ii) the practical feasibility of applying NTA in routine environmental monitoring. (For an example of current discourse, see Germany's guidelines for non-target screening in water analysis [52].) Regarding the former, this work demonstrates that NTA can be used to address the concerns of regulators by translating research questions arising from regulatory

priorities into peak-picking/mass prioritisation criteria: in this case, high concentration unknown pollutants with point sources that occurred persistently were taken to be high-intensity precursors found at one or few sampling sites at both sampling time points. Without the ability to perform quantification, the assumption that high ion intensity represents high concentration could be validated by using different chromatographic solvent systems as a test of ionisation efficiency in future work, or implementing ionisation efficiency models [32, 46].

On the feasibility of performing NTA as part of routine regulatory environmental monitoring, the overall method described here offers a highly *automated* approach via (i) feature prioritisation via *enviMass*, (ii) prescreening and quality control (plus a manual step), and (iii) *in silico* identification, of which (ii) and (iii) were developed in this work. The results interpretation and candidate recommendation processes performed manually in this work form the basis of future efforts towards automated reporting based on Scenario Analysis, *MetFrag* Score distributions, and evaluation of critical parameters like thresholds for potential toxicities and exposure levels. Such automated reporting would not only allow scalability of future regulatory NTA studies, but could also eliminate potential biases in unknown identification—analysts would not be able to ‘cherry-pick’ candidates based on their familiarity with certain compounds because uninformative identifiers, *e.g.*, DTX-SIDs would be used up until the final results are delivered at the end of the entire method. Furthermore, while the prescreening, quality control, and identification workflow was applied retrospectively, the improvements to workflow automation detailed here could allow for quicker data analysis turnaround in the future, which would help guide future sampling and measurements planned in the short–medium term and prevent the long delays between remeasurements still commonly observed in NTA investigations—effectively, moving towards ‘real-time’ instead of retrospective NTA approaches. Two concrete follow-up initiatives are foreseen: (i) build an interface connecting *ShinyScreen* and *MetFrag*, including automated reporting features as previously described, and (ii) develop a set of ‘default’ scoring terms and settings tailored for NTA of wastewater samples. Further collaborations involving non-target wastewater studies and database hosts will help augment expert knowledge on more use cases, which would be leveraged to develop this approach further.

On a community level, standardisation would play a role in increasing the feasibility of NTA as part of routine regulatory environmental monitoring. As previously mentioned, there exist considerable, albeit nascent, efforts towards standardising analytical protocols for

non-target screening on a national level in, *e.g.*, Germany in the form of guidelines [52]. Such activities suggest that standardisation is certainly of priority to the community and may be achievable over time. However, NTA may not be widely adopted by regulators in the short- to medium-term until analytical protocols are successfully standardised. In turn, it continues to be challenging from a data analysis perspective to implement standardised workflows if the analytical parameters used for measuring data are not themselves standardised. Thus, the status quo demands that current data processing methods remain flexible to accommodate the variety of analytical parameters used, as is the case with the method presented here.

## Conclusions

A prescreening and identification workflow for analysing non-target compounds was developed in this study to retrospectively identify unknowns detected in WWTP sites in the context of directly supporting regulatory decision-making for environmental monitoring. Using Open data and Open tools including the US EPA *CompTox* Chemicals Dashboard, NORMAN Network resources such as *SusDat* and the Suspect List Exchange, and *MetFrag*, tentative identifications for 21 unknown compounds were provided at Level 3 confidence, and 1 compound's identity was confirmed using a reference standard giving a Level 1 identification. These results were achieved despite limited data quality.

This study heavily emphasised results interpretation on two levels: on a global level across the chemical unknowns investigated, and on an individual candidate level. Through these analyses, specific candidates were recommended for further identification efforts, and transparent justifications were provided based on the *MetFrag* score breakdown (*i.e.*, spectral vs. metadata evidence). These recommendations, and not just *MetFrag*'s outputs, represent the final results in the regulatory and environmental monitoring context of this study, and may serve as a template to drive future developments in NTA.

The prescreening and quality control workflow developed here is embedded in the open R package *ShinyScreen* [24], which is freely available online, as is code from *ReSOLUTION* [55] and *RChemMass* [56] used for performing command-line *MetFrag* identification. The *CompTox* database version with the metadata terms used here is likewise also publicly available [54].

## Abbreviations

NTA: Non-target analysis; WWTP: Wastewater treatment plant; US EPA: United States Environmental Protection Agency; *CompTox*: US EPA *CompTox* Chemicals Dashboard; DTXSID: DSSTox Substance Identifier (from *CompTox*); CPDat: Chemicals and Products Database; REACH: Registration, Evaluation, Authorisation and Restriction of Chemicals; MoNA: MassBank of North America; UVCB:

Chemical substances of Unknown or Variable composition, Complex reaction products, and Biological materials.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12302-021-00475-1>.

**Additional file 1: Table S1.** enviMass Parameters used for Orbitrap measurements in this study. **Figure S1.** Screenshot of the MSConvert (v.3.0.19182-51f676fbc) Graphical User Interface showing settings used to convert the .RAW mass spectrometry data to .mzML format. **Table S2.** List of 22 *m/z* which had been prioritised by enviMass and passed Quality Control to qualify for MetFrag identification. **Table S3.** *m/z* 216.0930. **Table S4.** *m/z* 177.1126. **Table S5.** *m/z* 212.0889. **Table S6.** *m/z* 173.1649. **Table S7.** *m/z* 301.1396. **Table S8.** *m/z* 218.1040. **Table S9.** *m/z* 176.0707. **Table S10.** *m/z* 193.0721. **Table S10:** *m/z* 193.0721. **Table S11.** *m/z* 249.1848. **Table S12.** *m/z* 184.0427. **Table S13.** *m/z* 171.1492. **Table S14.** *m/z* 199.1190. **Table S15.** *m/z* 185.1033. **Table S16.** *m/z* 251.1491. **Table S17.** *m/z* 211.0285. **Table S18.** *m/z* 546.2622. **Table S19:** Candidate Recommendations for all 22 *m/z*.

## Acknowledgements

The authors acknowledge Dr. Martin Loos (enviBee GmbH) for his technical support with enviMass analyses. Contributors to CompTox, MetFrag, the suspect lists on the NORMAN Suspect List Exchange, the Open software used here, and Open Science in general are gratefully appreciated.

## Authors' contributions

LK conceived the study and set up the sampling campaigns; OJ measured the data; AL, ELS, RRS designed the workflow presented; AL, ELS, TK wrote the software; AL interpreted the data, AL drafted the manuscript with inputs from ELS, RRS, LK, and OJ; AL, LK, OJ, RRS, TK, and ELS revised the submitted version. All authors read and approved the final manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL. ELS, AL, and TK are supported by the Luxembourg National Research Fund (FNR) for project A18/BM/12341006.

## Availability of data and materials

The mass spectrometry dataset generated and analysed during the current study, including the complete MetFrag results for the 22 *m/z* that were tentatively identified, are available as an open MassIVE dataset (MSV000086631) via <https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=14f51e6ec99a42329e7a0eeead0e5824>. **Software** Project name: ShinyScreen. Project home page: <https://git-r3lab.uni.lu/eci/shinyScreen>. Archived version used in this study: ShinyScreen v.0.1.1-paper (<https://git-r3lab.uni.lu/eci/shinyScreen/-/tree/v.0.1.1-paper>). Operating system(s): Windows, Mac OSX, Linux. Programming language: R Other requirements: OpenJDK and other R package dependencies listed in ShinyScreen's README. License: Apache Version 2.0 (<https://www.apache.org/licenses/LICENSE-2.0>).

## Code availability

All codes used to run the prescreening and quality control workflow and MetFrag command-line analysis is open/publicly available via <https://github.com/schymane/ReSOLUTION>, <https://github.com/schymane/RChemMass>, and ShinyScreen (see below). All other datasets and databases used as part of MetFrag identification are open/publicly available (links available in References).

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

## Competing interests

The authors declare no competing interests.

## Author details

<sup>1</sup> Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 Avenue du Swing, 4367 Belvaux, Luxembourg. <sup>2</sup> Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller University, Lessing Strasse 8, 07743 Jena, Germany. <sup>3</sup> Present Address: IFREMER (Institut Français de Recherche pour l'Exploitation de la Mer), Laboratoire Biogéochimie des Contaminants Organiques, Rue de l'Île d'Yeu, BP 21105, 44311 Nantes Cedex 3, France. <sup>4</sup> Amt Für Abfall, Wasser, Energie Und Luft (AWEL), Walcheplatz 2, 8090 Zurich, Switzerland.

Received: 23 December 2020 Accepted: 9 March 2021

Published online: 04 April 2021

## References

- Albergamo V, Schollée JE, Schymanski EL et al (2019) Nontarget screening reveals time trends of polar micropollutants in a riverbank filtration system. *Environ Sci Technol*. <https://doi.org/10.1021/acs.est.9b01750>
- Alygizakis N, Slobodnik J (2018) S32 | REACH2017 | >68,600 REACH Chemicals (Version NORMAN-SLE-S32013). Zenodo. <https://doi.org/10.5281/zenodo.3653160>. Accessed 16 Aug 2020
- Anliker S, Loos M, Comte R et al (2020) Assessing emissions from pharmaceutical manufacturing based on temporal high-resolution mass spectrometry data. *Environ Sci Technol* 54:4110–4120. <https://doi.org/10.1021/acs.est.9b07085>
- Beckers L-M, Brack W, Dann JP et al (2020) Unraveling longitudinal pollution patterns of organic micropollutants in a river by non-target screening and cluster analysis. *Sci Total Environ* 727:138388. <https://doi.org/10.1016/j.scitotenv.2020.138388>
- Carpenter CMG, Wong LYJ, Johnson CA, Helbling DE (2019) Fall creek monitoring station: highly resolved temporal sampling to prioritize the identification of nontarget micropollutants in a small stream. *Environ Sci Technol* 53:77–87. <https://doi.org/10.1021/acs.est.8b05320>
- Chambers MC, Maclean B, Burke R et al (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* 30:918–920. <https://doi.org/10.1038/nbt.2377>
- ChemSpider | Search and share chemistry (2020). <http://www.chemspider.com/>. Accessed 13 Aug 2020
- Chiaia-Hernández AC, Günthardt BF, Frey MP, Hollender J (2017) Unraveling contaminants in the Anthropocene using statistical analysis of liquid chromatography–high-resolution mass spectrometry nontarget screening data recorded in lake sediments. *Environ Sci Technol* 51:12547–12556. <https://doi.org/10.1021/acs.est.7b03357>
- Choi Y, Kim K, Kim D et al (2020) Ny-Ålesund-oriented organic pollutants in sewage effluent and receiving seawater in the Arctic region of Kongsfjorden. *Environ Pollut* 258:113792. <https://doi.org/10.1016/j.envpol.2019.113792>
- Dionisio KL, Phillips K, Price PS et al (2018) The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products. *Sci Data* 5:180125. <https://doi.org/10.1038/sdata.2018.125>
- Faber A-H, Annevelink MPJA, Schot PP et al (2019) Chemical and bioassay assessment of waters related to hydraulic fracturing at a tight gas production site. *Sci Total Environ* 690:636–646. <https://doi.org/10.1016/j.scitotenv.2019.06.354>
- Fiehn Lab (2020) MassBank of North America. <https://mona.fiehnlab.ucdavis.edu/>. Accessed 3 Jun 2020
- Fischer S (2017) S17 | KEMIMARKET | KEMI Market List (Version NORMAN-SLE-S17013). Zenodo. <https://doi.org/10.5281/zenodo.3653175>. Accessed 8 May 2020
- Gerlich M, Neumann S (2013) MetFusion: integration of compound identification strategies. *J Mass Spectrom* 48:291–298. <https://doi.org/10.1002/jms.3123>
- Helmus R, ter Laak TL, van Wezel AP et al (2021) patRoom: open source software platform for environmental mass spectrometry based non-target screening. *J Cheminf* 13:1. <https://doi.org/10.1186/s13321-020-00477-w>



16. Hites RA, Jobst KJ (2018) Is nontargeted screening reproducible? *Environ Sci Technol* 52:11975–11976. <https://doi.org/10.1021/acs.est.8b05671>
17. Hollender J, Schymanski EL, Singer HP, Ferguson PL (2017) Nontarget screening with high resolution mass spectrometry in the environment: ready to go? *Environ Sci Technol* 51:11505–11512. <https://doi.org/10.1021/acs.est.7b02184>
18. Hollender J, van Bavel B, Dulio V et al (2019) High resolution mass spectrometry-based non-target screening can support regulatory environmental monitoring and chemicals management. *Environ Sci Eur* 31:42. <https://doi.org/10.1186/s12302-019-0225-x>
19. Hug C, Ulrich N, Schulze T et al (2014) Identification of novel micropollutants in wastewater by a combination of suspect and nontarget screening. *Environ Pollut* 184:25–32. <https://doi.org/10.1016/j.envpol.2013.07.048>
20. Human Metabolome Database (2020). <https://hmdb.ca/>. Accessed 13 Aug 2020
21. Kandie FJ, Krauss M, Beckers L-M et al (2020) Occurrence and risk assessment of organic micropollutants in freshwater systems within the Lake Victoria South Basin, Kenya. *Sci Total Environ* 714:136748. <https://doi.org/10.1016/j.scitotenv.2020.136748>
22. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27–30. <https://doi.org/10.1093/nar/28.1.27>
23. Kim S, Chen J, Cheng T et al (2021) PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res* 49:D1388–D1395. <https://doi.org/10.1093/nar/gkaa971>
24. Kondić T, Lai A, Schymanski E, et al (2020) Environmental cheminformatics/shiny screen. <https://git-r3lab.uni.lu/eci/shiny screen>. Accessed 16 Aug 2020
25. Köppe T, Jewell KS, Dietrich C et al (2020) Application of a non-target workflow for the identification of specific contaminants using the example of the Nidda river basin. *Water Res* 178:115703. <https://doi.org/10.1016/j.watres.2020.115703>
26. Krauss M, Hug C, Bloch R et al (2019) Prioritising site-specific micropollutants in surface water from LC–HRMS non-target screening data using a rarity score. *Environ Sci Eur* 31:45. <https://doi.org/10.1186/s12302-019-0231-z>
27. Krauss M, Singer H, Hollender J (2010) LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns. *Anal Bioanal Chem* 397:943–951. <https://doi.org/10.1007/s00216-010-3608-9>
28. Lara-Martín PA, Chiaia-Hernández AC, Biel-Maeso M et al (2020) Tracing urban wastewater contaminants into the Atlantic ocean by nontarget screening. *Environ Sci Technol* 54:3996–4005. <https://doi.org/10.1021/acs.est.9b06114>
29. Lege S, Eisenhofer A, Heras JEY, Zwiener C (2019) Identification of transformation products of denatonium—occurrence in wastewater treatment plants and surface waters. *Sci Total Environ* 686:140–150. <https://doi.org/10.1016/j.scitotenv.2019.05.423>
30. Letzel T (2021) FOR-IDENT—Fortschritte in der Identifizierung organischer Spurenstoffe: Zusammenführen der Hilfsmittel und Standardisierung der Suspected- und Non-Target Analytik. (Advances in the Identification of Organic Trace Pollutants: Merging Tools and Standardisation of Suspect and Non-target Analytics.) <https://www.for-ident.org/>. Accessed 28 Feb 2021
31. Li Z, Kaserzon SL, Plassmann MM et al (2017) A strategic screening approach to identify transformation products of organic micropollutants formed in natural waters. *Environ Sci Processes Impacts* 19:488–498. <https://doi.org/10.1039/C6EM00635C>
32. Liigand J, Wang T, Kellogg J et al (2020) Quantification for non-targeted LC/MS screening without standard substances. *Sci Rep* 10:5808. <https://doi.org/10.1038/s41598-020-62573-z>
33. Ljoncheva M, Stepišnik T, Džeroski S, Kosjek T (2020) Cheminformatics in MS-based environmental exposomics: current achievements and future directions. *Trends Environ Anal Chem* 28:e00099. <https://doi.org/10.1016/j.teac.2020.e00099>
34. Loos M, Schmitt U, Schollée JE (2018) biosloos/enviMass: enviMass version 3.5. <https://doi.org/10.5281/zenodo.1213098>. Accessed 13 Oct 2020
35. Luft A, Bröder K, Kunkel U et al (2017) Nontarget analysis via LC–QTOF-MS to assess the release of organic substances from polyurethane coating. *Environ Sci Technol* 51:9979–9988. <https://doi.org/10.1021/acs.est.7b01573>
36. MassBank Consortium, NORMAN Association (2021) MassBank | MassBank Europe Mass Spectral DataBase. <https://massbank.eu/MassBank/>. Accessed 28 Feb 2021
37. McEachran AD, Mansouri K, Grulke C et al (2018) “MS-Ready” structures for non-targeted high-resolution mass spectrometry screening studies. *J Cheminf*. <https://doi.org/10.1186/s13321-018-0299-2>
38. Mechelke J, Longrée P, Singer H, Hollender J (2019) Vacuum-assisted evaporative concentration combined with LC–HRMS/MS for ultra-trace-level screening of organic micropollutants in environmental water samples. *Anal Bioanal Chem* 411:2555–2567. <https://doi.org/10.1007/s00216-019-01696-3>
39. Menger F, Ahrens L, Wiberg K, Gago-Ferrero P (2021) Suspect screening based on market data of polar halogenated micropollutants in river water affected by wastewater. *J Hazard Mater* 401:123377. <https://doi.org/10.1016/j.jhazmat.2020.123377>
40. Miaz LT, Plassmann MM, Gyllenhammar I et al (2020) Temporal trends of suspect- and target-per/polyfluoroalkyl substances (PFAS), extractable organic fluorine (EOF) and total fluorine (TF) in pooled serum from first-time mothers in Uppsala, Sweden, 1996–2017. *Environ Sci Processes Impacts* 22:1071–1083. <https://doi.org/10.1039/C9EM00502A>
41. Moschet C, Anumol T, Lew BM et al (2018) Household dust as a repository of chemical accumulation: new insights from a comprehensive high-resolution mass spectrometric study. *Environ Sci Technol* 52:2878–2887. <https://doi.org/10.1021/acs.est.7b05767>
42. Muz M, Dann JP, Jäger F et al (2017) Identification of mutagenic aromatic amines in river samples with industrial wastewater impact. *Environ Sci Technol* 51:4681–4688. <https://doi.org/10.1021/acs.est.7b00426>
43. NORMAN Network, Aalizadeh R, Alygizakis N, et al (2019) SO | SUSDAT | Merged NORMAN Suspect List: SusDat (Version NORMAN-SLE-S0.0.2.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.3520132>. Accessed 8 May 2020
44. Oberacher H, Sasse M, Antignac J-P et al (2020) A European proposal for quality control and quality assurance of tandem mass spectral libraries. *Environ Sci Eur* 32:43. <https://doi.org/10.1186/s12302-020-00314-9>
45. Oetjen K, Blotevogel J, Borch T et al (2018) Simulation of a hydraulic fracturing wastewater surface spill on agricultural soil. *Sci Total Environ* 645:229–234. <https://doi.org/10.1016/j.scitotenv.2018.07.043>
46. Panagopoulos Abrahamsson D, Park J-S, Singh RR et al (2020) Applications of machine learning to in silico quantification of chemicals without analytical standards. *J Chem Inf Model*. <https://doi.org/10.1021/acs.jcim.9b01096>
47. Park N, Choi Y, Kim D et al (2018) Prioritization of highly exposable pharmaceuticals via a suspect/non-target screening approach: a case study for Yeongsan River, Korea. *Sci Total Environ* 639:570–579. <https://doi.org/10.1016/j.scitotenv.2018.05.081>
48. Pence HE, Williams A (2010) ChemSpider: an online chemical information resource. *J Chem Educ* 87:1123–1124. <https://doi.org/10.1021/ed100697w>
49. Purschke K, Zoell C, Leonhardt J et al (2020) Identification of unknowns in industrial wastewater using offline 2D chromatography and non-target screening. *Sci Total Environ* 706:135835. <https://doi.org/10.1016/j.scitotenv.2019.135835>
50. Ruff M, Mueller MS, Loos M, Singer HP (2015) Quantitative target and systematic non-target analysis of polar organic micro-pollutants along the river Rhine using high-resolution mass-spectrometry—Identification of unknown sources and compounds. *Water Res* 87:145–154. <https://doi.org/10.1016/j.watres.2015.09.017>
51. Ruttkies C, Schymanski EL, Wolf S et al (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform* 8:3. <https://doi.org/10.1186/s13321-016-0115-9>
52. Schulz W, Lucke T, et al. (2019) Non-target screening in water analysis—Guideline for the application of LC-ESI-HRMS for screening. [https://www.wasserchemische-gesellschaft.de/images/HAll/NTS-Guideline\\_EN\\_s.pdf](https://www.wasserchemische-gesellschaft.de/images/HAll/NTS-Guideline_EN_s.pdf). Accessed 27 Feb 2021
53. Schwarzbauer J, Ricking M (2010) Non-target screening analysis of river water as compound-related base for monitoring measures. *Environ Sci Pollut Res* 17:934–947. <https://doi.org/10.1007/s11356-009-0269-3>

54. Schymanski E (2019) MetFrag Local CSV: CompTox (7 March 2019 release) Wastewater MetaData File (Version WWMetaData\_4Oct2019). Zenodo. <https://doi.org/10.5281/zenodo.3472781>. Accessed 8 May 2020
55. Schymanski E (2020a) schymane/ReSOLUTION. Version 0.1.8 <https://github.com/schymane/ReSOLUTION>. Accessed 16 Aug 2020
56. Schymanski E (2020b) schymane/RChemMass. Version 0.1.27 <https://github.com/schymane/RChemMass>. Accessed 16 Aug 2020
57. Schymanski EL, Jeon J, Gulde R et al (2014) Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* 48:2097–2098. <https://doi.org/10.1021/es5002105>
58. Schymanski EL, Kondic T, Neumann S et al (2021) Empowering large chemical knowledge bases for exposomics: PubChemLite Meets MetFrag. *J Cheminform* 13:19. <https://doi.org/10.1186/s13321-021-00489-0>
59. Sousa JCG, Ribeiro AR, Barbosa MO et al (2018) A review on environmental monitoring of water organic pollutants identified by EU guidelines. *J Hazard Mater* 344:146–162. <https://doi.org/10.1016/j.jhazmat.2017.09.058>
60. Sun C, Zhang Y, Alessi DS, Martin JW (2019) Nontarget profiling of organic compounds in a temporal series of hydraulic fracturing flowback and produced waters. *Environ Int* 131:104944. <https://doi.org/10.1016/j.envint.2019.104944>
61. Tian Z, Peter KT, Gipe AD et al (2020) Suspect and nontarget screening for contaminants of emerging concern in an urban estuary. *Environ Sci Technol* 54:889–901. <https://doi.org/10.1021/acs.est.9b06126>
62. US EPA (2016) Chemical and Products Database (CPDat). US EPA. <https://www.epa.gov/chemical-research/chemical-and-products-database-cpdat>. Accessed 8 May 2020
63. Veenaas C, Bignert A, Liljelind P, Haglund P (2018) Nontarget Screening and time-trend analysis of sewage sludge contaminants via two-dimensional gas chromatography-high resolution mass spectrometry. *Environ Sci Technol* 52:7813–7822. <https://doi.org/10.1021/acs.est.8b01126>
64. Wagner TV, Helmus R, Quito Tapia S et al (2020) Non-target screening reveals the mechanisms responsible for the antagonistic inhibiting effect of the biocides DBNPA and glutaraldehyde on benzoic acid biodegradation. *J Hazard Mater* 386:121661. <https://doi.org/10.1016/j.jhazmat.2019.121661>
65. Wang Z, Walker GW, Muir DCG, Nagatani-Yoshida K (2020) Toward a Global understanding of chemical pollution: a first comprehensive analysis of national and regional chemical inventories. *Environ Sci Technol*. <https://doi.org/10.1021/acs.est.9b06379>
66. Williams AJ, Grulke CM, Edwards J et al (2017) The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminf* 9:61. <https://doi.org/10.1186/s13321-017-0247-6>
67. Wishart DS, Feunang YD, Marcu A et al (2018) HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* 46:D608–D617. <https://doi.org/10.1093/nar/gkx1089>
68. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinf* 11:148. <https://doi.org/10.1186/1471-2105-11-148>

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Overall, 21 compounds were tentatively identified with Level 3 confidence, and are suspected to be adhesives, pesticides, manufacturing reagents, and pharmaceuticals. One pesticide compound was identified with Level 1 confidence, *i.e.*, confirmed identification that was validated using a reference standard. The pre-screening step introduced in this work as a Quality Control algorithm effectively prioritised cases (sets of MS1 and corresponding MS2 signals) for non-target identification - out of 5,550 under consideration, only 22 were pursued for non-target identification that was performed using the *in silico* fragmenter MetFrag coupled to various regulatory chemical lists. These lists represent so-called 'environmental metadata' and their emerging status at the time was capitalised upon to aid non-target identification here.

Copious analysis of MetFrag's identification results were presented to help guide further non-target identification efforts such as obtaining reference standards for validation, as well as future sampling campaigns. The emphasis on results interpretation here was in effort to maintain transparency and curb the 'black box' phenomenon that often characterises non-target analysis identification workflows, particularly if the findings are intended to inform decisions in a regulatory context. All mass spectral data, R code, and supporting analysis and results are openly available online as an open MassIVE dataset (MSV000086631) via <https://doi.org/10.25345/C5CZ0K> and in a GitLab repository (<https://git-r3lab.uni.lu/eci/shinyscreen/-/tree/v.0.1.1-paper>).

Notably, one limitation of the work is its limited scope regarding compounds that occur in the environment as a result of transformations of their parent compounds. Myriad transformation products (TPs) are known to exist at detectable levels in the environment, and are in some cases potentially more bioactive and toxic than their parents. The systematic, large-scale identification of TPs remains elusive, partly because of the lack of reference standards available for purchase, and also the inaccessibility of TP information that can be used for screening environmental samples in *e.g.*, a consolidated open database. The work in the next chapter addresses these gaps by exploiting an up-and-coming database resource developed in-house for screening for the potential presence of TPs in environmental samples.

## Chapter 3

# Data Mining Transformation Product Information for Enhanced Suspect Screening

Pharmaceutical compounds are frequently detected in the environment as a result of human consumption and emission via e.g., wastewater treatment effluent. In Luxembourg as of 2019, 816 unique pharmaceutical compounds were approved for the domestic market and thus are potentially consumed by the local population. However, only five of the 92 chemicals regularly monitored by the Luxembourg Water Management Agency (*Administration de la Gestion de l'Eau*) as part of Target screening under the European Union Water Framework Directive (WFD) are pharmaceuticals, the rest being pesticides and related compounds. Because pharmaceuticals and their transformation products likely remain bioactive upon emission despite wastewater treatment, their presence in the environment may pose a threat to organisms living in Luxembourgish waters and potentially also human health. Therefore, this study focused on the identification and, where possible, quantification of pharmaceuticals and their transformation products in Luxembourgish surface water collected across various sites in the country over 2 years as part of the national monitoring campaign.

To identify and quantify pharmaceuticals in Luxembourgish waters, an augmented suspect screening of parent pharmaceutical compounds *and* their TPs was performed using two chemical lists: 1) the list of approved pharmaceuticals described above,<sup>92</sup> and 2) a newly-generated list of 82 pharmaceutical TPs resulting from data-mining two sources, PubChem and the scientific literature. PubChem is the largest open chemical database that not only contains basic information on chemicals, but also features cross-linked information such as provenance, patents, production, usage, links to disease *etc.* At the time of the study, integrating TP information into PubChem's backend was still in the early stages, so methods for mining PubChem's TP information were still in development, with new TP information



constantly being curated and added at the same time.<sup>93,94</sup> Besides mining PubChem, TP information was also obtained from Anliker et al., who published a list of pharmaceuticals and their TPs that were studied in Switzerland.<sup>95</sup> The TPs from these two sources were combined and curated to achieve a list of 82 unique TPs that was used in suspect screening.

A total of 94 pharmaceutical parent compounds, 86 of which were quantified, plus 16 transformation products were identified in this study. Considering the national scale of the monitoring campaign, the breadth of the pharmaceutical screening, and the copious pharmaceutical concentration data collected, data visualisations were developed to convey a spatio-temporal overview of pharmaceutical pollution in Luxembourg. Visualising data with four available dimensions (compound, concentration, location, time) is not trivial, and can potentially generate useful insights that may, amongst other things, inform future sampling campaigns for continued environmental monitoring. Advanced heatmaps and boxplots were used to display these data to facilitate comparison across the different pharmaceutical compounds, their locations, times of occurrence, and concentrations.

## Publication B

### Occurrence and Distribution of Pharmaceuticals and Their Transformation Products in Luxembourgish Surface Waters

Singh, R. R.<sup>1</sup>, Lai, A.<sup>2</sup>, Krier, J.<sup>3</sup>, Kondić, T<sup>4</sup>., Diderich, P.<sup>5</sup> & Schymanski, E. L.<sup>6</sup>

DOI: 10.1021/acsenvironau.1c00008

Reprinted with permission from *ACS Environ. Au* 2021, 1, 1, 58–70.  
Copyright 2022 American Chemical Society.

Selected for inclusion in the *ACS Environ. Au* “Rising Stars of 2022” collection.

<b>Author Contributions</b> (Underlined numbers refer to PhD students)						
<b>Author No.</b>	<b>1</b>	<b><u>2</u></b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>
Conceptual Research Design	x				x	x
Planning of Research Activities	x				x	x
Reviewing the Tools	x	x		x		x
Data Collection	x	x	x			
Data Analysis & Interpretation	x	x	x			x
Manuscript Writing	x	x				x
Suggested Publication Equivalence Value		0.5				

# Occurrence and Distribution of Pharmaceuticals and Their Transformation Products in Luxembourgish Surface Waters

Randolph R. Singh,\* Adelene Lai, Jessy Krier, Todor Kondić, Philippe Diderich, and Emma L. Schymanski\*



Cite This: *ACS Environ. Au* 2021, 1, 58–70



Read Online

ACCESS |

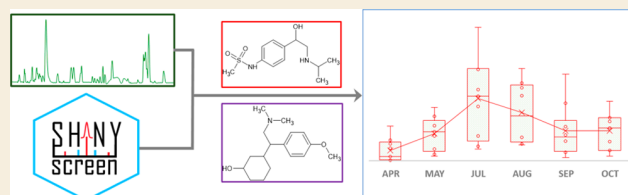
Metrics & More

Article Recommendations

Supporting Information

**ABSTRACT:** Pharmaceuticals and their transformation products (TPs) are continuously released into the aquatic environment via anthropogenic activity. To expand knowledge on the presence of pharmaceuticals and their known TPs in Luxembourgish rivers, 92 samples collected during routine monitoring events between 2019 and 2020 were investigated using nontarget analysis. Water samples were concentrated using solid-phase extraction and then analyzed using liquid chromatography coupled to a high-resolution mass spectrometer. Suspect screening was performed using several open source computational tools and resources including ShinyScreen (<https://git-r3lab.uni.lu/eci/shinyscreen/>), MetFrag (<https://msbi.ipb-halle.de/MetFrag/>), PubChemLite (<https://zenodo.org/record/4432124>), and MassBank (<https://massbank.eu/MassBank/>). A total of 94 pharmaceuticals, 88 confirmed at a level 1 confidence (86 of which could be quantified, two compounds too low to be quantified) and six identified at level 2a, were found to be present in Luxembourg rivers. Pharmaceutical TPs (12) were also found at a level 2a confidence. The pharmaceuticals were present at median concentrations up to 214 ng/L, with caffeine having a median concentration of 1424 ng/L. Antihypertensive drugs (15), psychoactive drugs (15), and antimicrobials (eight) were the most detected groups of pharmaceuticals. A spatiotemporal analysis of the data revealed areas with higher concentrations of the pharmaceuticals, as well as differences in pharmaceutical concentrations between 2019 and 2020. The results of this work will help guide activities for improving water management in the country and set baseline data for continuous monitoring and screening efforts, as well as for further open data and software developments.

**KEYWORDS:** pharmaceuticals, surface water, suspect screening, HRMS, transformation products, cheminformatics, open source, nontarget screening



## INTRODUCTION

The geography and history of Luxembourg have distinct implications on its environment and water quality: it borders Belgium, France, and Germany, and its rivers feed into the Rhine basin. Luxembourg has vineyards lining the Moselle River, agricultural activity in the north of the country, and a population largely centered in the capital, which together brings in a significant and varied chemical load into the environment. Previous studies have reported the presence of analgesics, antimicrobials, and estrogens in Luxembourgish surface water.<sup>1–3</sup> Aside from providing data on the level of xenobiotics in Luxembourgish waters, these studies have also demonstrated that the presence of these chemicals is due to inputs from land use, accidental spillage, wastewater effluent, and long-range transport.<sup>1,3–6</sup> Other studies have reported the measurement of 14 pesticides and their transformation products (TPs) in both surface water and drinking water.<sup>3,7</sup> The Luxembourg Water Management Agency (Administration de la Gestion de l'Eau, hereafter AGE), in compliance with the European Union Water Framework Directive (WFD), monitors different organic contaminants in Luxembourgish

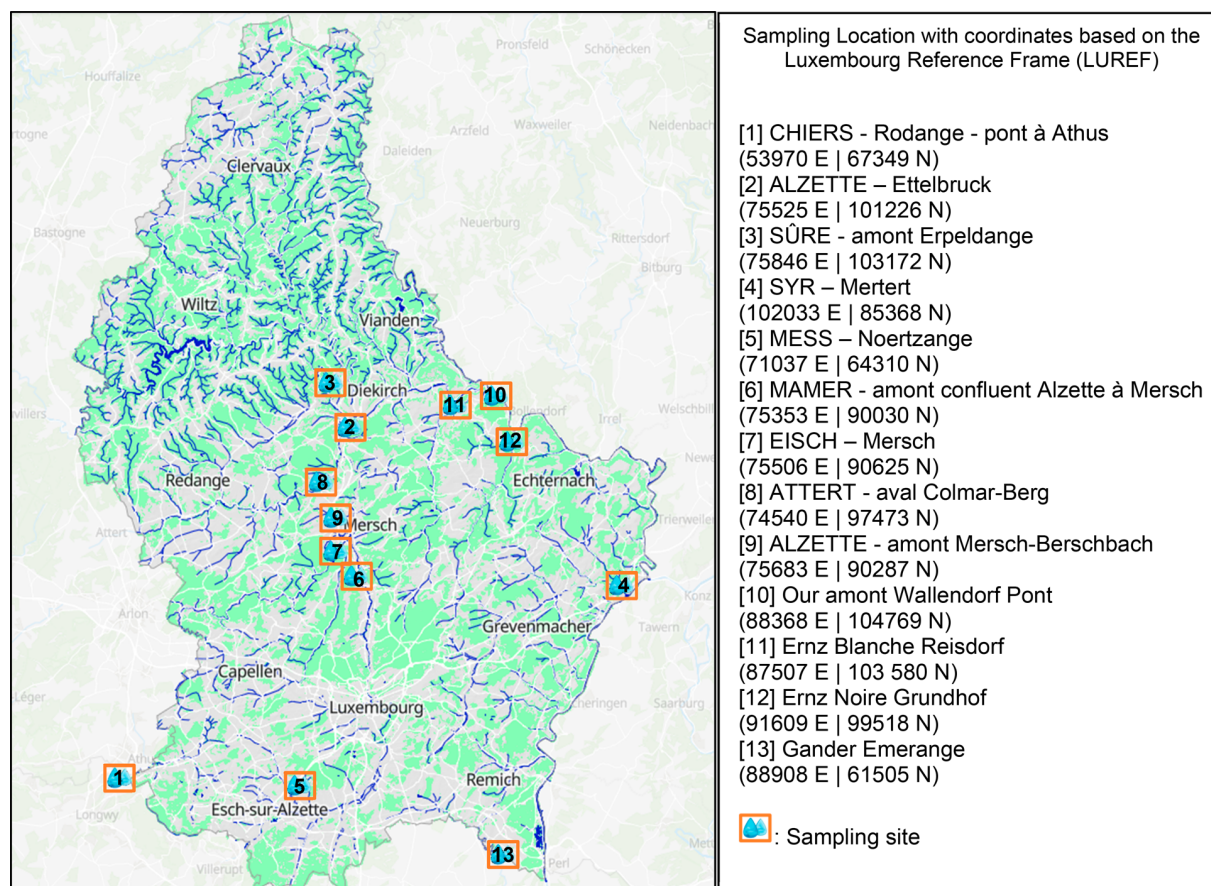
surface water.<sup>8</sup> Among the 92 compounds included in the targeted analysis performed by AGE, five are pharmaceuticals: carbamazepine, diclofenac, ibuprofen, ketoprofen, and lidocaine, while the rest the targeted organic contaminants are pesticides and related compounds.

As there are conceivably more pharmaceuticals than the five included in targeted monitoring that enter into the environment, it is important to determine which other pharmaceuticals may be present, to gain a more holistic idea of the pharmaceutical loading in Luxembourgish surface waters. The presence of pharmaceuticals in the aquatic environment poses a threat to human and environmental health due to exposure to either the pharmaceuticals themselves or their metabolites and TPs, which may still possess bioactivity.<sup>9–11</sup>

Received: May 12, 2021

Published: July 29, 2021





**Figure 1.** Sampling locations and their respective coordinates. Sampling locations 1–4 were sampled from 2019 to 2020; sampling locations 5–9 were sampled only in 2019, and sampling locations 10–13 were sampled only in 2020. Map generated using <https://www.geoportail.lu/en/>. Copyright MapTiler OpenStreetMap contributors.

These chemicals have potential negative impacts on human health and the environment through different routes of exposures.<sup>12,13</sup>

There are many approaches to account for the presence of xenobiotics in the environment, but recently, increasing effort has been in the use of nontargeted analysis (NTA) and/or suspect screening using high-resolution mass spectrometry (HRMS) specifically to support risk assessment efforts and regulatory institutions.<sup>14–16</sup> HRMS enables measurement of known pollutants, discovery of contaminants of emerging concern, as well as retrospective screening.<sup>17</sup> However, setting up analyses, both experimentally and computationally, is no trivial matter. Despite these challenges, the information that can be obtained from such analyses has a wide breadth of utility, especially for environmental studies. NTA and suspect screening are effective techniques for the monitoring and discovery of xenobiotics in the aquatic environment.<sup>17–20</sup> Nevertheless, the interpretation of HRMS data presents challenges that highlight the need for computational tools to enable the proper identification and annotation of the chemical components in environmental matrices.<sup>21</sup>

MetFrag (<https://ipb-halle.github.io/MetFrag/>)<sup>22</sup> is an open source tool for compound identification, including *in silico* fragmentation, mass spectral matching, and metadata functions.<sup>23,24</sup> MetFrag enables spectral matching with experimental data via the spectral library MassBank of North America (MoNA, <https://mona.fiehnlab.ucdavis.edu>)<sup>25</sup> and

prioritization using metadata from various sources. MetFrag first retrieves candidates by exact mass or molecular formula from one of many available compound databases. PubChem (<https://pubchem.ncbi.nlm.nih.gov/>)<sup>26</sup> is an open chemistry database at the National Institutes of Health (NIH) containing more than 110 million compounds.<sup>27</sup> While such a large database provides access to many chemicals, it can lead to (tens of) thousands of candidates per unknown when performing nontarget screening of hundreds of masses.<sup>28</sup> For this work, an early version of PubChemLite was used, which contains ~300,000 compounds selected to be highly relevant for environmental investigations based on annotation content, including information relevant for pharmaceuticals.<sup>28,29</sup> PubChemLite has been shown to outperform other databases such as the whole of PubChem and CompTox for well-known chemicals<sup>28</sup> and delivers important metadata that can be used during identification with MetFrag. PubChem and PubChemLite also contain information on environmental TPs contributed via the NORMAN Suspect List Exchange (<https://www.norman-network.com/nds/SLE/>).<sup>28,30</sup> This information can be exploited programmatically during the environmental screening of hundreds of compounds, together with their transformation products.

Considering the previously reported presence of chemicals in Luxembourg's environment<sup>2,4–7</sup> and the widespread use of chemicals in daily life, a large number of compounds could be considered as potential environmental pollutants in Luxem-



bourg. This work focuses on the presence of pharmaceuticals and known pharmaceutical TPs present in Luxembourg surface water systems using a mixture of instrumental measurements and cheminformatics approaches.

## MATERIALS AND METHODS

### Sample Collection and Processing

Surface water samples (1 L) were collected every 4 weeks, whenever physically possible, from nine different locations in Luxembourg from April to November 2019 (Figure 1) and eight different locations from April to August in 2020 in accordance with the triannual sampling strategy employed at AGE. In this strategy, four locations monitored in compliance with the WFD are consistently sampled every 4 weeks (locations 1–4, Figure 1), while the other locations throughout Luxembourg are divided into three regions and are alternately sampled during a 3 year cycle. The samples were filled in 1000 mL amber glass bottles and stored for up to 1 week at  $5 \pm 3$  °C in the dark until extraction. A method blank was prepared every month to account for potential contamination from sample handling using ultrapure water. Solid-phase extraction (SPE) was performed using Atlantic HLB SPE disks from Horizon (Salem, NH, USA) with a 47 mm diameter. The disks were conditioned twice for 1 min using acetonitrile and then twice for 1 min using Milli-Q water. The samples were pumped through each disk at a flow rate of roughly 30 mL/min, using the SPE-DEX 47900 system from Horizon (Salem, NH, USA). Sample loading was followed by washing the disks twice for 1 min with milli-Q water and drying by airflow for 15 min. The analytes were eluted for 1 min with cyclohexane, followed by an acetone elution for 1 min, then four times for 1 min with acetonitrile. After each elution step, the disks were air-dried for 1 min. The combined extracts were reduced to dryness under nitrogen flow in a water bath heated to 40 °C. The samples were resuspended in 2 mL of acetonitrile/water (10:90) by sonication for 5 min. Remaining particles were removed by passing the extracts through a 0.7  $\mu$ m glass-fiber filter (Sartorius, Brussels, BE) into 2 mL amber glass LC-MS vials. The filtered extracts were stored at  $-20$  °C until analysis.

### LC-HRMS Analysis

LC-HRMS analysis was performed on a Thermo QExactive HF mass spectrometer equipped with a Waters Acquity UPLC BEH C<sub>18</sub> column (1.7  $\mu$ m, 2.1  $\times$  150 mm) using both positive and negative electrospray ionization with the following spray settings (positive/negative): sheath gas flow rate (45/60 arbitrary units, AU), auxiliary gas flow rate (10/25 AU), sweep gas flow rate (2/2 AU), spray voltage (3.5/3.6 kV), capillary temperature (320/300 °C), S lens RF (50/50 AU), and auxiliary gas temperature (300/370 °C). Mobile phases A (water with 0.1% formic acid) and B (methanol) were mixed using the following LC gradient starting at 90A/10B at 0 min, 90/10 at 2 min, 0/100 at 15 min, 0/100 at 20 min, 90/10 at 21 min, and ending with 90/10 at 30 min at a flow rate of 0.200 mL/min. The following data-dependent (dd-)MS2 settings (in display order of instrumental acquisition method) were used: resolution (120,000 at  $m/z$  200), automatic gain control (AGC) target ( $1.0 \times 10^6$ ), maximum injection time (IT): (70 ms), and scan range ( $m/z = 60$ –900). For the selected ion monitoring of dd-MS2/ddSIM, the following were used: resolution (30,000 at  $m/z$  200), AGC target ( $5.0 \times 10^5$ ), maximum IT (70 ms), loop count (5), Top N (5), isolation window (1.0 Da), (N)CE (30). Lastly, the following dd settings were used: minimum AGC target ( $8.0 \times 10^3$ ), intensity threshold ( $1.1 \times 10^5$ ), apex trigger (4–6 s), exclude isotopes (On), and dynamic exclusion (10.0 s). The instrument was calibrated and optimized every time an analysis was performed using manufacturer settings to ensure consistent performance throughout the 2 year study. A 100  $\mu$ g/L standard mixture containing cyclizine, desipramine, nylidrin, amiloride, dibucaine, dothiepin, ethambutol, etofylone, mefruside, phenazone, phentermine, sulfamoxole, sulfamethoxazole, and metoclopramide obtained from Dr. Herbert Oberacher was used to monitor instrument performance between analyses.<sup>31</sup>

### Suspect Screening

Suspect screening was performed using two suspect lists. The first list contains 816 unique pharmaceutical compounds (Supporting Information, Table S1 CNS “Caisse Nationale de Santé” Suspects, also available on the NORMAN Suspect List Exchange, NORMAN-SLE)<sup>30,32</sup> that were curated from the Luxembourgish National Health Fund’s “List of marketed medications in Luxembourg”.<sup>33</sup> These drugs have marketing authorization in Luxembourg from the Ministry of Health and are therefore potentially in use domestically. For suspect screening, MS-ready SMILES of these compounds were obtained via the EPA CompTox Chemistry Dashboard’s batch search function.<sup>34,35</sup> Using MS-ready SMILES as a structural identifier ensures that the structure being used for data analysis is consistent with what is measured by the mass spectrometer and at the same time remains traceable within online chemical databases.<sup>35</sup>

The second suspect list consists of 82 pharmaceutical TPs. These TPs were derived from two sources: PubChem<sup>28</sup> and a recent study by Anliker et al.<sup>18</sup> From PubChem, TPs were obtained from the transformations table of a given compound (where available) using R scripts<sup>36</sup> written to programmatically download transformation product information.<sup>37</sup> The TP information in PubChem originates from the NORMAN Suspect List Exchange.<sup>28,30</sup> Sixty-seven TPs were extracted from PubChem in this way (coming from a total of 53 parents—44 parents were on the original CNS list of 816 parent compounds, while the remaining nine parents are actually themselves TPs with reciprocal transformations). The remaining 15 TPs were obtained from Anliker et al.<sup>18</sup> Curation of the final suspect list involved deduplication and multiple steps of interconversion between chemical identifiers (e.g., CAS to PubChem CID, InChIKey to CID) using PubChem’s Identifier Exchange Service<sup>38</sup> to facilitate compound comparisons and ensure that the final list of 82 TPs was unique. Then, the final SMILES (“parent SMILES” in PubChem terms, “MS-ready” SMILES in CompTox terms) were retrieved. More information and the full R code are available in the Supporting Information and on GitLab as a Jupyter Notebook.<sup>39</sup>

Prescreening was performed using ShinyScreen (<https://git-r3lab.uni.lu/eci/shinyScreen>),<sup>40</sup> an open source and freely available mass spectral processing software developed in house to extract MS1 data and the associated MS2 events and spectra. Detailed information on its functions, installation, and usage can be found by following the link provided above. The following settings for extraction and automatic quality control were used: coarse precursor  $m/z$  error ( $\pm 0.5$  Da), fine precursor  $m/z$  error ( $\pm 2.5$  ppm), extracted ion chromatogram (EIC)  $m/z$  error ( $\pm 0.001$  Da), retention time ( $t_r$ ) tolerance ( $\pm 0.5$  min), MS1 intensity threshold ( $1.0 \times 10^5$ ), MS2 intensity threshold relative to MS1 peak intensity (0.05), signal-to-noise ratio (3), and retention time shift tolerance ( $\pm 0.5$  min). Note that for suspect screening where  $t_r$  information is not available, the  $t_r$  tolerance on the MS1 level is still provided as a setting to ShinyScreen, but the whole chromatogram is screened. For suspect or target chemicals where the  $t_r$  is known from previous analysis (and provided in the input files), this threshold is then applied (e.g., in the suspect confirmation efforts). The “retention time shift” setting at the MS2 level controls the tolerance with regards to alignment of the MS1 and MS2 signals. Features that passed QC through manual curation including peak shape, peak width, peak intensity, and alignment of the MS1 and MS2 peaks were then analyzed using MetFrag to achieve tentative identifications. Scripts used for this work are available on GitLab.<sup>39</sup> PubChemLite was used as database, available as a local .csv file,<sup>29</sup> to find chemicals that match the exact mass (within 5 ppm) of the suspect pharmaceutical. Both in silico fragmentation ( $mzabs = 0.001$ ,  $frag\_ppm = 5$ ) and experimental MS/MS matching through MoNA records (built within MetFrag) were performed to obtain the fragmenter (scoring term 1) and MoNA (scoring term 2) scores. Metadata were also collected for the candidates by querying the database for patent count (scoring term 3), number of PubMed references (scoring term 4), PubChem annotation count (scoring term 5), pharmacology and biochemistry information (scoring term 6), and drug and medication information (scoring term 7). The latter

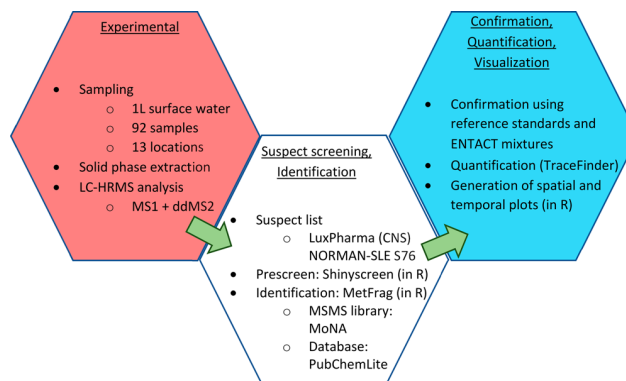
two scoring terms assist in the interpretation of the results where multiple relevant candidates occur per mass, as described recently elsewhere,<sup>28</sup> as well as in the retrieval of classification information (mentioned below). Candidates were ranked and given a score per category normalized to 1 and then added together to obtain the max\_score, with the highest possible score = 7. A more detailed explanation of the parameters used is available elsewhere.<sup>28,41</sup> Annotation confidence levels were determined using the scheme described by Schymanski et al.<sup>42</sup> Level 2a compounds were assigned when the MoNA score was greater than or equal to 0.9. Level 1 identifications were achieved using authentic standards and the ENTACT mixtures,<sup>43</sup> available in-house and analyzed using the same chromatographic method used for sample analysis. The ENTACT mixtures were obtained from participation in the EPA's non-targeted analysis collaborative trial.<sup>43</sup> Retention times were considered a match if the difference was less than 0.2 min. The compound classification for the compounds identified was obtained by consulting PubChem's "Drug and Medication Information" section, based on a specific drug's therapeutic use or function. Level 3 confidence was given for compounds with max\_score > 6.0 but with MoNA scores less than 0.9 (103 compounds); however, the scope of the paper has been limited to level 2a and level 1 chemicals at this stage due to their higher confidence.

Where reference standards were available, the concentration of the pharmaceuticals was quantified using an external calibration curve ranging from 1 to 1000  $\mu\text{g/L}$  spanning the linear dynamic range for the compounds quantified. Tracefinder (Thermo Scientific, version 5.1) was used for automatic peak integration and generation of the calibration curve. Concentrations below 1  $\mu\text{g/L}$  were reported to be below the quantifiable range. With the exception of nonanedioic acid, where the blank comprised <1% of the signal and was subtracted, no interference from the blank was observed for the other analytes identified in this work. After compound identification and quantification, a spatiotemporal analysis was performed to determine whether there were specific areas with higher pharmaceutical loading and/or monthly variability. The concentration of pharmaceuticals in surface waters is influenced by many factors such as matrix, precipitation, volume, wastewater effluent discharge, as well as significant changes in cross-border mobility in 2020 due to the pandemic (a dominating factor in Luxembourg where half of the workforce live outside the country). As a result, the spatial and temporal comparisons are limited to uncorrected concentration values here and should be interpreted accordingly. For spatial analysis, the median concentration of the identified compound across the different months was calculated and presented by sampling year. For temporal analysis, the median concentration of the identified compound across locations 1–4 was used, as these locations were sampled consistently irrespective of sampling year. A boxplot was also constructed to see which pollutants are consistently high and to show the difference in detected concentrations between 2019 and 2020. Heat maps and boxplots were generated using custom-made, openly accessible scripts in R.<sup>44</sup> Results were compared to pharmaceuticals found in the Meuse (Belgian and Dutch section) and Rhine (German section) rivers, which all have Luxembourgish rivers as tributaries. A simplified version of the workflow employed in this work is presented in Figure 2.

## RESULTS AND DISCUSSION

### Identification of Pharmaceuticals and Their TPs

After LC-HRMS analysis coupled with cheminformatics tools was performed, 88 compounds were confirmed at level 1 confidence; 86 of these could be quantified. Amantadine and 8-hydroxyquinoline concentrations were too low to be quantified. A further six compounds were identified at level 2a. These results are summarized in Tables 1 and 2. Among the detected compounds, only seven were detected in both positive and negative ionization: diclofenac, fluconazole, irbesartan, losartan, niflumic acid, oxazepam, and valsartan



**Figure 2.** Flow diagram of the experimental and data processing workflow employed in this work.

(further identifiers are provided in the Supporting Information, Tables S1 and S2). In terms of pharmaceutical class, many of the compounds identified in this work belong to drugs for the management of heart-related diseases (15), psychoactive drugs (15), antimicrobials (eight), and drugs for the management of pain (eight). All five chemicals monitored by AGE were also detected in this study. The number of analytes, including both levels 1 and 2a, found per location in this study ranged from 23 compounds (July 2020) to 52 compounds (May 2019). Thirty-eight pharmaceuticals were detected at least 90% of the time, accounting for 40% of the total compounds identified in this study.

Two TPs (3-hydroxycarbamazepine and *O*-desmethylvenlafaxine) were identified with level 1 confidence, whereas 12 TPs were identified at level 2a confidence and are listed including their parent compounds in parentheses: 4-acetamidoantipyrine (metamizole), 4-aminoantipyrine (metamizole), clopidogrel carboxylic acid (clopidogrel), cotinine (nicotine), D617 (verapamil), ritalinic acid (methylphenydate), fenofibric acid (fenofibrate), flucytosine (emtricitabine), guanylurea (metformin), morphine (codeine), N4-acetylsulfamethoxazole (sulfamethoxazole), 4-hydroxydiclofenac (diclofenac). Flucytosine on its own is used as an antifungal agent, whereas morphine can be used as the parent compound for pain management. In addition, two TPs (2-hydroxycarbamazepine and 10,11-dihydroxycarbamazepine) were tentatively identified (level 3) during the parent pharmaceutical screening because they were isobaric with some parent pharmaceuticals.

### Spatiotemporal Distribution of Pharmaceuticals in Luxembourg

The median concentrations of the different compounds identified in this work, irrespective of ionization polarity, were plotted to generate the spatial ( $N = 6$  time points for 2019,  $N = 5$  time points for 2020) and temporal ( $N = 4$  sampling points) heat maps presented in Figures 3, 4, and 5, respectively. Note that only locations 1–4 were sampled consistently between 2019 and 2020, in compliance with the WFD requirements; thus only data from these locations were used for the temporal analysis. Locations 5–9 were only sampled during 2019, whereas locations 10–13 were sampled in 2020. Tables S3 (negative mode) and S4 (positive mode) in the Supporting Information summarize the individual concentration of each pharmaceutical quantified from 2019 to 2020 from each location. The spatial heat maps (Figures 3 and 4) for both 2019 and 2020 consistently show that Chiers-Rodange-pont à Athus (location 1, Figure 1), followed by Alzette-

**Table 1. Summary of Pharmaceuticals and Pharmaceutical Transformation Products in Positive Mode Found in Luxembourgish River Water<sup>a</sup>**

m/z, [M+H] <sup>+</sup>	Name	t <sub>r</sub> , min	Level	MetFrag Score	MoNA score	PubChem CID
253.097	3-Hydroxycarbamazepine	14.9	1	-	-	135290
152.0706	Acetaminophen	8	1	6.54	0.9998	1983
152.1434	Amantadine	12.1	1	6.96	0.9980	2130
370.1795	Amisulpride	10.7	1	7	0.9993	2159
278.1903	Amitriptyline	14.3	1	5.83	0.9876	2160
267.1703	Atenolol	8.1	1	6.94	0.9363	2249
119.0604	Benzimidazole	2.5	1	6	0.9974	5798
326.2326	Bisoprolol	13.7	1	7	0.9997	2405
195.0877	Caffeine	11.2	1	6.92	0.9970	2519
237.1023	Carbamazepine	15.7	1	7	0.9999	2554
192.0768	Carbendazim	10.2	1	6.97	0.9999	25429
380.2544	Celiprolol	13	1	6.13	0.9934	2663
389.1627	Cetirizine	16.4	1	7	0.9999	2678
748.4842	Clarithromycin	16.2	1	7	0.9985	4663848
425.1871	Clindamycin	14.5	1	7	0.9985	2786
315.1623	Clomipramine	16.3	1	7	0.9993	2801
300.1594	Codeine	9.1	2a	6.81	0.9509	2828
177.1023	Cotinine	2.4	1	5.05	0.9896	408
296.024	Diclofenac*	18.6	1	7	0.9995	3033
415.1686	Diltiazem	14.9	1	7	0.9990	3076
271.1805	Doxylamine	10.4	1	7	0.9986	3162
330.0804	Epoiconazole	18.0	1	-	-	3317081
415.1451	Flecainide	14.3	1	6.85	0.9984	3356
307.1114	Fluconazole*	13	1	6.99	0.9901	3365
821.8876	Iohexol	5.7	1	6.9	0.9062	3730
429.2397	Irbesartan*	16.6	1	7	0.9993	3749
255.1016	Ketoprofen	16.9	1	6.93	0.9546	3825
256.0151	Lamotrigine	12.9	1	6.97	0.9693	3878
235.1805	Lidocaine	11.3	1	6.99	1.0000	3676
407.221	Lincomycin	10.6	1	7	0.9993	3928
321.0192	Lorazepam	16.5	2a	6.77	0.9952	3958
423.1695	Losartan*	16.5	1	7	0.9999	3961
180.1747	Memantine	1	-	-	-	4054
130.1087	Metformin	1.9	1	6.99	0.9856	4091
310.2166	Methadone	15.6	1	6.85	0.975	4095
300.1473	Metoclopramide	11.5	1	7	0.9999	4168
268.1907	Metoprolol	12.4	1	6.94	0.9357	4171
172.0717	Metronidazole	6.9	1	7	0.9983	4173
266.1652	Mirtazapine	12.2	1	6.69	0.9831	4205
269.1051	Moclobemide	11.8	1	7	0.9999	4235
286.1438	Morphine	17.8	2a	5.43	0.9943	4253
231.1016	Naproxen	17.3	1	6.48	0.8966	1302
123.0553	Niacinamide	2.5	1	6.94	0.9871	936
163.123	Nicotine	2.2	1	6.74	0.9952	942
124.0393	Nicotinic acid	2.5	1	5.9	0.9035	938
283.0689	Niflumic acid*	18.8	2a	7	0.9989	4488
264.1958	O-desmethylvenlafaxine	12.2	1	-	-	125017
287.0582	Oxazepam*	16.7	1	6.62	0.9172	4616
384.0824	Pantoprazole	14.6	1	6.98	0.9783	4679
369.2384	Perindopril	15	1	7	0.9996	4169159
189.1023	Phenazone	12.1	1	6.7	0.9578	2206
166.0863	Phenylalanine	6.3	1	6.95	0.9997	994
253.0977	Phenytol	15.5	1	-	-	1775
286.1438	Piperine	17.8	1	5.43	0.9943	4840
260.1645	Propranolol	14.4	1	6.99	0.9932	4946
325.1911	Quinine	11.5	1	6.74	0.8115	1065
315.1486	Ranitidine	7.9	1	6.99	0.9944	3001055
304.1543	Scopolamine	10.1	1	-	-	5184
408.1254	Sitagliptin	12.6	1	6.99	0.9889	11306691
273.1267	Sotalol	6.1	1	6.92	0.9188	5253
251.0597	Sulfadiazine	8	1	6.98	0.9819	5215
254.0594	Sulfamethoxazole	12.1	1	6.94	0.9766	5329
342.1482	Sulpiride	7.1	1	7	0.9996	5355
515.2442	Telmisartan	16.3	1	6.99	0.9929	65999
202.0434	Thiabendazole	11.2	1	6.59	0.9743	5430
329.153	Tiapropride	9	1	7	0.9999	5467
317.1642	Timolol	12.4	1	6.98	0.9840	5478
264.1958	Tramadol	12.3	1	7	0.9999	5523
291.1452	Trimethoprim	10.7	1	6.78	0.9619	5578
436.2343	Valsartan*	17.4	1	6.97	0.9717	5650
278.2115	Venlafaxine	14	1	6.9	0.9925	5656
130.0863	Vigabatrin	1.7	2a	5.32	0.9998	5665
304.202	Vildagliptin	8.1	1	7	0.9999	5251896
309.1122	Warfarin	17.4	1	6.94	0.9410	54678486

<sup>a</sup>An extended version with structural information is available in the Supporting Information Table S2, Pharma IDs. *t<sub>r</sub>* = retention time. \*Found in both positive and negative modes.

Ettelbruck (location 2, Figure 1) and Alzette-Mersch-Berschbach (location 9, Figure 1) that have higher levels of pharmaceutical contamination.

Among the pharmaceuticals found were antihypertensive drugs. In 2019, sotalol and telmisartan were the antihypertensive drugs detected at the highest concentration. In contrast,

**Table 2. Summary of Pharmaceuticals and Pharmaceutical Transformation Products in Negative Mode Found in Luxembourgish River Water<sup>a</sup>**

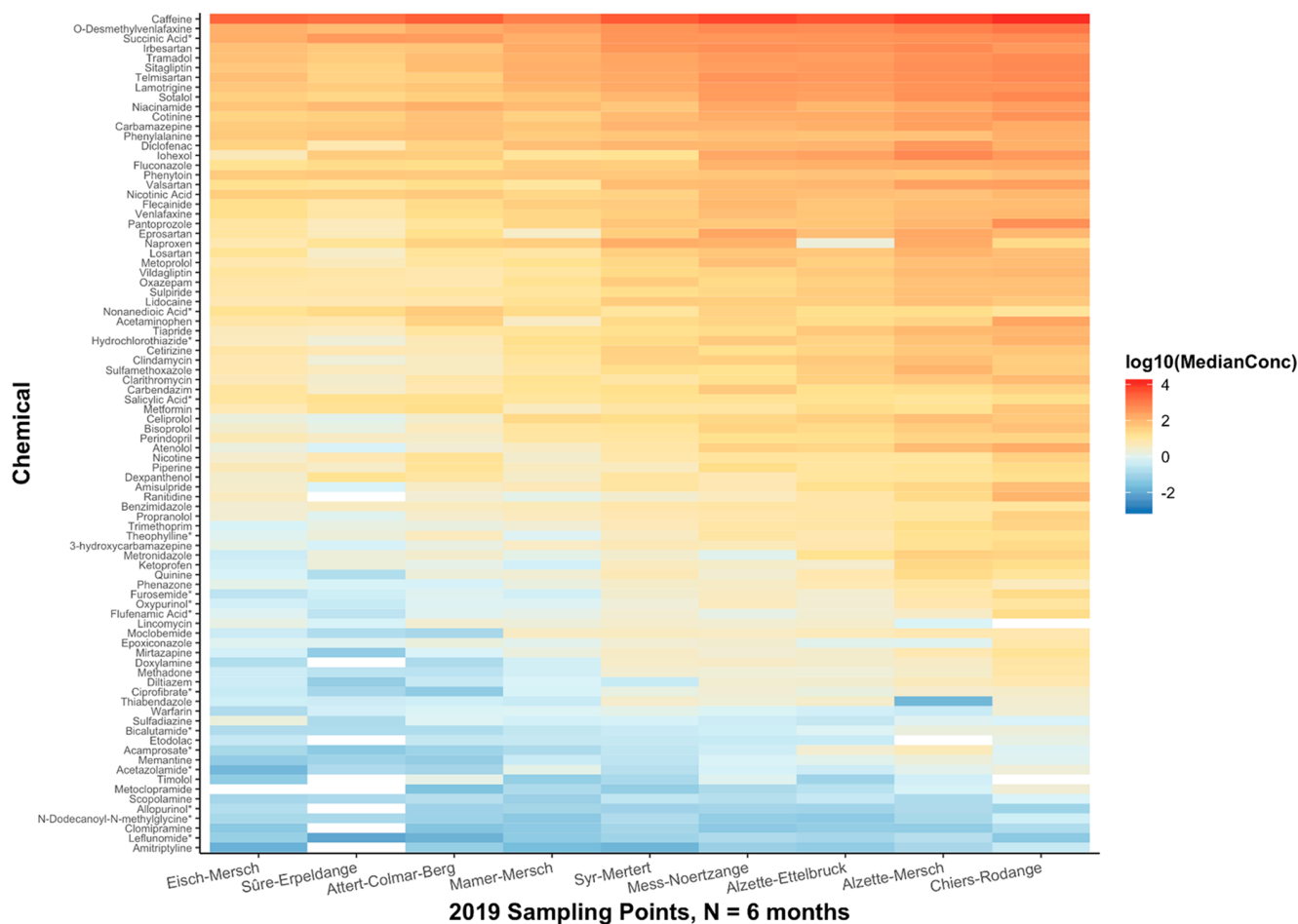
m/z, [M-H] <sup>-</sup>	Name	t <sub>r</sub> , min	level	MetFrag Score	MoNA score	PubChem CID
144.0455	8-Hydroxyquinoline	13.2	1	5.55	0.4714	1923
180.0334	Acamprosate	2.33	1	-	-	71158
220.9809	Acetazolamide	8.5	1	6.99	0.9888	1986
135.0310	Allopurinol	3.46	1	-	-	135401907
179.035	Aspirin	13.7	2a	5.99	0.9964	2244
429.0538	Bicalutamide	16.7	1	6.98	0.9868	2375
287.0247	Ciprofibrate	18.1	1	5.7	0.0000	2763
294.0094	Diclofenac*	18.6	1	7	0.9972	3033
423.1384	Eprosartan	14.2	1	6.83	0.8289	5281037
288.1594	Etodolac	18.5	1	-	-	3308
280.0591	Flufenamic acid	19.2	1	6	0.9998	3371
329.0004	Furosemide	14.7	1	6.94	0.9370	3440
295.9572	Hydrochlorothiazide	8.2	1	7	0.9972	3639
427.2252	Irbesartan*	16.6	1	7	0.9992	3749
269.0543	Leflunomide	17.6	1	5.79	0.0000	3899
421.1549	Losartan*	16.4	1	6.98	0.9844	3961
270.2075	N-Dodecanoyl-N-methylglycine	17.7	1	4.47	0.0000	7348
281.0543	Niflumic acid*	18.8	2a	6.99	0.9944	4488
187.0976	Nonanedioic acid	14.8	1	6	0.0000	2266
285.0436	Oxazepam*	16.6	1	5.27	0.0000	4616
151.0261	Oxypurinol	3.2	1	5.1	0.0246	1188
204.1241	(dex)Panthenol	8.2	1	6.75	0.7626	4678
137.0244	Salicylic acid	14.9	1	7	0.9997	338
117.0193	Succinic acid	3.4	1	6.08	0.9995	1110
179.0574	Theophylline	10.1	1	6.67	0.8883	2153
434.2198	Valsartan*	17.4	1	6.98	0.9830	5650

<sup>a</sup>An extended version with structural information is available in the Supporting Information Table S2, Pharma IDs. *t<sub>r</sub>* = retention time. \*Found in both positive and negative modes.

irbesartan was detected to have the highest concentration during 2020, followed by telmisartan. All three drugs were found to be highest in location 1 (Chiers-Rodange-pont à Athus) followed by location 2 (Alzette-Ettelbruck), irrespective of sampling year. Clarithromycin and clindamycin, on the other hand, were the antimicrobials detected with the highest concentration in 2019, respectively. However, in 2020, sulfamethoxazole and trimethoprim were the highest detected antimicrobials. These drugs are known to be used together for the treatment of bacterial infections. Locations 1 and 2 consistently showed the highest concentrations of the above-mentioned antimicrobials irrespective of year.

The Chiers river receives effluent from the Petange wastewater treatment plant (capacity: 70,000 population equivalents), which is close to the Chiers-Rodange-pont à Athus sampling point. This proximity is likely one of the reasons why Chiers-Rodange-pont à Athus was found to have the highest concentration of pharmaceuticals within this study. In comparison, both Alzette-Ettelbruck and Alzette-Mersch-Berschbach are downstream of the Beggen wastewater treatment plant<sup>45</sup> (capacity: 210,000 population equivalents), which receives sewage from Luxembourg City, the biggest and most populated city in Luxembourg. Despite the bigger capacity, both sampling points are not as close to the source as the Chiers location and thus may experience dilution. The lowest median concentrations for the pharmaceuticals quantified in this study were found at Eisch-Mersch (2019, location 7 in Figure 1), Sûre-amont Erpeldange (2020, location 3, Figure 1), and Our amont Wallendorf Pont (2020, location 10, Figure 1). Pharmaceutical compounds found in this study such as acetaminophen, caffeine, carbamazepine, clarithromycin, salicylic acid, and valsartan have been described before as markers of sewage or wastewater discharge into surface water,<sup>46,47</sup> further supporting the impact of wastewater effluents in Luxembourgish rivers.





**Figure 3.** Spatial heat map showing median concentration values (original units: ng/L) per compound measured per sampling location over 6 months in 2019, plotted using a base-10 logarithmic scale. Median values were calculated across the concentrations measured over the relevant months of sampling for the respective compound and location. Zero-value median concentrations are indicated by gray-shaded boxes. White boxes indicate that there were no concentration values within the quantification range. All compounds were measured in positive mode except for those marked with an asterisk, which were measured in negative mode.

Figures 3–6 show the dynamic nature of pharmaceutical contamination in surface water, demonstrating that aquatic organisms in these rivers are exposed to varying mixtures over time. Since recent studies have highlighted the ecological risks associated with exposure to mixtures in surface water systems,<sup>48,49</sup> this work helps show how suspect screening may support the identification of more chemicals in surface waters and thus help improve the ecological risk assessment of mixtures in future works.

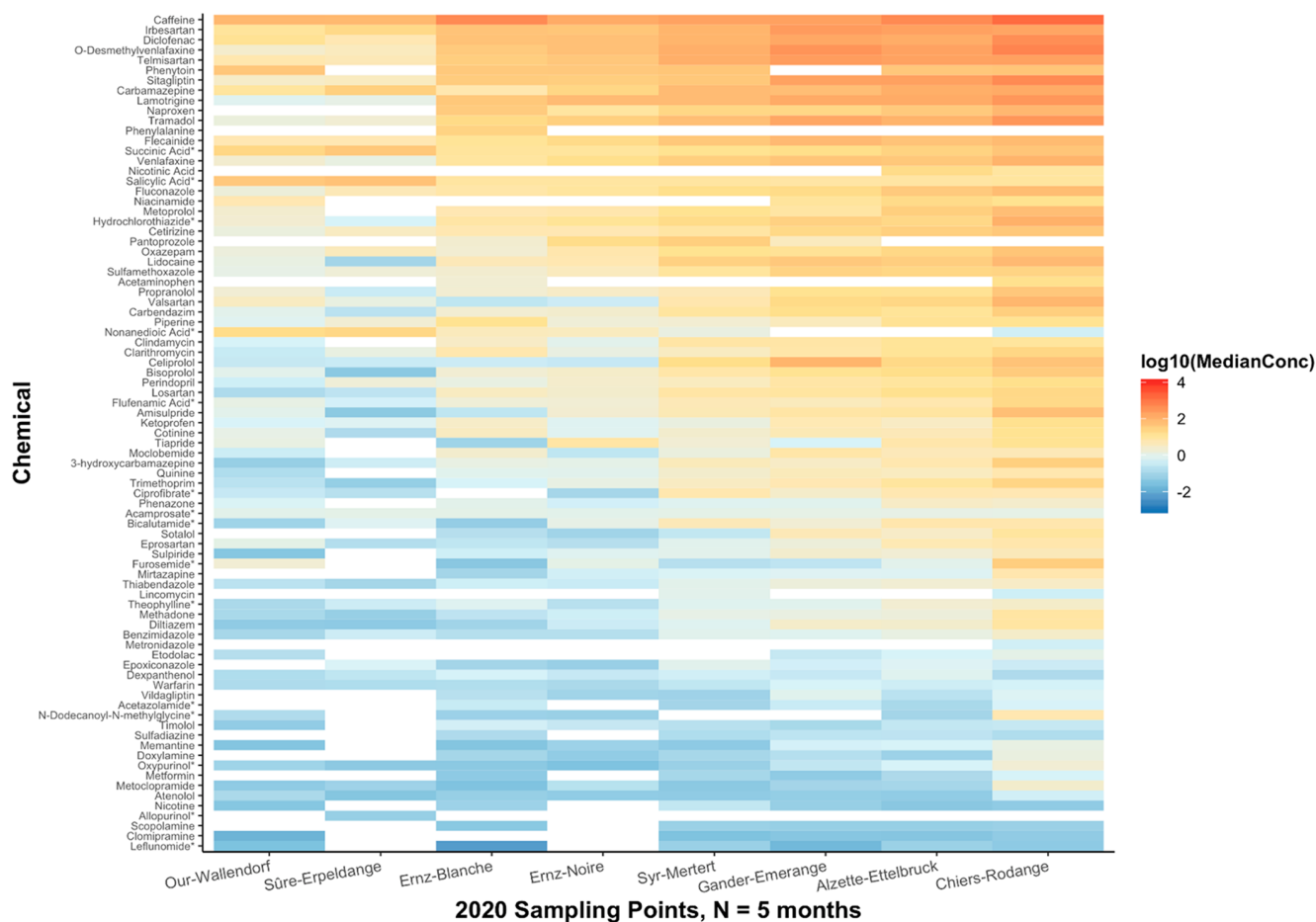
The stimulant caffeine, antidepressant metabolite *O*-desmethylvenlafaxine, antihypertensive drugs irbesartan and telmisartan, the antidiabetic drug sitagliptin, and the opioid analgesic tramadol were among the most concentrated pharmaceuticals found in Luxembourgish surface waters (Figures 3 and 4) in both 2019 and 2020. From a temporal point of view (Figure 5), the highest median concentrations of the pharmaceuticals were detected in September and October of 2019 and are consistently lower during the spring. The most visually obvious differences between the two sampling years include (1) amitriptyline, iohexol, phenylalanine, and ranitidine only detected at quantifiable levels in 2019 and (2) decreases in the median concentrations of dexamethasone, metformin, nicotine, sotalol, and vildagliptin. As an example, metformin had median concentrations of 3.0 ng/L (May) to

39 ng/L (October) in 2019, much higher than the highest detected median concentration of metformin in 2020 (0.62 ng/L in August 2020). Dexamethasone is a drug used for prophylactic purposes; both metformin and vildagliptin are drugs used for managing diabetes, sotalol is for the management of arrhythmia, while nicotine relates to smoking. A juxtaposition of data from 2019 and 2020 is presented as boxplots in Figure 6, showing the general decrease in many pharmaceutical concentrations in 2020 (green boxes). For simplicity, only the top 50 pharmaceuticals ranked by median concentration are presented. Some of the most notable drops in detected concentration were observed for dexamethasone, nicotine, metformin, and sotalol. The individual concentrations of the analytes per sampling location and time are summarized in Tables S3 and S4 in the Supporting Information.

#### Factors That Affected Pharmaceutical Concentrations in Luxembourg

Interestingly, lower median concentrations of the pharmaceuticals were measured in 2020 compared to those measured in 2019 (as shown in Figure 6), which may be partially due to the reduced presence of cross-border workers during the pandemic. COVID-19 has brought on a major shift in working practices, as more people were advised and allowed to work





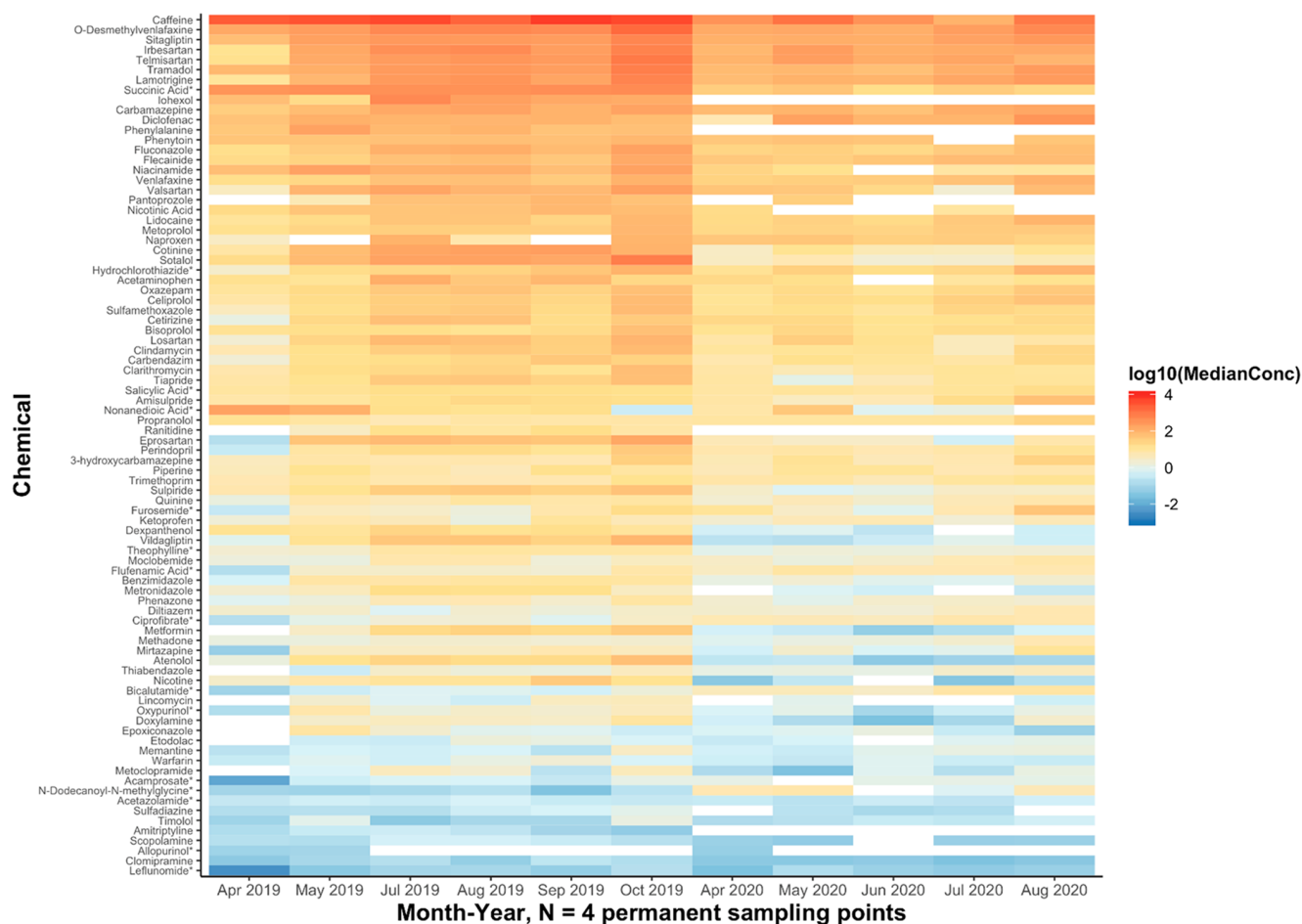
**Figure 4.** Spatial heat map showing median concentration values (original units: ng/L) per compound measured per sampling location over 5 months in 2020, plotted using a base-10 logarithmic scale. Median values were calculated across the concentrations measured over the relevant months of sampling for the respective compound and location. Zero-value median concentrations are indicated by gray-shaded boxes. White boxes indicate that there were no concentration values within the quantification range. All compounds were measured in positive mode except for those marked with an asterisk, which were measured in negative mode.

remotely. In Luxembourg, a major part of the workforce comprises cross-border workers (approximately 206,000 people in 2019).<sup>50</sup> This translates to an approximately 25% decrease in the daytime population, which may translate to reduced pharmaceutical loading in the sewage system. Two interesting features in Figure 6, also apparent in Figure 5, are the detections of iohexol and ranitidine in 2019 but not in 2020. Iohexol is a radiocontrast agent used for medical imaging. Due to the COVID-19 pandemic, there was a significant decrease in medical procedures for noncommunicable diseases, including radio imaging.<sup>51</sup> This decrease may explain why iohexol was not detected at a quantifiable level in 2020 despite having the sixth highest median concentration in 2019. Ranitidine use in the EU, on the other hand, was discontinued in 2020 because of the suspected carcinogen *N*-nitrosodimethylamine, an impurity present in ranitidine drugs.<sup>52</sup> It is interesting to see how changes in drug usage are abruptly reflected in their detection in the environment.

Changes in precipitation had been reported to affect contaminant levels in water, generally increasing with increased precipitation due to factors such as runoff and combined sewer overflow.<sup>53</sup> Compared to the long-term average (1981 to 2010), both 2019 and 2020 experienced a decrease in the annual precipitation (Table 3). For the samplings months that

were studied in both 2019 and 2020 (April, May, July, and August), 2020 showed the lowest amount of precipitation, which may have contributed to the lower concentration of pharmaceuticals detected. While there was not sufficient data available in this study to fully account for all factors influencing the concentration such as population, precipitation, matrix effects, and extraction recoveries, these results reveal interesting trends that will be the subject of further work.

While the Chiers flows into the Meuse River and the Alzette flows into the Sauer River (eventually leading into the Rhine), both rivers contribute to the chemical load that eventually ends up in the North Sea. Several studies have determined the presence of pharmaceuticals in the Meuse and Rhine rivers. A 2010 study by ter Laak et al. reported compounds such as caffeine, carbamazepine, lidocaine, and iohexol as some of the more concentrated pharmaceuticals in their study of the Rhine, with sulfamethoxazole as the most abundant antimicrobial.<sup>54</sup> The same study also found antihypertensive drugs such as atenolol, metoprolol, and sotalol. Despite being apart by almost a decade, similar trends can be observed in Luxembourgish waters. Later studies of different parts of the Rhine and Meuse rivers reported similar pharmaceuticals,<sup>55,56</sup> however, in some studies, the antidiabetic drug metformin and its TP guanylurea were found to be the most abundant



**Figure 5.** Temporal heat map showing median concentration values (original units: ng/L) per compound measured per sampling month–year plotted using a base-10 logarithmic scale. Median values were calculated across the concentrations measured at the four permanent sampling locations for the respective compound and month–year. Zero-value median concentrations are indicated by gray-shaded boxes. White boxes indicate concentration values that were below the respective quantification range, which were therefore discarded from median calculation. All compounds were measured in positive mode except for those marked with an asterisk, which were measured in negative mode.

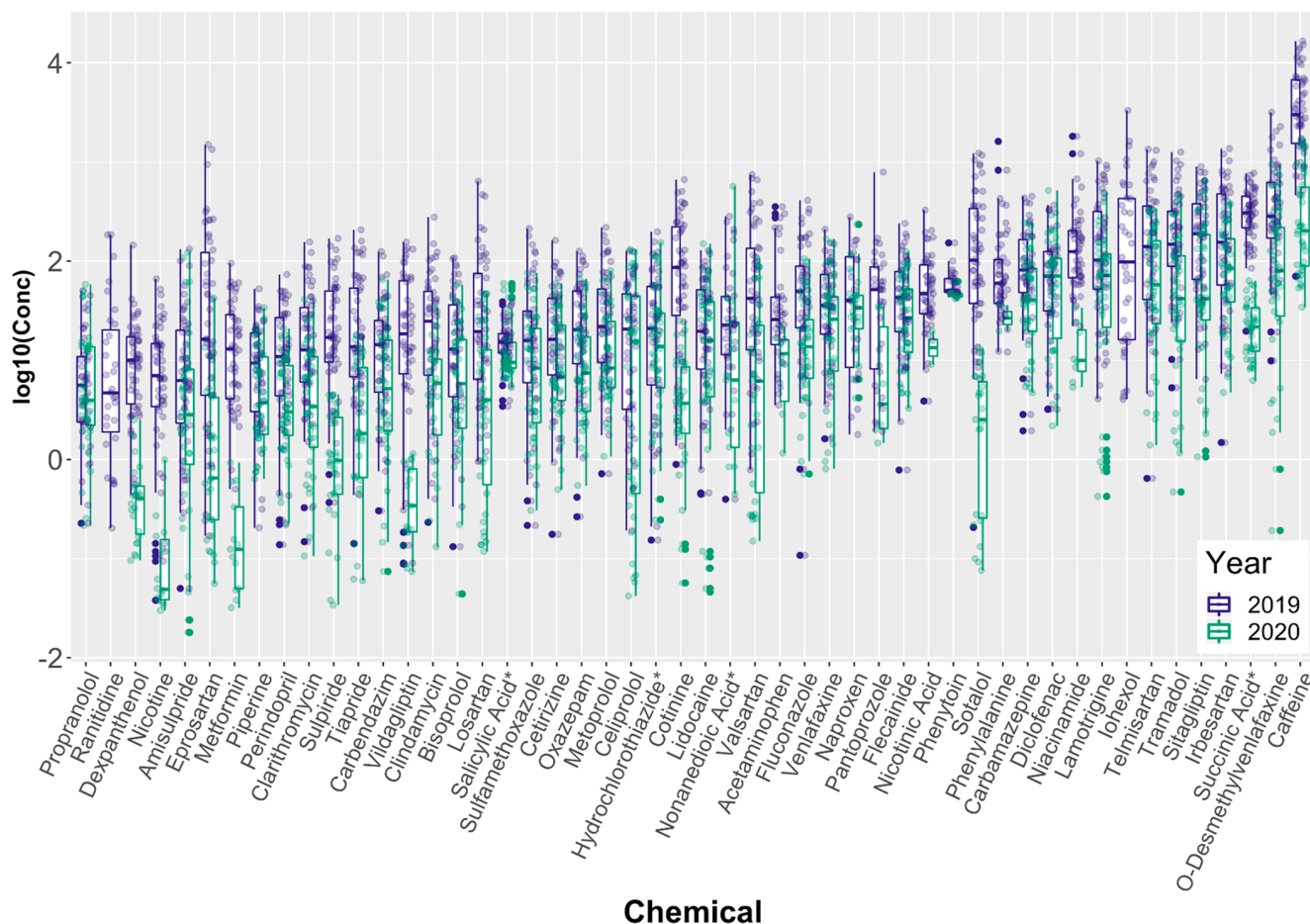
pharmaceutical in surface water samples.<sup>55,57,58</sup> While metformin was also quantified in this study, the median concentration only ranks 44<sup>th</sup> over both years among the pharmaceuticals found. Higher levels of the antidiabetic drug sitagliptin, fifth most abundant, were detected in Luxembourg. The two drugs differ in their mode of regulating sugar in the body.

### Challenges in Compound Identification

The presence of isobars, isomers, and in-source fragments complicates the identification of chemicals in HRMS data, sometimes even leading to these analytes to be excluded from HRMS analysis.<sup>59,60</sup> Several cases of isobars were encountered in this work including (a) acetaminophen and 1,2,3,6-tetrahydrophthalimide, (b) salicylic acid, 3-hydroxybenzoic acid, and 4-hydroxybenzoic acid, (c) piperine, morphine, and etodolac, (d) cocaine and scopolamine, (e) tramadol and *O*-desmethylvenlafaxine, and (f) phenytoin, 2-hydroxycarbamazepine, and 3-hydroxycarbamazepine. While cases a–d were easily resolved using authentic standards, cases e and f introduced specific challenges. Tramadol (parent compound) and *O*-desmethylvenlafaxine (TP of venlafaxine) are constitutional isomers whose extracted ion chromatogram shows two unresolved peaks that are both annotated by MetFrag as

tramadol (due to tramadol's higher metadata scores). Using standards, the first peak (12.2 min) was ultimately assigned to be *O*-desmethylvenlafaxine, while the second peak (12.4 min) was tramadol. In order to quantify both compounds, the peaks had to be manually integrated to avoid integrating the two peaks as one compound.

For the suspect screening of phenytoin, three prominent peaks ( $t_r$ : 13.95, 14.31, and 14.85 min) were observed in the positive mode extracted ion chromatogram of  $m/z$  253.0972 within 5 ppm error (Figure 7A). Looking at the structure of phenytoin, the absence of chiral carbons renders the possibility of diastereomers, which could explain the presence of multiple peaks, invalid. Analysis of the phenytoin standard showed that this compound elutes at 15.53 min, thus not matching any of the three peaks being investigated. Further inspection using MetFrag and database matching suggested that the second and third peaks belong to the positional isomers 2-hydroxycarbamazepine and 3-hydroxycarbamazepine, metabolites of the anticonvulsant carbamazepine. The  $t_r$  matching using a standard confirmed that the peak at 14.85 min is indeed 3-hydroxycarbamazepine, while the peak at 14.31 min can be assigned as 2-hydroxycarbamazepine (level 3), despite the lack of standards, due to the similarity of its mass spectrum with 3-hydroxycarbamazepine. However, the first and biggest peak



**Figure 6.** Boxplots showing the range of concentrations (original units: ng/L) measured for the top 50 highest concentration pharmaceutical chemicals across all months and sampling locations in 2019 and 2020, plotted using a base-10 logarithmic scale. Concentration values that were below the respective quantification ranges were excluded. All chemicals were measured in positive mode.

**Table 3. Precipitation Data for Luxembourg<sup>a</sup>**

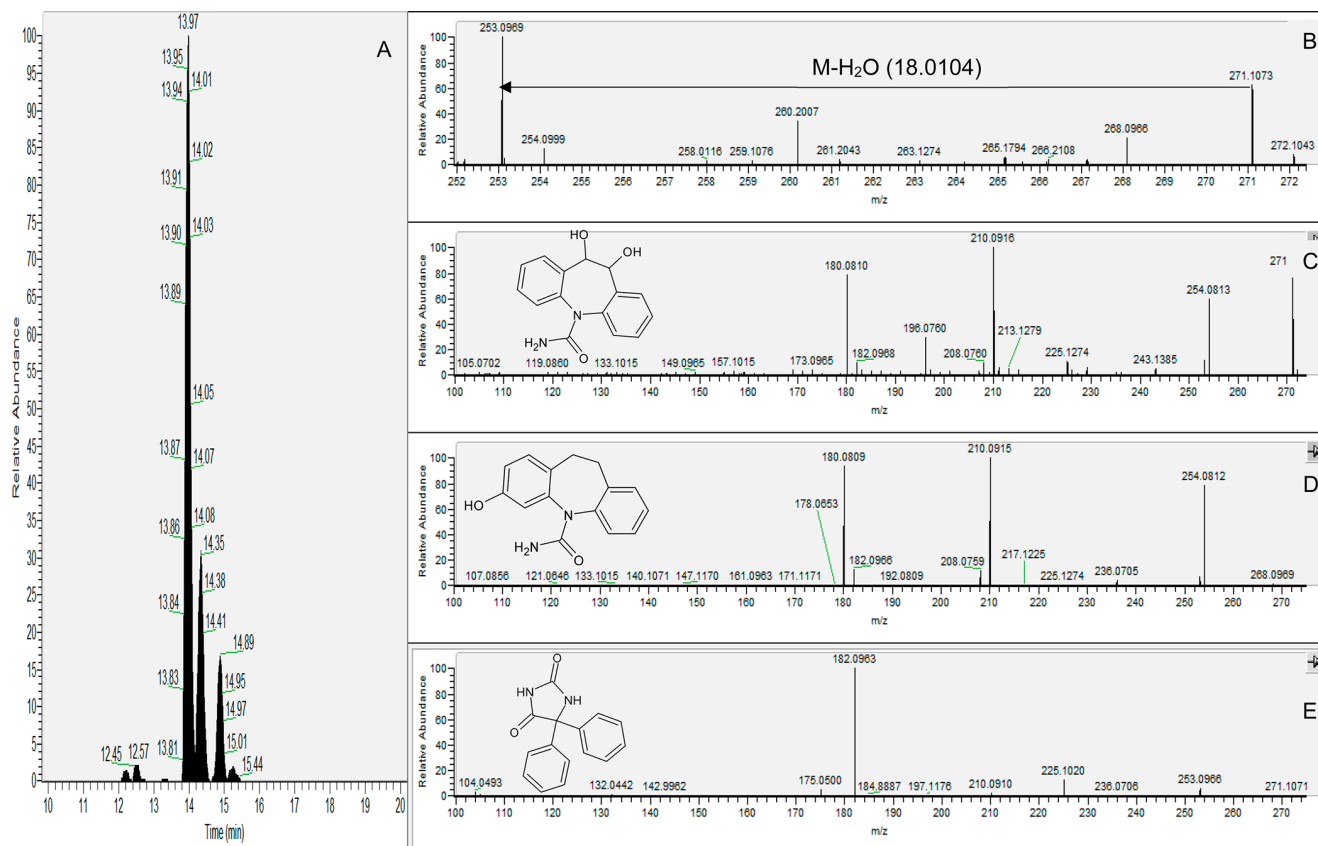
Luxembourg	Precipitation, mm												
	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	Year
Long term average (1981 to 2010)	77	63	69	58	79	80	71	75	76	87	76	87	898
2019	51	43	83	57	61	55	17	51	59	129	88	87	781
2020	47	148	66	20	36	114	8	30	54	113	33	119	788

<sup>a</sup>Source: <https://www.meteolux.lu>.

proved to be challenging. Inspection of the MS1 spectrum at 13.95 min shows another peak with  $m/z$  271.1075 (mass difference equivalent to the loss of water, Figure 7B) can be found whose MS2 spectrum is very similar to the 253.0972 peaks at 14.31 and 14.85 min (Figure 7C,D). Using these pieces of information, it can be suggested that the 253.0972 peak is potentially an in-source fragment of 271.1075. Using 271.1075 as the precursor ion, MetFrag suggests that the peak is potentially 10,11-dihydroxycarbamazepine (MoNA score: 0.8340) or phenytoin acid (MoNA score: 0.8076), which are TPs of carbamazepine and phenytoin, respectively. The presence of the 210.0915 and 180.0811 fragments, which match fragments of other carbamazepine metabolites, and the earlier elution suggesting that the molecule is more polar than the monohydroxylated analogs, supports the tentative identification of the 13.95 min peak as 10,11-dihydroxycarbamazepine (level 3).

One case that needs further inspection is the stereoisomers vidarabine and adenosine, which are impossible to separate using the chromatographic method employed in this study. While there are reports on the utility of ion mobility to discriminate between stereoisomers, it is still to be tested whether such resolution is practically achievable.<sup>61–63</sup> Published collisional cross sections of vidarabine ( $156.4 \text{ \AA}^2$  for  $[M + H]^+$ ) and adenosine ( $156.9 \text{ \AA}^2$  for  $[M + H]^+$ ) measured on the same instrument are available, revealing a difference of only  $0.5 \text{ \AA}^2$  or 0.3%, which is too close to distinguish currently within the typical resolving power of ion mobility spectrometers.<sup>64,65</sup>

This study documents suspect screening efforts thus far for pharmaceuticals and their known TPs as a starting point for further understanding pharmaceutical levels in Luxembourgish surface waters. Other activities looking into different chemical classes such as pesticides,<sup>66</sup> industrial chemicals, and other emerging pollutants are ongoing. The continuous analysis of surface water using HRMS as part of the routine monitoring efforts will enable retrospective screening<sup>67,68</sup> for newly identified contaminants that may impact local surface water quality and biota, such as the effect observed by city runoff on coho salmon.<sup>69</sup> Very recently, a portable HRMS setup for surface water monitoring was demonstrated to enable real-time pollutant analysis,<sup>70</sup> which would be interesting to consider in future efforts pending availability. This study reports primarily



**Figure 7.** (A) Extracted ion chromatogram of  $m/z = 253.0969$  in a surface water sample showing three distinct peaks. (B) MS1 spectrum of the 13.97 peak showing a higher peak that may have lost water to produce the 253.0969 peak. (C) MS2 spectrum of  $m/z = 271.1073$  (potentially 10,11-dihydroxycarbamazepine, structure on the same pane) showing similar fragments to the MS2 fragments of 3-hydroxycarbamazepine standard (structure on the same pane); see (D). (E) MS2 spectrum of the phenytoin standard (structure on the same pane).

level 1 and 2a identifications due to the hard filter of MoNA score of  $>0.9$  applied during the MetFrag analysis. Other tentative identifications have been communicated with AGE, and these, along with more detailed trend analysis as more temporal data points are collected, can be investigated in future works as resources allow. Quantification efforts could be further improved using the list of pharmaceuticals identified in this work as a target list, as well as investing in isotopically labeled standards (which was beyond the scope of the current works, as target analysis is performed by AGE). Finally, as experimental databases increase in size and coverage, the ability to screen for more compounds with higher confidence with these open source methods such as the one presented here will also increase, highlighting the need for the community at large to continue to contribute to publicly available databases.

One main factor limiting TP suspect screening is the lack of available information in open databases that is standardized and thus suitable to be extracted consistently and reproducibly to form meaningful suspect lists. Of the 816 parent compounds on the CNS list, only 44 had associated TP information (i.e., one or more TPs) that could be extracted from PubChem as performed in this study. Certainly, there are far more pharmaceutical metabolites/TPs than those that are identified here, but this information is not yet available in a readily extractable form suitable for an automated workflow within PubChem (the efforts within the NORMAN Suspect List Exchange have just commenced recently).<sup>28,66</sup> As more

information is added and as more environmental transformation studies are performed and deposited in a FAIR (findable, accessible, interoperable and reusable) manner,<sup>71</sup> the ability to screen for TPs in an automated fashion would also increase and support further research efforts.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acsenvironau.1c00008>. The suspect list used in this work is available online as LUXPHARMA (S76) on Zenodo (DOI: 10.5281/zenodo.4587356), CompTox ([https://comptox.epa.gov/dashboard/chemical\\_lists/LUXPHARMA](https://comptox.epa.gov/dashboard/chemical_lists/LUXPHARMA)), PubChem (<https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101>), and NORMAN-SLE (<https://www.norman-network.com/nds/SLE/>). The data (as .mzML files) are available as data set MSV000087190 from the GNPS MassIVE repository (<https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp>), citable under DOI: 10.25345/C5D81C and accessible via <ftp://massive.ucsd.edu/MSV000087190/> and <https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?accession=MSV000087190>. Both ShinyScreen (<https://git-r3lab.uni.lu/eci/shinyScreen/>) and MetFrag (<http://ipb-halle.github.io/MetFrag/>) are open source; additional support scripts mentioned are available from the ECI GitLab repository (<https://git-r3lab.uni.lu/eci/pubchem>). All code used to run MetFrag in the command line using R, generate the Transformation Products suspect list,



and plot Figures 3–6 is available via [https://git-r3lab.uni.lu/adelene.lai/additional\\_si\\_luxpharma\\_singh\\_et\\_al](https://git-r3lab.uni.lu/adelene.lai/additional_si_luxpharma_singh_et_al). All other code and databases used as part of MetFrag identification are likewise openly available (links inline throughout this article).

Tables of CNS suspects, pharma IDs, negative mode, positive mode, positive concentration, and the original file names and their corresponding names in this paper (the original file names were kept to allow traceability to the original sample files stored locally at the University of Luxembourg) (XLSX)

## AUTHOR INFORMATION

### Corresponding Authors

**Randolph R. Singh** – Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 4367 Belvaux, Luxembourg; IFREMER (Institut Français de Recherche pour l'Exploitation de la Mer), Laboratoire Biogéochimie des Contaminants Organiques, Nantes 44311, France; [orcid.org/0000-0003-4500-3400](https://orcid.org/0000-0003-4500-3400); Email: [randolph.singh@ifremer.fr](mailto:randolph.singh@ifremer.fr)

**Emma L. Schymanski** – Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 4367 Belvaux, Luxembourg; [orcid.org/0000-0001-6868-8145](https://orcid.org/0000-0001-6868-8145); Email: [emma.schymanski@uni.lu](mailto:emma.schymanski@uni.lu)

### Authors

**Adelene Lai** – Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 4367 Belvaux, Luxembourg; Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller University, 07743 Jena, Germany; [orcid.org/0000-0002-2985-6473](https://orcid.org/0000-0002-2985-6473)

**Jessy Krier** – Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 4367 Belvaux, Luxembourg

**Todor Kondić** – Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 4367 Belvaux, Luxembourg

**Philippe Diderich** – Administration de la gestion de l'eau, Ministère de l'Environnement, du Climat et du Développement durable, L-2918 Luxembourg, Luxembourg

Complete contact information is available at: <https://pubs.acs.org/10.1021/acsenvironau.1c00008>

### Author Contributions

E.L.S., P.D., and R.R.S. designed the study. P.D. prepared the samples. J.K. and R.R.S. performed instrumental analysis of samples and standards and suspect screening. A.L., E.L.S., and T.K. wrote the code/developed the computational pipeline used. A.L. and R.R.S. generated the figures. R.R.S. drafted the manuscript with contributions from all authors. All authors revised and approved the submitted version.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

A.L., E.L.S., R.R.S., and T.K. acknowledge support by the Luxembourg National Research Fund (FNR) for project A18/BM/12341006. The authors acknowledge the people in the background who tirelessly collect samples and have contributed to this work through healthy discussions in the office,

as well as all those contributing to the many open science efforts mentioned. E.L.S. gratefully acknowledges the contributions of several of the PubChem team to the TP and PubChemLite efforts, including Evan Bolton, Jeff Zhang, and Paul Thiessen (all NIH/NLM/NCBI).

## REFERENCES

- (1) Paillet, J.-Y.; Krein, A.; Pfister, L.; Hoffmann, L.; Guignard, C. Solid phase extraction coupled to liquid chromatography-tandem mass spectrometry analysis of sulfonamides, tetracyclines, analgesics and hormones in surface water and wastewater in Luxembourg. *Sci. Total Environ.* **2009**, *407* (16), 4736–4743.
- (2) Paillet, J.-Y.; Guignard, C.; Meyer, B.; Iffly, J.-F.; Pfister, L.; Hoffmann, L.; Krein, A. Behaviour and fluxes of dissolved antibiotics, analgesics and hormones during flood events in a small heterogeneous catchment in the Grand Duchy of Luxembourg. *Water, Air, Soil Pollut.* **2009**, *203* (1–4), 79–98.
- (3) Meyer, B.; Paillet, J.-Y.; Guignard, C.; Hoffmann, L.; Krein, A. Concentrations of dissolved herbicides and pharmaceuticals in a small river in Luxembourg. *Environ. Monit. Assess.* **2011**, *180* (1), 127–146.
- (4) Krein, A.; Keßler, S.; Meyer, B.; Paillet, J.-Y.; Guignard, C.; Hoffmann, L. Concentrations and loads of dissolved xenobiotics and hormones in two small river catchments of different land use in Luxembourg. *Hydrol Process* **2013**, *27* (2), 284–296.
- (5) Karier, P.; Kraus, G.; Kolber, I. Metazachlor traces in the main drinking water reservoir in Luxembourg: a scientific and political discussion. *Environ. Sci. Eur.* **2017**, *29* (1), 25.
- (6) Schummer, C.; Tuduri, L.; Briand, O.; Appenzeller, B. M.; Millet, M. Application of XAD-2 resin-based passive samplers and SPME-GC-MS/MS analysis for the monitoring of spatial and temporal variations of atmospheric pesticides in Luxembourg. *Environ. Pollut.* **2012**, *170*, 88–94.
- (7) Bohn, T.; Cocco, E.; Gourdol, L.; Guignard, C.; Hoffmann, L. Determination of atrazine and degradation products in Luxembourgish drinking water: origin and fate of potential endocrine-disrupting pesticides. *Food Addit. Contam., Part A* **2011**, *28* (8), 1041–1054.
- (8) E. Commission. Directive 2013/39/EU of the European Parliament and of the Council of 12 August 2013 amending Directives 2000/60/EC and 2008/105/EC as regards priority substances in the field of water policy *Off. J. Eur. Union* **2013**, *226*, 1–17; <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:226:0001:0017:EN:PDF>
- (9) Richardson, S. D.; Fasano, F.; Ellington, J. J.; Crumley, F. G.; Buettner, K. M.; Evans, J. J.; Blount, B. C.; Silva, L. K.; Waite, T. J.; Luther, G. W.; McKague, A. B.; Miltner, R. J.; Wagner, E. D.; Plewa, M. J. Occurrence and mammalian cell toxicity of iodinated disinfection byproducts in drinking water. *Environ. Sci. Technol.* **2008**, *42* (22), 8330–8338.
- (10) Escher, B. I.; Fenner, K. Recent advances in environmental risk assessment of transformation products. *Environ. Sci. Technol.* **2011**, *45* (9), 3835–3847.
- (11) Fenner, K.; Kooijman, C.; Scheringer, M.; Hungerbühler, K. Including transformation products into the risk assessment for chemicals: The case of nonylphenol ethoxylate usage in Switzerland. *Environ. Sci. Technol.* **2002**, *36* (6), 1147–1154.
- (12) Hollender, J.; Rothardt, J.; Radny, D.; Loos, M.; Epting, J.; Huggenberger, P.; Borer, P.; Singer, H. Comprehensive micro-pollutant screening using LC-HRMS/MS at three riverbank filtration sites to assess natural attenuation and potential implications for human health. *Water Res. X* **2018**, *1*, 100007.
- (13) Armnok, P.; Singh, R. R.; Burakham, R.; Pérez-Fuentetaja, A.; Aga, D. S. Selective Uptake and Bioaccumulation of Antidepressants in Fish from Effluent-Impacted Niagara River. *Environ. Sci. Technol.* **2017**, *51* (18), 10652–10662.
- (14) Hollender, J.; van Bavel, B.; Dulio, V.; Farnen, E.; Furtmann, K.; Koschorreck, J.; Kunkel, U.; Krauss, M.; Munthe, J.; Schlabach, M.; Slobodnik, J.; Stroomberg, G.; Ternes, T.; Thomaidis, N. S.; Togola, A.; Tornero, V. High resolution mass spectrometry-based

non-target screening can support regulatory environmental monitoring and chemicals management. *Environ. Sci. Eur.* **2019**, *31* (1), 42.

(15) Sobus, J. R.; Wambaugh, J. F.; Isaacs, K. K.; Williams, A. J.; McEachran, A. D.; Richard, A. M.; Grulke, C. M.; Ulrich, E. M.; Rager, J. E.; Strynar, M. J.; Newton, S. R. Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. *J. Exposure Sci. Environ. Epidemiol.* **2018**, *28* (5), 411–426.

(16) Pourchet, M.; Debrauwer, L.; Klanova, J.; Price, E. J.; Covaci, A.; Caballero-Casero, N.; Oberacher, H.; Lamoree, M.; Damont, A.; Fenaille, F.; Vlaanderen, J.; Meijer, J.; Krauss, M.; Sarigiannis, D.; Barouki, R.; Le Bizec, B.; Antignac, J.-P. Suspect and non-targeted screening of chemicals of emerging concern for human biomonitoring, environmental health studies and support to risk assessment: From promises to challenges and harmonisation issues. *Environ. Int.* **2020**, *139*, 105545.

(17) Brack, W.; Hollender, J.; de Alda, M. L.; Müller, C.; Schulze, T.; Schymanski, E.; Slobodnik, J.; Krauss, M. High-resolution mass spectrometry to complement monitoring and track emerging chemicals and pollution trends in European water resources. *Environ. Sci. Eur.* **2019**, *31* (1), 62.

(18) Anliker, S.; Loos, M.; Comte, R.; Ruff, M.; Fenner, K.; Singer, H. Assessing Emissions from Pharmaceutical Manufacturing Based on Temporal High-Resolution Mass Spectrometry Data. *Environ. Sci. Technol.* **2020**, *54* (7), 4110–4120.

(19) Jernberg, J.; Pellinen, J.; Rantalainen, A.-L. Identification of organic xenobiotics in urban aquatic environments using time-of-flight mass spectrometry. *Sci. Total Environ.* **2013**, *450–451*, 1–6.

(20) Carpenter, C. M. G.; Wong, L. Y. J.; Johnson, C. A.; Helbling, D. E. Fall Creek Monitoring Station: Highly Resolved Temporal Sampling to Prioritize the Identification of Nontarget Micropollutants in a Small Stream. *Environ. Sci. Technol.* **2019**, *53* (1), 77–87.

(21) Blaženović, I.; Kind, T.; Ji, J.; Fiehn, O. Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites* **2018**, *8* (2), 31.

(22) Leibniz Institute for Plant Biochemistry (IPB) Halle. MetFrag Web Interface, 2021; <https://msbi.ipb-halle.de/MetFrag/> (accessed 2021-07-15).

(23) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminf.* **2016**, *8* (1), 3.

(24) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinf.* **2010**, *11* (1), 148.

(25) FiehnLab, University of California Davis. MassBank of North America, 2021; <https://mona.fiehnlab.ucdavis.edu/> (accessed 2021-07-15).

(26) National Center for Biotechnology Information. National Library of Medicine, National Institutes of Health. PubChem, 2021; <https://pubchem.ncbi.nlm.nih.gov/> (accessed 2021-07-15).

(27) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **2021**, *49* (D1), D1388–D1395.

(28) Schymanski, E. L.; Kondić, T.; Neumann, S.; Thiessen, P. A.; Zhang, J.; Bolton, E. E. Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag. *J. Cheminf.* **2021**, *13* (1), 19.

(29) Bolton, E.; Schymanski, E. *PubChemLite tier0 and tier1*, version PubChemLite.0.2.0. Zenodo, 2020; DOI: DOI: 10.5281/zenodo.3611238.

(30) NORMAN Network. NORMAN Suspect List Exchange (NORMAN-SLE), 2021; <https://www.norman-network.com/nds/SLE/> (accessed 2021-07-15).

(31) Oberacher, H.; Sasse, M.; Antignac, J.-P.; Guitton, Y.; Debrauwer, L.; Jamin, E. L.; Schulze, T.; Krauss, M.; Covaci, A.; Caballero-Casero, N.; Rousseau, K.; Damont, A.; Fenaille, F.; Lamoree, M.; Schymanski, E. L. A European proposal for quality control and quality assurance of tandem mass spectral libraries. *Environ. Sci. Eur.* **2020**, *32* (1), 43.

(32) Singh, R. R. S76 LUXPHARMA Pharmaceuticals Marketed in Luxembourg, version NORMAN-SLE-S76.0.1.0. Zenodo, 2021; DOI: DOI: 10.5281/zenodo.4587356.

(33) National Health Fund of Luxembourg (Caisse Nationale de Santé, CNS). List of Commercial Medication in Luxembourg (in French), 2019; <https://cns.public.lu/en/professionnels-sante/medicaments/medicaments-commercialises.html> (accessed 2021-07-15).

(34) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, L.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J. Cheminf.* **2017**, *9* (1), 61.

(35) McEachran, A. D.; Mansouri, K.; Grulke, C.; Schymanski, E. L.; Ruttkies, C.; Williams, A. J. MS-Ready” structures for non-targeted high-resolution mass spectrometry screening studies. *J. Cheminf.* **2018**, *10* (1), 45.

(36) Schymanski, E. *Extract Annotations R Script*, University of Luxembourg, 2020; R3Lab GitLab Pages; <https://git-r3lab.uni.lu/eci/pubchem/-/blob/master/annotations/tps/extractAnnotations.R> (accessed 2021-07-15).

(37) National Center for Biotechnology Information. National Library of Medicine, National Institutes of Health. PubChem Documentation: PubChem Query Syntax, 2020; <https://pubchemdocs.ncbi.nlm.nih.gov/sdq-query-syntax> (accessed 2021-07-15).

(38) National Center for Biotechnology Information. National Library of Medicine, National Institutes of Health. PubChem Identifier Exchange Service, 2020; <https://pubchem.ncbi.nlm.nih.gov/idxexchange/idxexchange.cgi> (accessed 2021-07-15).

(39) Lai, A. *Mining TPs from PubChem. Jupyter Notebook*; University of Luxembourg, 2020; R3Lab GitLab Pages; [https://git-r3lab.uni.lu/adelene.lai/singh\\_et\\_al\\_2020/-/blob/master/Mining\\_TPs\\_from\\_Pubchem.ipynb](https://git-r3lab.uni.lu/adelene.lai/singh_et_al_2020/-/blob/master/Mining_TPs_from_Pubchem.ipynb) (accessed 2021-07-15).

(40) Kondić, T.; Elapavalore, A.; Krier, J.; Lai, A.; Mohammed Taha, H.; Narayanan, M.; Warmoes, M.; Schymanski, E. L. *ShinyScreen*. University of Luxembourg, 2020; R3Lab GitLab Pages; <https://git-r3lab.uni.lu/eci/shinyScreen> (accessed 2021-07-15).

(41) Lai, A.; Singh, R. R.; Kovalova, L.; Jaeggi, O.; Kondić, T.; Schymanski, E. L. Retrospective non-target analysis to support regulatory water monitoring: from masses of interest to recommendations via in silico workflows. *Environ. Sci. Eur.* **2021**, *33* (1), 43.

(42) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **2014**, *48* (4), 2097–2098.

(43) Ulrich, E. M.; Sobus, J. R.; Grulke, C. M.; Richard, A. M.; Newton, S. R.; Strynar, M. J.; Mansouri, K.; Williams, A. J. EPA’s non-targeted analysis collaborative trial (ENTACT): genesis, design, and initial findings. *Anal. Bioanal. Chem.* **2019**, *411* (4), 853–866.

(44) Singh, R. R.; Lai, A.; Krier, J.; Kondić, T.; Diderich, P.; Schymanski, E. L. *Supplemental Information for Occurrence and Distribution of Pharmaceuticals and their Transformation Products in Luxembourgish Surface Waters*. University of Luxembourg, 2021; R3Lab GitLab Pages; [https://git-r3lab.uni.lu/adelene.lai/si\\_luxpharma\\_singh\\_et\\_al](https://git-r3lab.uni.lu/adelene.lai/si_luxpharma_singh_et_al) (accessed 2021-07-15).

(45) The Government of the Grand Duchy of Luxembourg. The National Geoportal of the Grand Duchy of Luxembourg: Wastewater Treatment Plant Overview. Geoportal Luxembourg, 2021; [https://map.geoportail.lu/theme/eau?lang=en&version=3&zoom=12&X=681318&Y=6386989&rotation=0&layers=645&opacities=1&bgLayer=topo\\_bw\\_jpeg](https://map.geoportail.lu/theme/eau?lang=en&version=3&zoom=12&X=681318&Y=6386989&rotation=0&layers=645&opacities=1&bgLayer=topo_bw_jpeg) (accessed 2021-07-15).

(46) Celić, M.; Gros, M.; Farré, M.; Barceló, D.; Petrović, M. Pharmaceuticals as chemical markers of wastewater contamination in the vulnerable area of the Ebro Delta (Spain). *Sci. Total Environ.* **2019**, *652*, 952–963.

(47) Tran, N. H.; Reinhard, M.; Khan, E.; Chen, H.; Nguyen, V. T.; Li, Y.; Goh, S. G.; Nguyen, Q. B.; Saeidi, N.; Gin, K. Y.-H. Emerging contaminants in wastewater, stormwater runoff, and surface water:

Application as chemical markers for diffuse sources. *Sci. Total Environ.* **2019**, *676*, 252–267.

(48) Zhi, H.; Kolpin, D. W.; Klaper, R. D.; Iwanowicz, L. R.; Meppelink, S. M.; LeFevre, G. H. Occurrence and spatiotemporal dynamics of pharmaceuticals in a temperate-region wastewater effluent-dominated stream: variable inputs and differential attenuation yield evolving complex exposure mixtures. *Environ. Sci. Technol.* **2020**, *54* (20), 12967–12978.

(49) Topaz, T.; Boxall, A.; Suari, Y.; Egozi, R.; Sade, T.; Chefetz, B. Ecological Risk Dynamics of Pharmaceuticals in Micro-Estuary Environments. *Environ. Sci. Technol.* **2020**, *54* (18), 11182–11190.

(50) The Government of the Grand Duchy of Luxembourg. The Statistics Portal of the Grand Duchy of Luxembourg: Overview of the Labour Market 2000–2020 (in French). Statistics Portal Luxembourg, 2021; [https://statistiques.public.lu/stat/TableViewer/tableView.aspx?ReportId=12951&IF\\_Language=fr&MainTheme=2&FldrName=3](https://statistiques.public.lu/stat/TableViewer/tableView.aspx?ReportId=12951&IF_Language=fr&MainTheme=2&FldrName=3) (accessed 2021-07-15).

(51) Cavallo, J. J.; Forman, H. P. The Economic Impact of the COVID-19 Pandemic on Radiology Practices. *Radiology* **2020**, *296* (3), E141–E144.

(52) European Medicines Agency. Suspension of ranitidine medicines in the EU, 2020; <https://www.ema.europa.eu/en/news/suspension-ranitidine-medicines-eu> (accessed 2021-07-15).

(53) Zhu, L.; Jiang, C.; Panthi, S.; Allard, S. M.; Sapkota, A. R.; Sapkota, A. Impact of high precipitation and temperature events on the distribution of emerging contaminants in surface water in the Mid-Atlantic, United States. *Sci. Total Environ.* **2021**, *755*, 142552.

(54) ter Laak, T. L.; van der Aa, M.; Houtman, C. J.; Stoks, P. G.; van Wezel, A. P. Relating environmental concentrations of pharmaceuticals to consumption: A mass balance approach for the river Rhine. *Environ. Int.* **2010**, *36* (5), 403–409.

(55) Ruff, M.; Mueller, M. S.; Loos, M.; Singer, H. P. Quantitative target and systematic non-target analysis of polar organic micro-pollutants along the river Rhine using high-resolution mass-spectrometry – Identification of unknown sources and compounds. *Water Res.* **2015**, *87*, 145–154.

(56) de Jongh, C. M.; Kooij, P. J. F.; de Voogt, P.; ter Laak, T. L. Screening and human health risk assessment of pharmaceuticals and their transformation products in Dutch surface waters and drinking water. *Sci. Total Environ.* **2012**, *427–428*, 70–77.

(57) ter Laak, T. L.; Kooij, P. J. F.; Tolcamp, H.; Hofman, J. Different compositions of pharmaceuticals in Dutch and Belgian rivers explained by consumption patterns and treatment efficiency. *Environ. Sci. Pollut. Res.* **2014**, *21* (22), 12843–12855.

(58) Houtman, C. J.; ten Broek, R.; de Jong, K.; Pieterse, B.; Kroesbergen, J. A multicomponent snapshot of pharmaceuticals and pesticides in the river Meuse basin. *Environ. Toxicol. Chem.* **2013**, *32* (11), 2449–2459.

(59) Sobus, J. R.; Grossman, J. N.; Chao, A.; Singh, R.; Williams, A. J.; Grulke, C. M.; Richard, A. M.; Newton, S. R.; McEachran, A. D.; Ulrich, E. M. Using prepared mixtures of ToxCast chemicals to evaluate non-targeted analysis (NTA) method performance. *Anal. Bioanal. Chem.* **2019**, *411* (4), 835–851.

(60) Singh, R. R.; Chao, A.; Phillips, K. A.; Xia, X. R.; Shea, D.; Sobus, J. R.; Schymanski, E. L.; Ulrich, E. M. Expanded coverage of non-targeted LC-HRMS using atmospheric pressure chemical ionization: a case study with ENTACT mixtures. *Anal. Bioanal. Chem.* **2020**, *412*, 4931–4939.

(61) Colson, E.; Decroo, C.; Cooper-Shepherd, D.; Caulier, G.; Henoumont, C.; Laurent, S.; De Winter, J.; Flammang, P.; Palmer, M.; Claereboudt, J.; Gerbaux, P. Discrimination of regioisomeric and stereoisomeric saponins from *Aesculus hippocastanum* seeds by ion mobility mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (11), 2228–2237.

(62) McCooeye, M.; Ding, L.; Gardner, G. J.; Fraser, C. A.; Lam, J.; Sturgeon, R. E.; Mester, Z. Separation and quantitation of the stereoisomers of ephedra alkaloids in natural health products using flow injection-electrospray ionization-high field asymmetric waveform

ion mobility spectrometry-mass spectrometry. *Anal. Chem.* **2003**, *75* (11), 2538–2542.

(63) Hofmann, J.; Hahm, H. S.; Seeberger, P. H.; Pagel, K. Identification of carbohydrate anomers using ion mobility–mass spectrometry. *Nature* **2015**, *526* (7572), 241–244.

(64) Hines, K. M.; Ross, D. H.; Davidson, K. L.; Bush, M. F.; Xu, L. Large-Scale Structural Characterization of Drug and Drug-Like Compounds by High-Throughput Ion Mobility-Mass Spectrometry. *Anal. Chem.* **2017**, *89* (17), 9023–9030.

(65) Celma, A.; Sancho, J. V.; Schymanski, E. L.; Fabregat-Safont, D.; Ibáñez, M.; Goshawk, J.; Barknowitz, G.; Hernández, F.; Bijlsma, L. Improving Target and Suspect Screening High-Resolution Mass Spectrometry Workflows in Environmental Analysis by Ion Mobility Separation. *Environ. Sci. Technol.* **2020**, *54* (23), 15120–15131.

(66) Krier, J.; Singh, R.; Kondic, T.; Lai, A.; Diderich, P.; Zhang, J.; Thiessen, P.; Bolton, E.; Schymanski, E. Discovering Pesticides and their Transformation Products in Luxembourg Waters using Open Cheminformatics Approaches. *Research Square* **2021**, DOI: 10.21203/rs.3.rs-478324/v1.

(67) Creusot, N.; Casado-Martinez, C.; Chiaia-Hernandez, A.; Kiefer, K.; Ferrari, B. J. D.; Fu, Q.; Munz, N.; Stamm, C.; Tlili, A.; Hollender, J. Retrospective screening of high-resolution mass spectrometry archived digital samples can improve environmental risk assessment of emerging contaminants: A case study on antifungal azoles. *Environ. Int.* **2020**, *139*, 105708.

(68) Alygizakis, N. A.; Samanipour, S.; Hollender, J.; Ibanez, M.; Kaserzon, S.; Kokkali, V.; van Leerdam, J. A.; Mueller, J. F.; Pijnappels, M.; Reid, M. J.; Schymanski, E. L.; Slobodnik, J.; Thomaidis, N. S.; Thomas, K. V. Exploring the potential of a global emerging contaminant early warning network through the use of retrospective suspect screening with high-resolution mass spectrometry. *Environ. Sci. Technol.* **2018**, *52* (9), 5135–5144.

(69) Tian, Z.; Zhao, H.; Peter, K. T.; Gonzalez, M.; Wetzel, J.; Wu, C.; Hu, X.; Prat, J.; Mudrock, E.; Hettinger, R.; Cortina, A. E.; Biswas, R. G.; Kock, F. V. C.; Soong, R.; Jenne, A.; Du, B.; Hou, F.; He, H.; Lundeen, R.; Gilbreath, A.; Sutton, R.; Scholz, N. L.; Davis, J. W.; Dodd, M. C.; Simpson, A.; McIntyre, J. K.; Kolodziej, E. P. A ubiquitous tire rubber-derived chemical induces acute mortality in coho salmon. *Science* **2021**, *371* (6525), 185–189.

(70) Stravs, M. A.; Stamm, C.; Ort, C.; Singer, H. Transportable Automated HRMS Platform “MS2field” Enables Insights into Water-Quality Dynamics in Real Time. *Environ. Sci. Technol. Lett.* **2021**, *8* (5), 373–380.

(71) GO FAIR. Fair Principles, 2021; <https://www.go-fair.org/fair-principles/> (accessed 2021-07-15).

The Level 1 identification of 2 TPs, Level 2a identification of the 12 TPs, and Level 3 identification of 2 TPs in this study is particularly significant because it proves the notion that these species do exist in the environment and that only considering parent compounds would neglect the 'big picture' of environmental chemical pollution and subsequent exposures. Thus, including TPs as suspects in future screenings is imperative, particularly as our collective knowledge of TP processes and compounds grows and becomes more readily accessible in a systematic manner, as was the case in this study via querying PubChem. All spectral data, suspect lists, and R code developed to mine TP information from PubChem, plus further analysis (including visualisations, described below), are openly available via the respective URL links listed in the publication.

Spatio-temporal analyses via data visualisation compared the occurrence and concentrations of the different pharmaceuticals, which allowed for numerous observations that could be explained by epidemiological and social phenomena. For example, the highest levels of pharmaceutical pollution overall were recorded in Chiers-Rodange, Alzette-Ettelbruck, and Alzette-Mersch, which likely reflects high loads of wastewater due to the high population density in those regions. Further observations from these analyses include: pharmaceutical concentrations in 2020 were lower overall than in 2019, possibly due to the drop in office workers commuting during COVID-19 pandemic-induced quarantine; iohexol was lower in concentration in 2020 than 2019 likely because of reductions in radioimaging due to the pandemic; and lower concentrations of ranitidine in 2020 compared to 2019, as ranitidine was suspended by the European Medicines Agency in 2020.

Critically, the consideration of TPs and data visualisations developed for the spatio-temporal analyses generated multiple insights that not only serve as additional validation of the identifications and quantifications achieved, but that could also guide future water monitoring campaigns and regulatory activities in Luxembourg. For example, regulatory screenings could be expanded to include pharmaceuticals and their TPs instead of just the chemicals officially listed in the WFD. Additionally, candidate locations for potential upgrades to wastewater treatment systems may be proposed based on the spatio-temporal analysis.



Nevertheless, this study only dealt with single compounds. In terms of pharmaceuticals, the compounds studied here were the active ingredients that convey most of the potent effects, but in reality, pharmaceutical products are delivered as formulations, which are essentially mixtures. In fact, nearly all chemical products, virtually all environmental samples, and thus sources of chemical exposure to humans are mixtures. Environmental chemical mixtures represent an active but extremely challenging area of research because of the sheer variability of mixtures that can possibly exist. However, one tractable form of mixture exists within regulatory frameworks, namely a specific class of chemicals called substances of Unknown or Variable composition, Complex reaction products, or Biological materials (UVCBs). Regulators find assessing the risks of UVCBs particularly challenging compared to single compounds because multiple confounding factors, not least their ambiguous or unknown identities, come into play that make these substances difficult to deal with. Thus, a holistic review of UVCBs, their properties, and their challenges is warranted.

## Chapter 4

# Tackling the Next Frontier of Environmental Unknowns - UVCBs

As demonstrated in the previous two chapters, the ability to identify all the unknown chemicals in a given environmental sample is still elusive, despite using state-of-the-art open software, tools, and in particular environmental chemical lists and databases. Often, chemical registries are an important source of information when trying to identify environmental chemical unknowns. However, at least 20-40% of substances within these registries have ambiguous or non-existent chemical structures associated with them. These substances are classified within regulatory frameworks as having Unknown or Variable composition, Complex reaction products, or Biological materials (UVCB).

In many regulatory frameworks around the world, UVCB substances are subject to chemicals assessment. However, the fact that their chemical identities are mostly unknown or ambiguous makes it very challenging for regulators to assess these substances. Multiple interconnected and interdisciplinary issues contribute to the challenge of dealing with UVCBs, which calls for a holistic overview of these substances, their characteristics, and a critical review of strategies for assessing and managing these substances. The work in this chapter is the first of its kind in terms of its breadth and interdisciplinarity, as it goes beyond discussing risk assessment and additionally addresses topics in analytical chemistry, toxicity, and mixture cheminformatics, towards developing enhanced methodologies for dealing with UVCBs.

## Publication C

### The Next Frontier of Environmental Unknowns: Substances of Unknown or Variable Composition, Complex Reaction Products, or Biological Materials (UVCBs)

Lai, A.<sup>1</sup>, Clark, A.M.<sup>2</sup>, Escher, B. I.<sup>3</sup>, Fernandez, M.<sup>4</sup>, McEwen, L. R.<sup>5</sup>, Tian, Z.<sup>6</sup>, Wang, Z.<sup>7</sup>, & Schymanski, E. L.<sup>8</sup>

DOI: 10.1021/acs.est.2c00321

Reprinted with permission from *Environ. Sci. Technol.* 2022, 56, 12, 7448–7466.  
Copyright 2022 American Chemical Society.

<b>Author Contributions</b> (Underlined numbers refer to PhD students)								
<b>Author No.</b>	<b><u>1</u></b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>
Conceptual Research Design	x							x
Planning of Research Activities	x							x
Reviewing the Tools	x	x	x	x	x	x	x	x
Data Analysis & Interpretation	x							
Manuscript Writing	x	x	x	x	x	x	x	x
Suggested Publication Equivalence Value	0.5							

# The Next Frontier of Environmental Unknowns: Substances of Unknown or Variable Composition, Complex Reaction Products, or Biological Materials (UVCBs)

Adelene Lai,\* Alex M. Clark, Beate I. Escher, Marc Fernandez, Leah R. McEwen, Zhenyu Tian, Zhanyun Wang, and Emma L. Schymanski\*



Cite This: *Environ. Sci. Technol.* 2022, 56, 7448–7466



Read Online

ACCESS |



Metrics & More



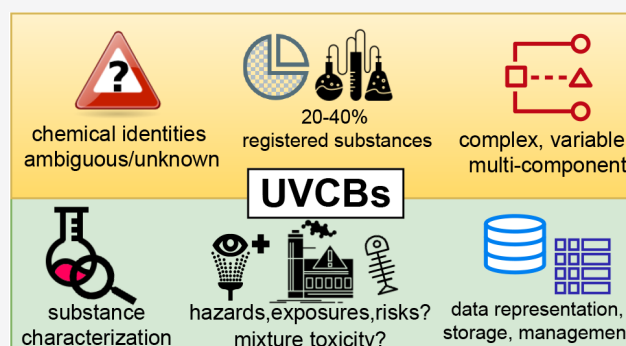
Article Recommendations



Supporting Information

**ABSTRACT:** Substances of unknown or variable composition, complex reaction products, or biological materials (UVCBs) are over 70 000 “complex” chemical mixtures produced and used at significant levels worldwide. Due to their unknown or variable composition, applying chemical assessments originally developed for individual compounds to UVCBs is challenging, which impedes sound management of these substances. Across the analytical sciences, toxicology, cheminformatics, and regulatory practice, new approaches addressing specific aspects of UVCB assessment are being developed, albeit in a fragmented manner. This review attempts to convey the “big picture” of the state of the art in dealing with UVCBs by holistically examining UVCB characterization and chemical identity representation, as well as hazard, exposure, and risk assessment. Overall, information gaps on chemical identities underpin the fundamental challenges concerning UVCBs, and better reporting and substance characterization efforts are needed to support subsequent chemical assessments. To this end, an information level scheme for improved UVCB data collection and management within databases is proposed. The development of UVCB testing shows early progress, in line with three main methods: whole substance, known constituents, and fraction profiling. For toxicity assessment, one option is a whole-mixture testing approach. If the identities of (many) constituents are known, grouping, read across, and mixture toxicity modeling represent complementary approaches to overcome data gaps in toxicity assessment. This review highlights continued needs for concerted efforts from all stakeholders to ensure proper assessment and sound management of UVCBs.

**KEYWORDS:** mixtures, UVCB, complex substances, testing and assessment, cheminformatics, environmental pollutants



## 1. INTRODUCTION

Anthropogenic chemical pollution is pervasive and has been found in multiple environments,<sup>1–5</sup> animals,<sup>6–9</sup> and humans<sup>10–14</sup> worldwide, with at least 16% of global premature deaths attributed to diseases caused by pollution.<sup>15</sup> Chemical pollutants originate from the production, use, and disposal of diverse chemical products. The most familiar and well-studied are single chemical compounds, but these form only a part of the bigger picture of chemical pollution. In practice, many pollutants come from chemical products consisting of mixtures. While some of these mixtures are well-defined, many are poorly characterized or contain constituents with unknown or variable chemical identities, and they are classified as substances of unknown or variable composition, complex reaction products, or biological materials (UVCBs).

UVCBs are considered chemical substances within multiple legal frameworks,<sup>16–18</sup> and thus they are subject to various registration, hazard evaluation, and risk assessment require-

ments. UVCBs can be found everywhere: within detergents, fragrances, and personal care products, and even within fuel and starting materials for chemical manufacturing. A broad range of substances are considered UVCBs, e.g., those of natural origin such as petroleum fractions and essential oils, synthetic products such as technical mixtures of specialty copolymers, and reaction products such as medium-chain chlorinated paraffins (MCCPs; CASRN 85535-85-9) and substances such as “Rape oil, reaction products with diethylenetriamine” (CASRN 91081-13-9; all UVCBs mentioned in this review are detailed in Table S1). As such, UVCBs may contain structurally similar (e.g., isomers,

Received: January 14, 2022

Revised: April 2, 2022

Accepted: April 4, 2022

Published: May 9, 2022



homologues, congeners), or entirely dissimilar chemical constituents. Variations in their composition may arise from fluctuations in production processes, starting materials, or the presence of transformation products formed from spontaneous reactions.

UVCBs are highly prevalent on the global market: 20–40% of chemicals registered in the European Union and in the United States comprise UVCBs.<sup>19–21</sup> A recent global inventory found over 70 000 UVCBs and polymers within over 235 000 registered chemicals with Chemical Abstracts Service Registry Numbers (CASRNs).<sup>22</sup> Additionally, many UVCBs are produced and used at high volumes globally. Annual production of MCCPs in China alone was estimated to be 600 000 t in 2013,<sup>23</sup> and 1027 million metric tons of petroleum substances were manufactured or imported into the European Union in 2018.<sup>24</sup>

Given their significant proportion within chemical registries, high production volumes, and wide usage patterns, UVCBs are highly environmentally relevant. While certain UVCBs such as linear alkylbenzenesulfonate surfactants were found at high intensity in wastewater,<sup>25</sup> the chemical identities of most UVCBs remain unknown or poorly characterized. These critical information gaps limit their detection and identification in the environment and biota, and hinder assessment of their hazards and risks, particularly as most existing testing methods were originally designed for discrete compounds. Meanwhile, current information systems and cheminformatic representations are ill-equipped to store, index, and retrieve information on UVCBs from databases. Consequently, UVCBs are commonly omitted from scientific studies for the sake of simplicity,<sup>26–28</sup> and regulators around the world face challenges in assessing and managing their environmental and health risks.<sup>29</sup>

Rather than tackle UVCBs as a substance *class*, previous reviews focused on specific substances using a single disciplinary lens: e.g., analytical characterization of chondroitin sulfate<sup>30</sup> and surfactants,<sup>31</sup> health assessment of endocrine-disrupting chemicals in oil and natural gas,<sup>32</sup> environmental risks of MCCPs,<sup>23</sup> and toxicology and epidemiology of bentonite.<sup>33</sup> The sole review that tackles UVCBs as a substance class only addresses aspects of its risk assessment.<sup>29</sup> Meanwhile, reviews on chemical mixtures typically mention UVCBs only superficially<sup>34,35</sup> or do not explicitly address them at all.<sup>36,37</sup>

In this review, UVCBs are treated as a substance class as a means of addressing common challenges across UVCBs from the perspectives of cheminformatics, analytical chemistry, toxicology, and regulatory science. This review aims to (1) provide an overview of methodological developments for addressing UVCBs across the different domains, (2) summarize general approaches taken, (3) highlight challenges and gaps, and (4) identify further areas of research toward developing shared good practices. UVCBs warrant urgent attention from both scientific and regulatory communities, and this review aims to provide tractability in tackling this next frontier of environmental unknowns.

## 2. CHARACTERIZATION, IDENTIFICATION, AND REPRESENTATION OF UVCBS

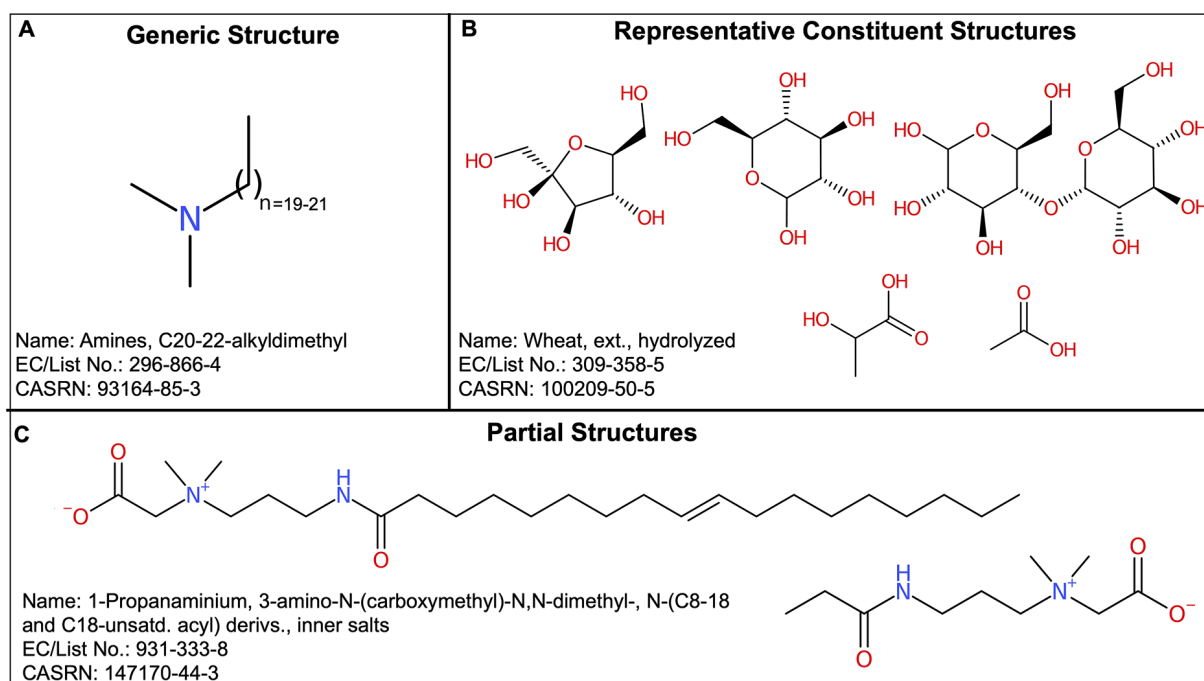
Meaningful structural representation of a chemical is important for connecting its detection in the environment or biota to chemicals registered on the global market and subsequent assessment of hazard, fate, exposure, and risks to human health and the environment. While chemical characterization (the process of obtaining information about a substance's constitu-

ents and composition), identification (unambiguous and precise recognition of the same substance by all stakeholders), and representation (how a chemical's identity is communicated) are typically clear for single compounds, they are not clear for UVCBs due to the lack of structural information available on these multiconstituent substances. Consequently, there exist challenges in chemically representing UVCBs using currently established formats: as text via its name, synonym, or description; structurally as structural diagrams, Simplified Molecular Input Line Entry System (SMILES),<sup>38</sup> molecular data files such as Molfile (MOL) and Structure Data File (SDF);<sup>39</sup> or by identifiers such as the International Chemical Identifier (InChI),<sup>40</sup> its hashed version InChIKey, and other database or registry specific identifiers, e.g., CASRN, Distributed Structure-Searchable Toxicity Substance Identifier (DTXSID), PubChem Compound Identifier (CID), and European Community List Number (EC/List No.).

**2.1. Current State of Available Structural Information on UVCBs in the Public Domain.** The current availability of UVCB structural information has largely been determined by registration requirements. A substance is categorized as UVCB during chemical registration if it adheres to UVCB specifications, as was historically the case in the United States, where nearly 10 000 UVCBs were listed in the original Toxic Substances Control Act Inventory dating back to 1979.<sup>41,42</sup> Similarly in Canada and Europe, substances are determined to be UVCBs if they meet the formal definition specified in the 1999 Canadian Environmental Protection (CEPA) Act<sup>18</sup> and 2017 Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) Guidance, respectively.<sup>43</sup>

In most cases, the initial information that can be used to identify UVCBs depends upon what registrants provide via the registration systems. For example, under EU REACH legislation, registrants can report multiple constituents, concentrations, and manufacturing process details of their UVCB within the International Uniform Chemical Information Database (IUCLID).<sup>44</sup> However, not all information submitted during registration is necessarily made publicly available at a level that allows for unambiguous identification of a given UVCB.<sup>45</sup> Furthermore, registration frameworks in most parts of the world tend to focus on new substances, despite the existence of many older substances with little to no available information that were already on the market before registration frameworks entered into force.

Presently, UVCBs are included in both national chemical registries and certain public databases. The major relevant databases, types of information available, and chemical representations are summarized in Table S2. Substance name is the most widely available identifier of UVCBs across all databases, and some substances have registry numbers (CASRN and/or EC No.) and/or an additional database identifier. Notably, however, substance name and identifiers for UVCBs can be ambiguous in nature.<sup>18,43,46,47</sup> Complete structural diagrams are frequently optional to provide upon registration; instead, descriptive information on chemical composition, source, processing, and/or partial structural diagrams are usually accepted.<sup>18,43,46,48</sup> Consequently, the vast majority of UVCBs have little to no detailed structural information (at least in the public domain), whether in the form of SMILES, InChI, structural diagram, or molecular formula. This lack of structural information is a fundamental knowledge gap concerning UVCB identities. For the few UVCBs that do have some associated



**Figure 1.** Examples of chemical structure representations for UVCBs available in REACH registration dossiers, depicted using CDK Depict.<sup>49</sup>

structural information, their chemical representation can be single and/or multiple structure(s) as illustrated in Figure 1.

Generic structures (Figure 1A) typically encompass a range of homologues with varying chain length at a certain site or sites on the molecule. Representative constituents for UVCBs (Figure 1B) can be chosen in multiple ways, e.g., as the predominant constituent by percent composition reported in the literature, to reflect a specific end point such as toxicity, of median chain length to represent homologous constituents of varying chain lengths, or two compounds with the shortest and longest chain lengths defining the range of constituents. Representatives resulting from grouping (sections 2.2.2 and 3) or statistical selection<sup>21,50</sup> are also possible. Lastly, partial structures (Figure 1C) represent one or more chemically interpretable aspects described in the substance name. Regardless of representation type, varying levels of specificity in structures (i.e., specific compound versus chemical class the compound belongs to) have been reported, resulting from being registered under the same registry number<sup>51</sup> or cheminformatic import issues across various databases causing inadvertent removal of undefined substituents (“Rgroup”) or imprecise polymer (“Sgroup”) definitions.<sup>39</sup>

**2.2. UVCB Characterization.** UVCB characterization has been driven by increased regulatory assessments of UVCBs,<sup>29</sup> developments in chemical database infrastructure,<sup>52</sup> and increasing awareness of the need to identify problematic chemicals in the environment.<sup>53</sup> Characterization initiatives have emerged in two main areas: cheminformatics (section 2.2.1) and analytical chemistry (section 2.2.2).

**2.2.1. Cheminformatics Approaches to Characterize UVCBs. Linking Preexisting Chemicals to UVCBs.** This cheminformatics approach involves linking preexisting structures of discrete compounds to UVCBs within chemical databases. A prominent example is the CompTox Chemicals Dashboard of the United States Environmental Protection Agency (U.S. EPA),<sup>54</sup> where constituents are linked to UVCBs via manually curated relationship mappings in its database. The

Dashboard also includes generic (Markush) representations and so-called “Markush Children” for UVCBs with generic structures.<sup>52</sup> Besides enumeration using Markush technology,<sup>55</sup> molecular structure generation methods such as MOLGEN<sup>56,57</sup> and simple SMILES expansion<sup>58</sup> have also been explored.<sup>59</sup> Another example is SciFinder’s<sup>60</sup> approach: SciFinder parses a UVCB name into its individual constituents and then provides the constituent structures as output to the UVCB queried. The drawbacks of this method are that the constituents must be present in the database to begin with (or new entries need to be registered), linking is time-consuming if performed manually or more prone to errors through automatic name parsing, and final structures are not necessarily achieved. Finally, the European Chemicals Agency Database (ECHA) has a section on “Group Members” within certain Substance Infocards, which may consist of UVCB constituents (e.g., MCCP<sup>61</sup>), and is curated either by official sources, expert judgment, or algorithm proposed judgment. However, this grouping is intended for specific regulatory activities instead of purely linking constituents to UVCBs. Therefore, groups may also contain substances that are not constituents if these substances fall within the same regulatory group.

**Elucidation of Chemical Structures.** For certain UVCB names containing chemically interpretable parts, e.g., “Quaternary ammonium compounds, coco alkyl(2,3-dihydroxypropyl)-dimethyl, 3-phosphates (esters), chlorides, sodium salts” (CASRN 173010-79-2), a trained analyst can manually elucidate (sub)structures using basic knowledge of chemical nomenclature, database searches, and depiction tools such as CDK Depict.<sup>49</sup> Representative structures are chosen where necessary, and proposed structures should be chemically feasible (e.g., obey basic chemistry principles such as valence rules). In this way, the analyst effectively manually generates new structural information. However, such structure elucidation can only be validated with analytical studies<sup>62</sup> and would not be applicable to UVCBs with names containing chemically uninterpretable elements such as unknown or variable starting



materials, biological species, or reaction processes, e.g., “Juniper, *Juniperus mexicana*, ext., isomerized, acetylated” (CASRN 91053-33-7) or “Distillates, petroleum, steam-cracked” (CASRN 64742-91-2).

An alternative approach involves extensively searching the literature for constituent structures and their “structural variability characteristics” (e.g., physicochemical properties inferred from spectral or chromatographic data), encoding these pieces of information into formats such as generic SMILES (section 2.3), and then generating all possible constituent structures accordingly.<sup>21,50,51</sup> This approach relies heavily on the availability of constituent information in the literature or from industry collaborators as well as curators’ knowledge and expert judgment to use this information, which may explain why its applicability has been limited to mostly petroleum substances so far, as expertise and information on their constituents are highly available compared to other substances.

**2.2.2. Analytical Chemistry Approaches to Characterize UVCBs. Elucidating Chemical Structures and Composition.** General discussions of analytical techniques applicable to characterizing UVCBs are available elsewhere,<sup>63,64</sup> but since these techniques are typically chemical class and property dependent, they must be tailored to specific UVCBs. Additionally, certain UVCBs such as petroleum substances that contain mostly hydrocarbons may be less challenging to characterize compared to UVCBs containing multiple chemical classes such as essential oils. Overall, petroleum substances appear to be the most extensively characterized UVCBs: constituent identification commonly by gas chromatography–mass spectrometry (GC–MS) and ion mobility spectrometry–mass spectrometry, and relative quantification by GC(xGC) flame ionization detection.<sup>65–70</sup> Essential-oil UVCBs were characterized using low resolution GC–MS aided by available library spectra and reference standards of constituents.<sup>71–73</sup> Among high resolution mass spectrometry methods, one example used five different techniques to characterize a polyhalogenated flame retardant UVCB, concluding that it is “dominated by C<sub>18</sub> carbon chain lengths, substituted with 3–7 chlorine atoms and 1–3 bromine atoms on an alkane chain”.<sup>62</sup> Unambiguous structural identification is often not feasible for many UVCBs such as these, as “no individual or mixed standards for [polyhalogenated (bromo-chloro) *n*-alkanes] exist”.<sup>62</sup> A similarly broad characterization of chlorinated paraffins revealed the composition of the constituents’ different chain lengths.<sup>74</sup> Constituent percentage compositions were also derived for organic metal salt UVCBs that required pretreatment steps for amenability to GC–MS and nuclear magnetic resonance analyses.<sup>75</sup>

In general, analytical characterization of UVCBs is technically challenging: first, the commercial availability of standards is limited. Petroleum UVCBs are the exception, as direct provision of standards by industry stakeholders supporting research likely contributed to intense characterization efforts over the years. Second, choosing appropriate test material may be difficult because of possible variability in substance composition. In a dossier screening study of 155 UVCB registration dossiers under REACH, 49% on average were found to have materials used for ecotoxicological end point testing that did not match the UVCB actually being registered.<sup>76</sup> Biological materials in particular can have high variability. For example, chondroitin sulfate (CASRN 9007-28-7) is a polymeric UVCB isolated from animals, whose diet and lifestyle, in addition to material extraction and processing, may affect polymer composition.<sup>30</sup> Likewise, a given petroleum substance produced using the same refinery

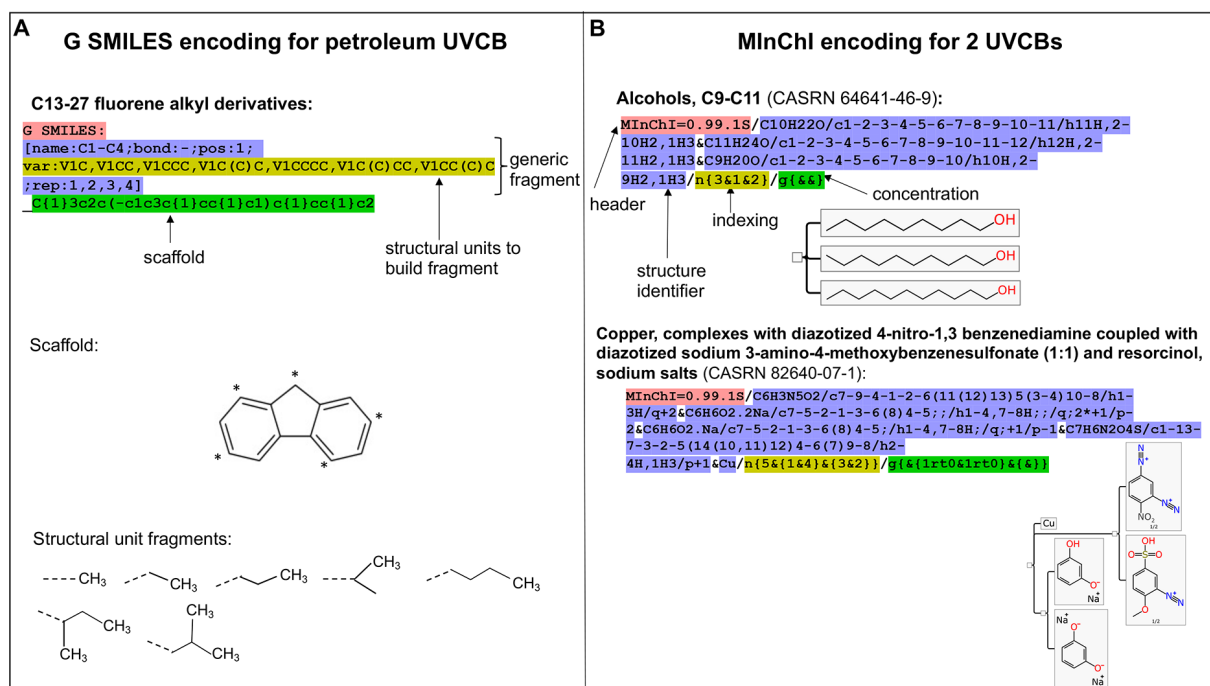
process could have different compositions within or across refineries depending on the operating conditions of the processing plant and chemical composition of the crude oil feedstocks.<sup>77</sup> Harmonized criteria with composition ranges<sup>76</sup> for selecting UVCB reference materials could be developed, and reference material manufacturers should provide detailed characterizations of their substances that have ideally been standardized, pooled, or homogenized across multiple batches.

Selecting appropriate sample preparation, separation, and analytical methods can be especially challenging for UVCBs, as there is little, if any, prior knowledge of substance identity to guide decisions in analytical strategy. Similar to typical nontarget studies, multiple analytical techniques and an iterative approach are often needed to provide as much complementary information as possible when dealing with UVCBs.<sup>30,62</sup> Ideally, both qualitative (constituent identity or bulk identities) and quantitative (constituent percent composition/concentration) characterization would be performed, highlighting the importance of both high mass resolution and chromatography (multidimensional if necessary for highly complex substances) in UVCB characterization. Where complete characterization is not possible, sum parameters (e.g., total carbon content, extractable organic chlorine, or total molar concentrations) can be used as intermediate descriptions.<sup>78</sup>

Overall, more studies and experience are needed for the analytical characterization of UVCBs, as they are so chemically diverse that there is no one method suitable for all. To date, most efforts have focused on some UVCBs of economic interest, i.e., petroleum products, and therefore other UVCBs may warrant more attention from the analytical chemistry community. A scheme prioritizing UVCBs by, e.g., known toxicity, high exposure, high production volume, or least complexity in terms of number/type of constituents may guide researchers in this area, as could the tiered approach for substance identification and characterization necessary to support ecological risk assessment that is currently under development.<sup>29</sup>

**Grouping.** Besides revealing compositions and information on chemical identities of individual UVCBs, analytical characterization of UVCBs enables grouping of substances and/or constituents based on common analytical features measured. Grouping helps mitigate substance complexity and multiplicity<sup>79</sup> through simplifying a UVCB down to representative constituents or fractions, or a group of UVCBs to a representative UVCB, thus allowing for more efficient testing, hazard assessment, and risk assessment (section 3), and read across (i.e., using available data to predict properties of analogous substances and fill data gaps),<sup>20,69,80</sup> while facilitating structural representation in databases (Figure 1B). In general, grouping should be fit for purpose as there are many strategies for and applications of grouping,<sup>81</sup> such that rationale, decisions, and uncertainties should be communicated transparently.

Establishing similarity is a prerequisite for grouping. Guidance specific to oleochemicals,<sup>82</sup> hydrocarbon solvents,<sup>83</sup> and petroleum substances<sup>84</sup> and for general chemicals<sup>80,85</sup> recommends grouping based on similar structural/physicochemical properties such as the presence of common functional groups, length and branching of carbon chains, aromaticity, etc. Ion mobility and GC–MS were used to group petroleum substances on this basis, as indicated by measured features in common such as carbon chain length, double bond equivalents, and H:C ratio.<sup>68,70</sup>



**Figure 2.** Examples of cheminformatics representations of UVCBs: (A) G SMILES. (Modified with permission from ref 51. Copyright 2015 John Wiley and Sons.) (B) Mixture InChI (MInChI).<sup>91</sup> The highlighted character strings are machine-readable formats, color coded according to the different components of G SMILES and MInChI, respectively, as indicated by their labels.

**Addressing Substance Variability.** Analytical characterization may reveal the extent of substance variability across different samples of the same UVCB, which may affect the applicability of available data on end points, properties, and/or substance identity. For example, despite observing some variation in hydrocarbon content and composition of the solvent “White Spirit” over multiple years and geographical samples, researchers concluded that the fluctuation was so minimal that its “technical properties and toxicological effects have not substantially changed”.<sup>86</sup> Conversely, insufficient similarity found among various *Ginkgo biloba* extracts may limit the applicability of toxicological data collected for the tested reference to untested samples.<sup>87</sup>

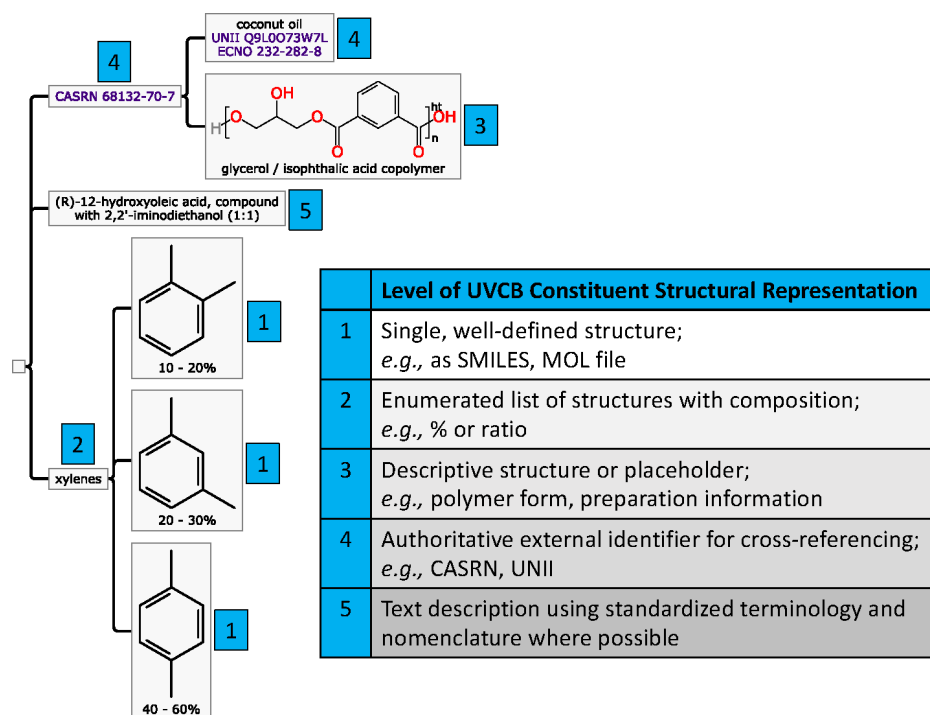
**2.3. UVCB Identification and Representation.** An appropriate representation for UVCBs is needed to facilitate unambiguous and precise identification, which in turn enables searchability. Currently, substance name is the most universally available representation across all UVCBs. However, name is problematic for searching as multiple synonyms may exist, names are sensitive to typographical errors, and they are often inconsistent across different registries/databases because there are multiple, inherently ambiguous UVCB nomenclature specifications across different jurisdictions.<sup>18,43,46,47</sup> Strategies to exploit this ambiguity have been developed, e.g., using generic descriptors to mask specific chemical identities.<sup>88,89</sup> Certain UVCBs such as essential oils face specific challenges: a combination of commercial, botanical, and chemical names can be used,<sup>90</sup> such that the same substance can have multiple different names. Additionally, curation inaccuracies and/or quality control issues can make identification even more difficult; e.g., within the ECHA database some substances have names such as “As UVCB, this information cannot be provided” (EC No. 942-495-4), or “the substance is UVCB” (EC No. 939-895-6). After name, CASRN is the second most used representation of UVCBs, but like substance name it is

imprecise and ambiguous<sup>29,47</sup> and is not an open identifier. Further compounding ambiguity issues, the same combination of CASRN and substance name can be used to represent different substances.<sup>47</sup>

For improved UVCB identification and searchability, there are currently two (complementary) alternative cheminformatics representations capable of capturing the multiconstituent, multifaceted nature of UVCB chemical systems in a machine-readable way (Figure 2). The first is generic SMILES (G SMILES), a method for structurally describing UVCBs and their variable compositions to facilitate hazard assessment via selection of representative constituents.<sup>21,51</sup> G SMILES relies on a dictionary of predefined descriptors to convey generic fragment information, derived from a scaffold-fragment approach. Nonstructural descriptors such as physicochemical properties and substance formation processes are encoded in a so-called G graph. However, since this format deliberately focuses on hazard assessment, it focuses on capturing relevant structures and disregards those considered irrelevant or computationally too expensive to manage. Additionally, it may not be easily applicable to substances whose names inherently contain little chemical information and thus no structural representation as it relies on the premise of an existent molecular scaffold. The format, proposed in 2015, has yet to be formally adopted in major databases.

The second approach applies the open InChI identifier to the latest developments in mixture cheminformatics, first proposed in 2019.<sup>92</sup> (Note: “mixture” is used here in the cheminformatics context of having multiple components, unrelated to the regulatory definition of mixture.) Mixture InChI (MInChI) provides a standardized definition of a given mixture that incorporates three essential properties within its notation: compound, quantity, and hierarchy. Incorporation of the InChI standard facilitates searching and linking of constituent information to public databases (e.g., PubChem). As for G





**Figure 3.** (left) Graphical illustration of the proposed UVCB data structure expressing constituents, concentrations/composition, and hierarchy, shown representing a “mixture of ‘coconut oil, polymer with glycerol and isophthalic acid’ (CASRN 68132-70-7) and ‘(R)-12-hydroxyoleic acid, compound with 2,2'-iminodiethanol (1:1)’ (CASRN 94232-00-5) dissolved in xylenes (CASRN 1330-20-7)” for demonstrative purposes.<sup>93</sup> (bottom right) Different specificity levels of available information on UVCB constituent structural representation, in decreasing order of preference from 1 to 5.

SMILES, knowledge of structure is necessary to generate InChIs, which may have limited application for many UVCBs. MInChI is in active development and has an open source editor and tools to generate an upstream “Mixfile” format for additional metadata.<sup>93</sup> A preliminary study has been initiated;<sup>91</sup> discussions within the International Union of Pure and Applied Chemistry (IUPAC)’s MInChI project are ongoing.

**UVCB Information Management.** Improved systematic representation of UVCBs as multicomponent substances is much needed to properly manage their multifaceted information properties toward supporting chemicals assessment and monitoring. In particular, the ability to link single components and their reported characteristics back to “source” substances would support the identification and tracking of UVCBs in environmental samples—an issue that has received little concerted attention so far. Ultimately, the goal for representation of UVCBs in databases is to make them as accurate, nonambiguous, and machine readable as possible, so that entries can be easily searched, classified, and analyzed—including by constituents and between databases. Proper quality control during registration, substance representation, and database curation will be crucial to avoid “inaccurate and unrepresentative structures in databases” (as discovered for CASRN 68527-01-5).<sup>62</sup>

Many UVCBs are intentional mixtures of poorly defined substances (e.g., plant extracts) with well-defined and characterized adjuncts (e.g., solvents). Breaking these up into separate components hierarchically allows known properties such as toxicity to be ascribed to either individual constituents, a group thereof, or an entire substance, which would eliminate ambiguity between individual and aggregate properties and facilitate analysis at the appropriate hierarchy level.

The data structure similar to the Mixfile format described by Clark et al.<sup>92</sup> could be used to achieve such systematic cheminformatic representation. Based on the principles of MInChI, the framework provided by Mixfile can be adapted to represent UVCBs at the *substance* level in terms of constituent, composition/concentration, and hierarchy. Additional metadata can be managed around these properties that facilitates cheminformatics operations and is able to handle missing or incomplete information about a given constituent. Importantly, whatever relevant chemical information available contributing to substance characterization (e.g., physicochemical properties, substance source, physical state/form, and toxicity) should be represented in a way that supports derivation of further properties via, e.g., modeling. Furthermore, especially for reaction product UVCBs, parameters such as reaction precursors, intermediates, reaction processes, and conditions of formation can be incorporated into substance characterization profiles.

For any given constituent in a mixture hierarchy, the specificity of constituent structural information available can be roughly characterized into five levels that indicate what types of cheminformatics functions can be applied (Figure 3).

Ideally, sufficiently characterized UVCBs have enough associated structural information to achieve level 1 and/or level 2 for individual constituents. With a single, well-defined structure (level 1), almost all structure-related derived properties can be calculated: names and identifiers via algorithms; database identity via lookup; and numerous search types, e.g., structure equivalence, similarity, substructure. Most importantly for chemical assessment, prediction of physicochemical, degradation, and (eco)toxicological properties via quantitative structure–activity relationships becomes possible. The same is generally true for level 2, but it is only viable up to the point

when enumerating all isomers/congeners/homologues is practical.

Level 3 captures the essence of the UVCB problem: there is something known *about* the chemical entities present but this information often cannot be readily converted into a manageable set of discrete constituents. For poorly defined constituents, chemical information is often reported in a form accessible to the experimentalist to a certain extent,<sup>47</sup> such as classes of chemical functionality (e.g., a form of starch is known to contain carbohydrate substructures), an industrial mixture described as the reaction products of certain input structures, polymers that may be indicated by providing the repeating units, and a constituent that may be described as all of the molecules from a source which distilled within a certain temperature range. The information known to the creator of the UVCB entry is sometimes only sufficient to enumerate a representative selection of molecules, but even when it is not, there might still be possibilities to narrow down what the molecules could be (e.g., by considering typical outputs from a given reaction type) and, subsequently, the appropriate queries and comparisons.

UVCBs may contain constituents that are defined in some sense other than chemical characteristics, which is commonly the case when using biologically sourced materials, corresponding to level 4. Many materials have an officially defined provenance and can be linked to a formal description using an identifier maintained or used by an authoritative organization, e.g., CASRN<sup>94</sup> or International Nomenclature Cosmetic Ingredient names.<sup>95</sup> These identifiers may be traced to the primary literature or preparation description (e.g., how to extract a fraction from a plant grown under certain conditions), but often they do not always provide meaningful, unambiguous chemical information, as discussed elsewhere.<sup>47</sup>

The final fallback, level 5, is to provide a text description of the substance, which facilitates keyword searching but is likely only understandable by domain experts. Very few higher-order text analyses are possible with current methods. However, such text-based fields could be supported by the development of ontologies or standardized terms (e.g., “acetylated”, “sulfurized”, or examples from European Union guidance<sup>43</sup>) that have formal definitions and should be used consistently by all stakeholders.

The above scheme is intended to be applicable to all UVCBs as a means of systematizing whatever information is currently available albeit possibly incomplete, for quality control of future reporting and to guide future characterization initiatives. Overall, but especially for chemicals assessment, levels 1 and 2 represent the most desirable levels of detail and should ideally be reflected in corresponding substance registration and characterization efforts.

### 3. HAZARD ASSESSMENT OF UVCBS

Different regulatory approaches exist around the world concerning the hazard assessment of UVCBs, some of which were reviewed elsewhere.<sup>63</sup> In the United States, the EPA has not issued any guidelines specifically addressing UVCB testing and instead relies on a case-by-case approach.<sup>29,63</sup> In Canada, UVCBs were prioritized<sup>96</sup> within the ecological risk classification approach under the Chemicals Management Plan<sup>97</sup> and assessed case by case using a weight of evidence approach (section 3.3.1), typically within chemical class specific groups, e.g., quaternary ammonium compounds, resins and rosins, etc. The groupings were identified on the basis of structural or functional similarities and were chosen according to several factors related to assessment efficiencies and avoiding regret-

table substitution, among others. Alternative grouping strategies by common fate properties and ecotoxicological effects have also been recommended<sup>63,84</sup> and performed based on common biological activity signatures,<sup>69</sup> toxicological and biodegradability end points,<sup>50</sup> and industrial use/emission patterns (section 4.1). Under the European Union’s REACH framework, certain hazard information must be provided with all registered UVCBs depending on the registration tonnage band and uses.<sup>98</sup> Multiple UVCBs have been assessed under the Australian Inventory Multitiered Assessment and Prioritisation Tier 1 framework,<sup>99</sup> but there is no specific UVCB guidance. Overall, UVCBs present challenges to regulatory frameworks concerning hazard assessment and communication, with specific issues related to testing strategies.

**3.1. Overarching Hazard Classification and Communication: GHS.** A primary outcome of hazard assessment is hazard classification, e.g., following the conventions of the Globally Harmonized System of Classification and Labeling (GHS). There is still no specific official guidance on UVCBs in the latest (ninth) revision of the GHS,<sup>100</sup> despite early initiatives to develop GHS guidance for petroleum UVCBs,<sup>101</sup> though a whole-mixture toxicity assessment is recommended for hazard classification of environmental and human health hazards and skin corrosion/irritation as well as for whole-mixture environmental biodegradation.<sup>100</sup> If only part of the mixture is known, a suite of bridging principles can be applied to predict the mixture classification. However, explicit guidance for mixtures exists. Applying GHS guidance for mixtures requires knowledge of all constituents present so that all the respective hazards can be evaluated, which may be possible for certain UVCBs. For example, the MeClas tool, used for hazard identification and classification, assumes all metal constituents are known in complex inorganic UVCBs.<sup>102</sup> Similarly, an adapted implementation of GHS was proposed for petroleum UVCBs,<sup>103</sup> where petroleum streams are considered unique substances each having individual CASRNs, which can be sorted into categories based on similar physicochemical/toxicological profiles and then evaluated for hazard accordingly. Implementing this same method for hydrocarbon solvents has been deemed feasible by Mckee et al.<sup>104</sup> However, for most other types of UVCBs, detailed knowledge of constituents may not be available, thus limiting the applicability of current GHS mixtures guidance to UVCBs because GHS requires all constituents to be known.

Despite the lack of UVCB-specific GHS guidance, testing strategies for hazard classification of UVCBs are under development.<sup>105</sup> Moreover, there is some evidence of partial GHS classification of certain UVCBs such as “Juniper, *Juniperus virginiana*” (CASRN 85085-41-2),<sup>106</sup> however, it is not clear how such classification was achieved, further supporting the need for specific transparent guidance for classifying UVCBs under GHS. In future guidance, some element to encode uncertainty could be introduced, e.g., as pictograms/classification/hazard statements to reflect uncertainty or incomplete understanding of the given UVCB composition and thus hazards.

**3.2. General Approaches to Assess Persistence, Bioaccumulation, and Toxicity (PBT).** Three main approaches have been prescribed for empirical testing of P, B, and/or T properties of UVCBs:<sup>63,107</sup> whole substance, known constituents, and fraction profiling (Figure 4). The European Union’s REACH encourages a combination thereof where necessary, for example, when knowledge of the substance evolves during assessment or if tested constituents are

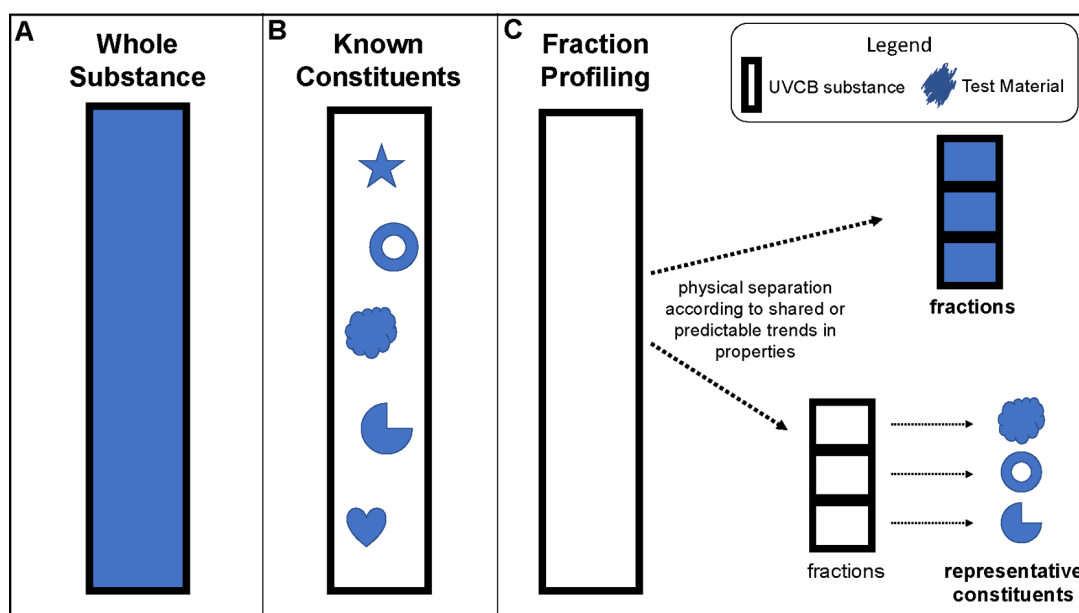


Figure 4. Schematic representation of the three main experimental approaches prescribed for PBT assessment of UVCB substances.<sup>63,107</sup>

sufficiently different from the remaining composition of the substance.<sup>107</sup> In the whole-substance approach, the entire UVCB undergoes testing and assessment (Figure 4A). However, because of substance complexity and potentially variable constituent solubilities that can cause challenging test conditions for the whole-substance approach, the known-constituent approach may be favored (Figure 4B). Known constituents can represent the entire UVCB in testing and assessment if they can be isolated, are present at relevant concentrations within the substance, and represent worst-case characteristics. Alternatively, the fraction-profiling approach involves splitting the whole substance into so-called “fractions”, and either the fractions themselves or representative constituent(s) of each fraction are tested (Figure 4C). The latter is also known as the “block method”. Physical separation of the whole substance into fractions is performed such that constituents within each fraction show a predictable trend in properties, e.g., physicochemical, structural, mode of action (MoA), and degradation.<sup>63,107</sup> Read across is expected to be applicable within the constituents of a given fraction.<sup>107</sup> The hydrocarbon block method (HBM)<sup>84</sup> is a specific form of fraction profiling for petroleum UVCBs and, together with its associated assessment tools (e.g., PetroTox,<sup>108</sup> a spreadsheet model designed to calculate the toxicity of petroleum products to aquatic organisms), has been the result of 30 years of work in the petroleum sector. In the first EU Technical Guidance Document, HBM was prescribed for assessing environmental risks of petroleum substances.<sup>109</sup>

Detailed discussions of the advantages and disadvantages of each approach are available elsewhere.<sup>63,107</sup> Briefly, testing whole substances does not require generation of new test material, but results may not be representative of all constituents; known constituents are relatively easy to test as they are discrete and well-characterized but may require more effort to characterize up front and may not ultimately be representative of the whole substance; and fraction profiling allows more targeted assessment than whole substance but requires generation of test material, i.e., the fractions.

A fourth, less common approach consists of *in silico* PBT screening, as recently performed for 884 constituents in the same hydrocarbon block of alkylated three-ring PAHs via relative trend analysis of experimental and modeled data.<sup>110</sup> The half-lives of petroleum products modeled by BioHCWin were validated by newly generated empirical data, suggesting that preliminary persistence screening of petroleum UVCBs is feasible using models.<sup>111</sup> Although *in silico* PBT screening may circumvent experimental difficulties associated with dealing with complex UVCBs, it ultimately requires experimental validation, is extremely data-intensive, and thus is only viable for well-studied UVCBs whose constituents are well-characterized and chemically similar.

The availability of PBT-related studies for a given UVCB is highly dependent on the nature of the substance itself and factors such as the substance’s practical applications, economic/industrial importance, availability of reference material, and overall environmental relevance. For example, there has been relatively more research on the degradation, bioaccumulation, and toxicity behaviors of petroleum substances and chlorinated paraffins,<sup>23</sup> as reflected in extant prioritization schemes for PBT assessment, likely because these are well-known UVCBs.<sup>112</sup> In comparison, there is little knowledge of the PBT characteristics of lesser-known UVCBs such as “Morpholine, 4-C<sub>12–14</sub>-alkyl derivs.” (CASRN 1402434-48-3), “Alcohols, lanolin” (CASRN 8027-33-6), or “Fatty acids C<sub>18</sub> unsat, reaction products with pentaethylenhexamine” (CASRN 1224966-13-5).

The following sections focus on recent studies highlighting universal issues affecting PBT testing strategies for UVCBs.

**3.2.1. Persistence.** Generally, ISO- and OECD-standardized tests for degradability were originally developed for fully characterized substances and by default adopt a whole-substance approach. The biodegradation screening tests, e.g., ready biodegradability (OECD 301A–301F) and inherent degradability (OECD 302A–302F), typically measure CO<sub>2</sub> formation, theoretical oxygen demand, or substrate decay. These methods can be applied to UVCBs, although these screening tests may not accurately reflect whole-substance persistence. The simulation biodegradation tests in soil, sediment, and surface



water (OECD 304, 307, 308, 309) require  $^{14}\text{C}$  labeled compounds to quantify loss of the parent and identify transformation products. While these are more challenging to perform for UVCBs, efforts involving, e.g., fully labeled chlorinated paraffin mixtures already exist.<sup>113</sup>

Screening tests based on  $\text{CO}_2$  formation or oxygen demand quantification can be applied to UVCBs, but it is possible that the persistence of a whole UVCB could be incorrectly determined by assessing its more degradable constituents, despite the UVCB containing persistent constituents. As these tests do not provide detailed persistence information at the constituent level, the true degradability of a UVCB can be subject to interpretation and may have to be evaluated on a case-to-case basis.<sup>63</sup> An alternative measure for testing a UVCB's ready biodegradability has been proposed, where a carbon balance approach is used to derive the level of ultimately transformed organic carbon (sum of mineralized carbon and carbon converted to biomass) in aerobic biodegradation tests as a measure of ready biodegradability, but it may be limited to only substances whose carbon content can be measured.<sup>114</sup>

In certain cases where the UVCB has a relatively simple chemical composition, it may be justifiable to apply bulk degradation test results to the entire UVCB substance. For example gas-to-liquid synthetic hydrocarbons were deemed "sufficiently homologous", such that nonspecific results from ready biodegradability tests "can be used to conclude on their biodegradability as a whole".<sup>115</sup> Alternatively, if tested known constituents cover an appropriately broad and relevant chemical space that would account for substance variability, degradation results could be extrapolated to other substances within that applicability domain, as performed with kinetic studies of test chemicals commonly found in petroleum substances.<sup>116,117</sup>

Overall, evaluating UVCB persistence is still in the method development stage, as there are many technical and analytical challenges, e.g., possible impact of mixture effects (where certain constituents may enhance or diminish the biodegradation kinetics of other constituents present), for which whole-substance testing is necessary to evaluate.<sup>72,118,119</sup> An important outcome of these works for informing future studies is that test concentrations should be kept at low, environmentally relevant concentrations to avoid mixture toxicity affecting biodegradation kinetics. To date, most studies focused on developing persistence tests for hydrophobic UVCBs. Testing strategies for UVCBs with other types of challenging physicochemical properties (e.g., hydrophilic, volatile) should be developed to enable the persistence testing of UVCBs with different properties.

**3.2.2. Bioaccumulation.** Initial bioaccumulation screening relies on the octanol–water partition coefficient ( $K_{ow}$ ), but as with persistence testing, different constituents may have different  $K_{ow}$  values and thus different bioaccumulative properties that could complicate results interpretation for whole UVCBs. Initial estimates of whole-substance bioaccumulation potential could be inductively concluded if analytical methods such as high performance liquid chromatography capable of capturing multiple constituents indicate whether all constituents either exceed or are below the common regulatory  $\log K_{ow}$  4.5 threshold for screening bioaccumulation assessment.<sup>107</sup> However, as equilibrium partitioning may not be the only process determining bioaccumulation,  $\log K_{ow} > 4.5$  does not imply that a chemical is bioaccumulative, but further evaluations are required. In the case of UVCBs, different constituents may undergo active uptake, metabolism, and/or excretion to varying

extents.<sup>29</sup> The recommended approach<sup>63</sup> has been to consider the bioaccumulative properties of a UVCB's representative/main constituents instead of those of the whole substance itself. Bioconcentration factors (BCFs) were successfully determined for the main constituents of "cedarwood Virginia oil" (CASRN 8000-27-9) in rainbow trout this way,<sup>120</sup> and continued work by the same authors developed an analytical technique within a suspect-screening approach that circumvents the need to have *a priori* knowledge of constituent identities and available analytical standards.<sup>121</sup> Several technical substance mixtures of chlorinated paraffins, typically already subdivided according to chain length, were found to be bioaccumulative in *Daphnia magna*.<sup>122</sup>

An extended discussion of measuring UVCB bioaccumulation is available elsewhere.<sup>29</sup> Overall, there are very few bioaccumulation studies of UVCBs and their constituents, and more work is needed to develop methods for future bioaccumulation studies of other UVCBs, such as testing the suitability of *in vitro* methods.<sup>29</sup>

**3.2.3. Toxicity.** Toxicity assessment requires aquatic toxicity testing and/or the evaluation if the substance poses a human health hazard, namely if it is carcinogenic, mutagenic, or reproduction toxic (CMR), an endocrine disrupting compound (EDC), or mediates specific target organ toxicity (STOT). Aquatic toxicity testing of UVCBs is challenging from two perspectives. First, it involves the ability to correctly define the dose of the substance and make sure a constant test concentration is maintained over the testing period. Second, the constituents of many UVCBs can be very hydrophobic, making dosing challenging even for single compounds. Toxicity is mediated by bioavailability, which is limited by solubility and the sample preparation methods used. Interestingly, very hydrophobic chemicals are of such low solubility that toxic concentrations cannot be achieved for single compounds but can be achieved for mixtures.<sup>123</sup> As UVCBs have multiple constituents of likely varying solubilities and percentage compositions, aquatic toxicity testing of UVCBs poses technical challenges for hydrophobic and/or volatile constituents. Thus, considerable studies in recent years have focused on developing improved toxicity testing methodologies for UVCBs, especially with respect to dosing of volatile, hydrophobic, and volatile and hydrophobic UVCBs,<sup>124–126</sup> as well as analyzing the effect of sample preparation on bioavailability.<sup>127</sup>

Overall, modeling toxicity and testing of UVCBs have mostly focused on petroleum substances,<sup>124–126,128</sup> solvents,<sup>86,129</sup> and chlorinated paraffins.<sup>23,130–132</sup> Future method development and toxicity evaluations of other UVCBs are warranted.

**3.3. Additional Considerations for Comprehensive Effect Assessment.** Exposure to a UVCB substance results in combined exposure to more than one chemical at the same time. Therefore, from a chemical and toxicological perspective, UVCBs are mixtures despite the legal distinction drawn between UVCBs and mixtures within regulatory frameworks.<sup>17,133,134</sup> Thus, for the purposes of comprehensive effect assessment, the same established principles for assessing mixture toxicity are applicable to assessing UVCBs.<sup>135</sup>

**3.3.1. Whole-Mixture Testing.** Comprehensive effect assessment requires a whole-substance approach where the effect of the mixture is tested. In principle, dosing mixtures into bioassays follows the sample principles as for single chemicals, and since solubility of each compound is additive in a mixture, overall, more chemicals can be brought into solution in the case of UVCBs as compared to single chemicals. However, there are challenges because the mixture composition must not be

changed since the exposure concentrations of mixtures cannot be confirmed analytically.

Dosing remains a particular challenge for UVCBs that contain many low solubility components because the solubility of whole mixtures depends on the solubility of the least soluble constituent during aquatic toxicity testing. Therefore, there is a danger that the more hydrophobic chemicals are not dissolved and hence not bioavailable, and the effect is dominated by the more soluble constituent. As more hydrophobic chemicals are typically more potent than more hydrophilic chemicals, this may lead to dramatic underestimation of toxicity.

Another complication is UVCBs with volatile components or volatile and hydrophobic components. For such UVCBs, the water accommodated fraction (WAF) approach is intended as a “last resort” if all other means of ensuring stable substance concentrations during testing have been exhausted,<sup>107</sup> or as an “additional supporting line of evidence” to empirical and modeled data.<sup>136</sup> It involves expressing aquatic toxicity in terms of loading rate (ratio of test substance to aqueous medium used to make the aquatic toxicity test medium), thereby providing a measure of relative toxicity at concentrations equating to the apparent solubility of each component and not their actual abundance in the mixture. However, WAF has fundamental drawbacks: it represents only a fraction and not the whole substance (whose chemical identity is subject to uncertainty), mixture composition may be altered compared to the UVCB it is prepared from, and the WAF composition depends on preparation techniques. Issues related to WAF results interpretation for coal tar pitch and kerosene/jet fuel UVCBs within regulatory processes of the U.S. EPA and REACH have been reported.<sup>137</sup> Alternatives to WAF include solvent extraction followed by solvent spiking, generator systems, saturator columns, and passive dosing methods, the last of which has been in active development in recent years with respect to UVCBs.<sup>124–126</sup>

On balance, results from whole-mixture testing could be integrated into a weight of evidence (WoE) approach for UVCB assessment. In Canada’s WoE approach, multiple lines of evidence are considered in the assessment of a UVCB: for example, besides considering WAF test results, other aspects such as representative structures, individual constituent toxicity, and additive toxicity may also be evaluated together when deciding on a substance’s toxicity and capacity to cause adverse effects in the environment.

**3.3.2. Mixture Toxicity Models: Toxic Equivalence Approach for UVCBs.** Ideally, choosing an appropriate mixture toxicity model for a given UVCB would be determined by knowledge of its constituents and composition. For example, UVCBs containing chemically diverse constituents with different MoAs would follow an independent action (IA) model of toxicity, while those with the same MoA would follow concentration addition (CA), whereas mixtures with known interactions between their constituents might cause synergistic or antagonistic effects. However, these effects are rare and typically happen in mixtures with few components and for highly specifically acting compounds such as in pesticide formulations;<sup>138,139</sup> therefore synergism and antagonism are unlikely for UVCBs.

The simple CA model can be applied to UVCBs with relatively simple compositions and chemically similar constituents (e.g., UVCBs such as “Alcohols, C<sub>9</sub>–C<sub>11</sub>”). Even independently acting compounds often have mixture predictions very similar to CA or converge to the same mathematical

model at low effect levels (<10%).<sup>140,141</sup> Very complex UVCBs with many diverse constituents, albeit each individually present at very low concentrations below effect levels (e.g., petroleum, or biological materials like essential oils), would also follow CA. Provided that relative effect potencies (REPs) are independent of effect level or concentration in these cases, a toxic equivalency approach can be applied.<sup>142</sup>

The toxic equivalent concentration (TEQ) of a UVCB or any chemical mixture is the sum of the products of the concentration of each constituent *i* and its respective toxic equivalency factor (TEF<sub>*i*</sub>), where TEF<sub>*i*</sub> is defined as the ratio of the effect of a reference compound to the effect of the constituent *i*. Such a reference compound could be a known representative constituent. TEFs are consensus values for dioxin-like chemicals,<sup>143</sup> but a conceptually and mathematically identical approach could be taken using REP<sub>*i*</sub>’s from the same toxicity test<sup>142</sup> (eq 1), where *C<sub>i</sub>* is the concentration of constituent *i* in the mixture, EC<sub>*i*</sub> is its effect concentration in the given bioassay, and EC<sub>ref</sub> is the effect concentration of the reference compound.

$$\text{TEQ} = \sum_{i=1}^n C_i \cdot \text{REP}_i = \sum_{i=1}^n C_i \cdot \frac{\text{EC}_{\text{ref}}}{\text{EC}_i} \quad (1)$$

The TEQ approach was mentioned in the official European Union opinion on mixtures<sup>144</sup> but no practical examples for UVCBs exist in the public domain as of yet. Currently, the whole-mixture approach is recommended in regulatory risk assessment of mixtures.<sup>35</sup> In practice, if not all the EC<sub>*i*</sub> values of the mixture components are known, they can be approximated by similar constituents, as was successfully demonstrated for the human health risk assessment of brominated flame retardant mixtures.<sup>145</sup>

In multiconstituent mixtures, not only does CA likely apply, but toxicity of complex mixtures is often reduced to baseline toxicity,<sup>146</sup> which is the minimum toxicity triggered by nonspecific interactions of chemicals with biological membranes leading to disturbance of structure and functioning of cell and organelle membranes.<sup>147</sup> Since all chemicals are equipotent with respect to baseline toxicity if effects are expressed as internal concentrations, there is a critical molar membrane burden above which effects can be observed. This level is around 200–500 mmol/kg lipid for LC<sub>50</sub> of aquatic animals.<sup>147–149</sup> The chemical properties of the chemicals decide only how much is taken up by the organism and ultimately distributed into the membranes, but once they are in the membrane all chemicals act close to equipotent. This means that critical membrane burdens or, for all practical matters, critical or lethal body burdens can easily be applied to mixtures.<sup>150</sup> This principle has also been extended to mixtures in the so-called target lipid model (TLM).<sup>151,152</sup>

## 4. EXPOSURE AND RISK ASSESSMENT OF UVCBS

**4.1. Exposure Assessment.** Within regulatory frameworks, exposure assessment of UVCBs is not always considered necessary and is highly dependent on the framework in question. For example, in the European Union, the outcomes of initial hazard assessments may already be enough to initiate risk management measures without having to assess UVCB exposure. However, in many other jurisdictions such as the United States, Canada, and Australia, a full risk assessment of chemical substances that includes exposure assessment is generally required to determine whether risk management measures should be triggered.

In cases where exposure assessment of UVCBs is necessary, regulators must deal with multiple challenging aspects of UVCB exposure, particularly with regard to environmental monitoring and biomonitoring. First, it is difficult to measure UVCBs in the environment because of their multiconstituent nature. Environmental monitoring typically only tracks single compounds, but because UVCBs comprise multiple constituents, validation issues may arise as it is difficult to attribute the detection of a particular constituent to the emission of a UVCB containing that constituent. Furthermore, environmental transformations of these constituents and potentially different fate and transport properties resulting in different exposure pathways could complicate this attribution further.<sup>19,29</sup> Therefore, ideally full knowledge of constituent identities and compositions is needed for exposure assessment of UVCBs. However, as this has been difficult to achieve in practice, refining exposure scenarios by, e.g., considering the magnitude of emissions and current mitigation measures in place may help prioritize substance characterization efforts (section 2) needed for exposure assessment. Overall, some uncertainty will remain regarding unknown constituents and their unknown environmental fate and exposure properties, which is challenging to capture in the overall exposure assessment. It is important to convey this gap in knowledge/uncertainty as part of assessment outcomes.

Concepts for exposure assessment and fate and transport modeling of UVCBs are currently under active development.<sup>29</sup> A review of publicly available electronic registration dossiers and risk assessment reports revealed three main approaches for estimating exposures of UVCBs: whole substance (section 4.1.1), constituent (section 4.1.2), and expert judgment (section 4.1.3).

**4.1.1. Whole-Substance Approach.** UVCBs whose constituents are not clearly defined or are too complex in composition can be assessed as a whole. Relevant information such as import and manufacturing volumes, consumer uses, product use scenarios, and percent concentration within products are considered. An example is the assessment of the organic anthraquinone UVCB “9,10-Anthracenedione, 1,4-diamino-, N,N'-mixed 2-ethylhexyl and Me and pentyl derivs.” (CASRN 74499-36-8) by the Government of Canada (GoC) using the ConsExpo model to estimate oral and dermal exposures.<sup>153</sup>

Whole-substance exposure assessment can also be performed for groups of substances within, e.g., a common sector of industrial activity, as their exposures are considered very similar or identical. GoC assessed 57 sector-specific inorganic UVCBs used in metals, paper, and cement processing and manufacturing in this way.<sup>154</sup> Exposure potential was evaluated on the basis of the status of the substance (e.g., “waste”, “byproduct”), and whether there were any preexisting measures to limit environmental exposure. In this example, exposure was emphasized over hazard in the overall characterization of risk, and as exposure was deemed negligible, regardless of hazard, risks to human health were considered low and harm to the environment not expected. However, such grouping and disproportionate emphasis on exposure over hazard could be detrimental for substances with specific MoAs and/or high toxicity, uncertainties in assessing exposure potential persist, and there may be caveats in assuming the preexisting measures to limit exposure were adequate.

**4.1.2. Constituent Approach.** Each constituent and/or representative constituents must be known and should undergo individual exposure assessment (or the relevant information gathered from the literature) before the assessments are combined to give an overall exposure assessment of the

UVCB. This approach has been recommended for inorganic UVCBs, where assessing constituents would be similar to “standard metal exposure assessment” and should take into account speciation behavior, assuming the worst-case scenario where information is incomplete.<sup>155</sup> In the final combination step of the parallel constituent assessments, multidimensional risk characterization ratio tables (constituent × exposure route × local/systemic effects, short term/long term) are generated.<sup>155</sup> Examples exist under the EU REACH, e.g., the inorganic UVCB “Lead alloy, base, Pb, Sn, dross” (CASRN 69011-60-5), whose dossier states “assessing transport and distribution of the UVCB substance has no meaning”, as the “metals contained in the UVCB have been assessed in the respective risk assessments”.<sup>156</sup>

**4.1.3. Expert Judgment.** Expert judgment can be used where there is insufficient knowledge of hazard and exposure and no representative structure(s) to describe the substance. Qualitative exposure classification was performed for 192 organic UVCBs<sup>157</sup> and an anthraquinone UVCB (CASRN 74499-36-8) by GoC.<sup>153</sup> Supporting information, e.g., industry surveys and consideration of similar substances, was also taken into account. However, more information is needed to transparently illustrate how these expert judgments are carried out and validated, and to assess whether such judgments can be automated in the future.

**4.2. Risk Assessment of UVCBs.** Risk characterization traditionally involves the calculation of a risk quotient: the outcome of exposure assessment (e.g., predicted environmental concentration, PEC) is divided by that of effect assessment (e.g., predicted no effect concentration, PNEC). Risk quotients of individual components of a mixture are additive to yield the risk index (RI) if CA applies for the mixture effect (eq 2).

$$RI = \sum_{i=1}^n RQ_i = \sum_{i=1}^n \frac{PEC_i}{PNEC_i} \quad (2)$$

Hence for mixtures and therefore also for UVCBs, one could calculate the TEQ as described above and use that in relation to the PNEC of the reference compounds used to derive the TEQ (eq 3).

$$RI = \frac{TEQ_i}{PNEC_{\text{reference chemical}}} \quad (3)$$

Comprehensive environmental risk assessments including both effect and exposure of whole substances have been developed for two particular types of UVCBs: petroleum products (PETRORISK)<sup>158</sup> and hydrocarbon solvents.<sup>159</sup> While substance complexity and variability are reflected in hazard and risk predictions by PETRORISK,<sup>160</sup> careful ongoing evaluation of these models is necessary, as PETRORISK was found to underestimate the environmental risks of petroleum use and production.<sup>161</sup> Methods for other UVCBs have yet to be established.

**4.3. Current Regulatory Activities, Perspectives, and Priorities.** Overall, many regulatory authorities have endeavored over the past decade to develop scientifically sound and consistent approaches for the assessment of UVCBs. However, the availability of specific (standardized) guidance to achieve this is still limited to date. In practice, both whole-substance<sup>136,153,154,162</sup> and constituent-based<sup>156</sup> approaches are being used in current regulatory assessments, informed by established principles such as those of HBM (but tailored to suit chemistries other than petroleum), as well as guidance on mixtures.<sup>163–166</sup> Given the large range in complexity, chemical



classes, and data availability for UVCBs, it is not always possible to be prescriptive for all aspects of hazard, exposure, and/or risk assessment. Therefore, a case-by-case approach is still the preferred and potentially only viable approach for certain UVCBs, but it may pose a burden for risk assessors and result in less predictability for stakeholders.

## 5. DISCUSSION: CHALLENGES AND OPPORTUNITIES

Several systemic factors contribute to the challenges posed by UVCBs: information gaps in chemical identities and compositions stemming from the registration process, inadequate chemical representation and nomenclature hindering identification and database searchability, lack of analytical standards and methods tailored specifically to UVCBs, challenging conditions for PBT testing, and the sheer number of UVCBs to be assessed. Below, key opportunities and steps forward in addressing these challenges are summarized.

**5.1. Registration.** Fundamental knowledge gaps in UVCB identities could be avoided from the start if information requirements to register UVCBs were increased, in tandem with implementing better methods for chemical representation. Requiring machine-readable structural information, including representative or generic structures for constituents, and compliance and quality checks during registration may assist with this. Standardized description terminology should be developed toward improving UVCB nomenclature for registration, possibly with the support of IUPAC and CAS. Potential avenues to implement these information types include GHS, OECD, and IUCLID.

**5.2. Chemical Representation and Information Management.** Chemical representation issues linked to nomenclature, structure, and use of closed identifiers such as CASRN still hinder precise identification of UVCBs. Machine-readable representations to enable unambiguous substance identification and searchability such as G SMILES and the open MInChI represent possible solutions. Future initiatives to improve chemical representation of UVCBs could be spearheaded by organizations such as IUPAC's InChI Subcommittee focusing on capturing mixture composition using MInChI.<sup>167</sup>

UVCB information must be better organized to enable (1) capture of their multiconstituent and multifaceted properties, (2) quality checks, and (3) detection of information gaps. A hierarchical data format and associated constituent representation scheme were proposed to achieve this (Figure 3). It is important for stakeholders to consider this format in further discussions toward achieving a standardized system so that future reporting, storing, and exchanging of UVCB information become more accurate and precise. Future research in this area such as proofs of concept and analyses on how our proposed format could function for several types of UVCBs is highly anticipated.

**5.3. UVCB Characterization: Toward Environmental Detection and Monitoring.** UVCB characterization is currently achieved by two means: cheminformatics and analytical chemistry. Cheminformatics methods rely on text parsing and cross-linking information that already exists in databases and, because these are often done *in silico*, are potentially the fastest and most scalable characterization approach. However, these methods are fundamentally limited by the availability and quality<sup>62</sup> of preexisting UVCB information in the public domain.

Ultimately, analytical characterization will be necessary to generate (new) knowledge on UVCB identities and composi-

tions. UVCBs other than petroleum substances warrant characterization, particularly if they are high production, toxic, or heavily emitted into the environment. Given their complex and unknown characteristics, nontarget strategies<sup>168,169</sup> involving multiple analytical techniques to give complementary information will be required to elucidate UVCBs, especially as they may have generic elemental compositions (e.g., only C, H, O, N) and molecular formulas similar to hundreds of natural products, making them hard to distinguish from environmental matrices. Chemometrics or cheminformatics tools could be used for prioritization based on substructure or toxicity.<sup>170,171</sup>

Overall, UVCB characterization is a prospective area of dynamic research, especially as knowledge of their identities becomes indispensable for answering "bigger questions" such as investigating known toxicity end points associated with constituents requiring identification. Successful characterization efforts and analytical method development contributing to better knowledge of UVCB identities will likely open more avenues for their environmental detection and monitoring. Chlorinated paraffins<sup>23,74,172</sup> are a good example, as their constituents are known and have distinctive analytical signatures (e.g., homology, multiple halogens present) which facilitate identification.<sup>173–175</sup> However, for UVCBs with very different constituents, new concepts and analytical methods for their environmental detection will be necessary. Several open questions remain, such as how many constituents must be co-detected to conclude on the detection of a specific UVCB, how potential transformations<sup>176</sup> and partitioning of different constituents across multiple environmental compartments can be accounted for, etc.

**5.4. Hazard, Exposure, and Risk Assessment.** Existing testing strategies for single-compound end-point assessments should be adapted to the multiconstituent characteristics of UVCBs following one of three approaches: whole substance, known constituent, and fraction profiling. Standardized testing methods are needed, requiring cooperation among the relevant stakeholders to develop them. Strategies such as grouping and read across may help streamline chemicals assessment, especially for UVCBs with similar constituents or properties, as would applying appropriate mixture toxicity models (i.e., CA and/or TEQ) for comprehensive effect assessment in a complementary approach to further substance characterization.

To support chemicals assessment of UVCBs, current priorities for future research and action include the following: (1) improving the quality and availability of information on UVCB components, (2) deepening the understanding of manufacturing and use practices and the release potential of UVCBs to the environment, (3) developing tools to estimate exposure of multiconstituent substances in environmental matrices and biota, (4) developing standard hazard and fate test and assessment methods for UVCBs, and (5) improving approaches to communicating complex risk assessment findings to stakeholders.

Concerted efforts from all stakeholders are needed to systematically address UVCBs, particularly in identifying and managing those that present unacceptable risks. There are tens of thousands of UVCBs on the market, and risk assessment prioritization schemes such as those available for petroleum substances<sup>112</sup> should be devised for other UVCBs based on, e.g., detection in the environment, highest production volumes, and known toxicity and/or exposures (preliminary initiatives within NORMAN Network activities are underway<sup>177</sup>). Meanwhile, stakeholders may also consider simplification<sup>79</sup> and sustainable

circular use<sup>178</sup> principles of UVCBs toward their sound management in the medium to long term.

## ■ ASSOCIATED CONTENT

### SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.est.2c00321>.

Details of individual UVCBs mentioned in this review (TXT)

Summary table of major databases and inventories containing UVCBs (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Authors

**Adelene Lai** – Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 4367 Belvaux, Luxembourg; Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller University, 07743 Jena, Germany; [orcid.org/0000-0002-2985-6473](https://orcid.org/0000-0002-2985-6473); Email: [adelene.lai@uni.lu](mailto:adelene.lai@uni.lu)

**Emma L. Schymanski** – Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 4367 Belvaux, Luxembourg; [orcid.org/0000-0001-6868-8145](https://orcid.org/0000-0001-6868-8145); Email: [emma.schymanski@uni.lu](mailto:emma.schymanski@uni.lu)

### Authors

**Alex M. Clark** – Collaborative Drug Discovery Inc., Burlingame, California 94010, United States; [orcid.org/0000-0002-3395-4666](https://orcid.org/0000-0002-3395-4666)

**Beate I. Escher** – Helmholtz Centre for Environmental Research GmbH—UFZ, 04318 Leipzig, Germany; Environmental Toxicology, Center for Applied Geosciences, Eberhard Karls University Tübingen, 72076 Tübingen, Germany; [orcid.org/0000-0002-5304-706X](https://orcid.org/0000-0002-5304-706X)

**Marc Fernandez** – Environment and Climate Change Canada, Vancouver, British Columbia V6C 3R2, Canada

**Leah R. McEwen** – Cornell University, Ithaca, New York 14850, United States; International Union of Pure and Applied Chemistry, Research Triangle Park, North Carolina 27709, United States; [orcid.org/0000-0003-2968-1674](https://orcid.org/0000-0003-2968-1674)

**Zhenyu Tian** – Department of Chemistry and Chemical Biology, Department of Marine and Environmental Sciences, Northeastern University, Boston, Massachusetts 02115, United States; [orcid.org/0000-0002-7491-7028](https://orcid.org/0000-0002-7491-7028)

**Zhanyun Wang** – Empa—Swiss Federal Laboratories for Materials Science and Technology, Technology and Society Laboratory, 9014 St. Gallen, Switzerland; Chair of Ecological Systems Design, Institute of Environmental Engineering, ETH Zürich, 8093 Zürich, Switzerland; [orcid.org/0000-0001-9914-7659](https://orcid.org/0000-0001-9914-7659)

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.est.2c00321>

### Notes

The authors declare the following competing financial interest(s): A.M.C. declares that Collaborative Drug Discovery is involved in developing commercial products to deal with mixtures. No other authors declare any competing interests.

## ■ ACKNOWLEDGMENTS

Randolph R. Singh and Corey M. Griffith are acknowledged for helpful discussions. We also thank the three anonymous

reviewers for their constructive feedback. A.L. and E.L.S. are supported by the Luxembourg National Research Fund (FNR) for Project A18/BM/12341006. A.M.C. and L.R.M. are supported by National Institutes of Health Grant 2R44TR002528-02. Z.W. gratefully acknowledges financial support by the European Union under the Horizon 2020 Research and Innovation Programme (Grant Agreement No. 101036756), and his work at ETH Zurich as part of the NCCR Catalysis (Grant No. 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

## ■ REFERENCES

- (1) Ssebugere, P.; Sillanpää, M.; Matovu, H.; Wang, Z.; Schramm, K.-W.; Omwoma, S.; Wanasolo, W.; Ngeno, E. C.; Odongo, S. Environmental Levels and Human Body Burdens of Per- and Polyfluoroalkyl Substances in Africa: A Critical Review. *Sci. Total Environ.* **2020**, *739*, 139913.
- (2) Hoang, A. Q.; Tran, T. M.; Tu, M. B.; Takahashi, S. Polybrominated Diphenyl Ethers in Indoor and Outdoor Dust from Southeast Asia: An Updated Review on Contamination Status, Human Exposure, and Future Perspectives. *Environ. Pollut.* **2021**, *272*, 116012.
- (3) Panagopoulos Abrahamsson, D.; Warner, N. A.; Jantunen, L.; Jahnke, A.; Wong, F.; MacLeod, M. Investigating the Presence and Persistence of Volatile Methylsiloxanes in Arctic Sediments. *Environ. Sci.: Processes Impacts* **2020**, *22* (4), 908–917.
- (4) Köck-Schulmeyer, M.; Ginebreda, A.; Petrovic, M.; Giulivo, M.; Aznar-Alemay, O.; Eljarrat, E.; Valle-Sistac, J.; Molins-Delgado, D.; Diaz-Cruz, M. S.; Monllor-Alcaraz, L. S.; Guillem-Arghiles, N.; Martínez, E.; Miren, L. de A.; Llorca, M.; Farré, M.; Peña, J. M.; Mandaric, L.; Pérez, S.; Majone, B.; Bellin, A.; Kalogianni, E.; Skoulikidis, N. Th; Milačić, R.; Barceló, D. Priority and Emerging Organic Microcontaminants in Three Mediterranean River Basins: Occurrence, Spatial Distribution, and Identification of River Basin Specific Pollutants. *Sci. Total Environ.* **2021**, *754*, 142344.
- (5) Abayi, J. J. M.; Gore, C. T.; Nagawa, C.; Bandowe, B. A. M.; Matovu, H.; Mubiru, E.; Ngeno, E. C.; Odongo, S.; Sillanpää, M.; Ssebugere, P. Polycyclic Aromatic Hydrocarbons in Sediments and Fish Species from the White Nile, East Africa: Bioaccumulation Potential, Source Apportionment, Ecological and Health Risk Assessment. *Environ. Pollut.* **2021**, *278*, 116855.
- (6) Routti, H.; Atwood, T. C.; Bechshoft, T.; Boltunov, A.; Ciesielski, T. M.; Desforges, J.-P.; Dietz, R.; Gabrielsen, G. W.; Jenssen, B. M.; Letcher, R. J.; McKinney, M. A.; Morris, A. D.; Rigét, F. F.; Sonne, C.; Styriehave, B.; Tartu, S. State of Knowledge on Current Exposure, Fate and Potential Health Effects of Contaminants in Polar Bears from the Circumpolar Arctic. *Sci. Total Environ.* **2019**, *664*, 1063–1083.
- (7) Wang, Y.; Yao, J.; Dai, J.; Ma, L.; Liu, D.; Xu, H.; Cui, Q.; Ma, J.; Zhang, H. Per- and Polyfluoroalkyl Substances (PFASs) in Blood of Captive Siberian Tigers in China: Occurrence and Associations with Biochemical Parameters. *Environ. Pollut.* **2020**, *265*, 114805.
- (8) Medici, E. P.; Fernandes-Santos, R. C.; Testa-José, C.; Godinho, A. F.; Brand, A.-F. Lowland Tapir Exposure to Pesticides and Metals in the Brazilian Cerrado. *Wildl. Res.* **2021**, *48* (5), 393–403.
- (9) Winfield, Z. C.; Mansouri, F.; Potter, C. W.; Sabin, R.; Trumble, S. J.; Usenko, S. Eighty Years of Chemical Exposure Profiles of Persistent Organic Pollutants Reconstructed through Baleen Whale Earplugs. *Sci. Total Environ.* **2020**, *737*, 139564.
- (10) Long, M.; Wielsøe, M.; Bonefeld-Jørgensen, E. C. Time Trend of Persistent Organic Pollutants and Metals in Greenlandic Inuit during 1994–2015. *Int. J. Environ. Res. Public Health* **2021**, *18* (5), 2774.
- (11) Govarts, E.; Iszatt, N.; Trnovec, T.; de Cock, M.; Eggesbø, M.; Palkovicova Murinova, L.; van de Bor, M.; Guxens, M.; Chevrier, C.; Koppen, G.; Lamoree, M.; Hertz-Picciotto, I.; Lopez-Espinosa, M.-J.; Lertxundi, A.; Grimalt, J. O.; Torrent, M.; Goñi-Irigoyen, F.; Vermeulen, R.; Legler, J.; Schoeters, G. Prenatal Exposure to Endocrine Disrupting Chemicals and Risk of Being Born Small for Gestational



Age: Pooled Analysis of Seven European Birth Cohorts. *Environ. Int.* **2018**, *115*, 267–278.

(12) Bai, X.; Zhang, B.; He, Y.; Hong, D.; Song, S.; Huang, Y.; Zhang, T. Triclosan and Triclocarbon in Maternal-Fetal Serum, Urine, and Amniotic Fluid Samples and Their Implication for Prenatal Exposure. *Environ. Pollut.* **2020**, *266*, 115117.

(13) Wu, Z.; He, C.; Han, W.; Song, J.; Li, H.; Zhang, Y.; Jing, X.; Wu, W. Exposure Pathways, Levels and Toxicity of Polybrominated Diphenyl Ethers in Humans: A Review. *Environ. Res.* **2020**, *187*, 109531.

(14) Matovu, H.; Ssebugere, P.; Sillanpää, M. Prenatal Exposure Levels of Polybrominated Diphenyl Ethers in Mother-Infant Pairs and Their Transplacental Transfer Characteristics in Uganda (East Africa). *Environ. Pollut.* **2020**, *258*, 113723.

(15) Landrigan, P. J.; Fuller, R.; Acosta, N. J. R.; Adeyi, O.; Arnold, R.; Basu, N.; Baldé, A. B.; Bertollini, R.; Bose-O'Reilly, S.; Boufford, J. I.; Breyse, P. N.; Chiles, T.; Mahidol, C.; Coll-Seck, A. M.; Cropper, M. L.; Fobil, J.; Fuster, V.; Greenstone, M.; Haines, A.; Hanrahan, D.; Hunter, D.; Khare, M.; Krupnick, A.; Lanphear, B.; Lohani, B.; Martin, K.; Mathiasen, K. V.; McTeer, M. A.; Murray, C. J. L.; Ndahimananjara, J. D.; Perera, F.; Potočnik, J.; Preker, A. S.; Ramesh, J.; Rockström, J.; Salinas, C.; Samson, L. D.; Sandilya, K.; Sly, P. D.; Smith, K. R.; Steiner, A.; Stewart, R. B.; Suk, W. A.; van Schayck, O. C. P.; Yadama, G. N.; Yumkella, K.; Zhong, M. The Lancet Commission on Pollution and Health. *Lancet.* **2018**, *391* (10119), 462–512.

(16) European Chemicals Agency. *What Is a Substance?* <https://echa.europa.eu/support/substance-identification/what-is-a-substance> (accessed 2020-10-16).

(17) United States Environmental Protection Agency. *Toxic Substances Control Act Inventory Representation for Chemical Substances of Unknown or Variable Composition, Complex Reaction Products and Biological Materials: UVCB Substances.* <https://www.epa.gov/sites/default/files/2015-05/documents/uvcb.pdf> (accessed 2021-08-14).

(18) Government of Canada. *Guidelines for the Notification and Testing of New Substances: Chemicals and Polymers, Pursuant to Section 69 of the Canadian Environmental Protection Act, 1999.* <https://publications.gc.ca/collections/Collection/En84-25-2005E.pdf> (accessed 2022-01-01).

(19) Sauer, U. G.; Barter, R. A.; Becker, R. A.; Benfenati, E.; Berggren, E.; Hubesch, B.; Hollnagel, H. M.; Inawaka, K.; Keene, A. M.; Mayer, P.; Plotzke, K.; Skoglund, R.; Albert, O. 21st Century Approaches for Evaluating Exposures, Biological Activity, and Risks of Complex Substances: Workshop Highlights. *Regul. Toxicol. Pharmacol.* **2020**, *111*, 104583.

(20) European Chemicals Agency. *Read-Across Assessment Framework (RAAF): Considerations on Multi Constituent Substances and UVCBs;* European Chemicals Agency: 2017. DOI: 10.2823/794394.

(21) Kutsarova, S. S.; Yordanova, D. G.; Karakolev, Y. H.; Stoeva, S.; Comber, M.; Hughes, C. B.; Vaiopoulou, E.; Dimitrov, S. D.; Mekenyan, O. G. UVCB Substances II: Development of an Endpoint-Nonspecific Procedure for Selection of Computationally Generated Representative Constituents. *Environ. Toxicol. Chem.* **2019**, *38* (3), 682–694.

(22) 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.* **2020**, *54*, 2575.

(23) Glüge, J.; Schinkel, L.; Hungerbühler, K.; Cariou, R.; Bogdal, C. Environmental Risks of Medium-Chain Chlorinated Paraffins (MCCPs): A Review. *Environ. Sci. Technol.* **2018**, *52* (12), 6743–6760.

(24) Conca. *REACH Roadmap for Petroleum Substances 2019 Update.* <https://www.conca.eu/wp-content/uploads/REACH-Roadmap-for-Petroleum-Substances-2019-update.pdf> (accessed 2021-10-21).

(25) Schymanski, E. L.; Singer, H. P.; Longrée, P.; Loos, M.; Ruff, M.; Stravs, M. A.; Ripollés Vidal, C.; Hollender, J. Strategies to Characterize Polar Organic Contamination in Wastewater: Exploring the Capability of High Resolution Mass Spectrometry. *Environ. Sci. Technol.* **2014**, *48* (3), 1811–1818.

(26) Muir, D. C. G.; Howard, P. H. Are There Other Persistent Organic Pollutants? A Challenge for Environmental Chemists. *Environ. Sci. Technol.* **2006**, *40* (23), 7157–7166.

(27) Zhang, X.; Sun, X.; Jiang, R.; Zeng, E. Y.; Sunderland, E. M.; Muir, D. C. G. Screening New Persistent and Bioaccumulative Organics in China's Inventory of Industrial Chemicals. *Environ. Sci. Technol.* **2020**, *54* (12), 7398–7408.

(28) Stempel, S.; Scheringer, M.; Ng, C. A.; Hungerbühler, K. Screening for PBT Chemicals among the “Existing” and “New” Chemicals of the EU. *Environ. Sci. Technol.* **2012**, *46* (11), 5680–5687.

(29) Salvito, D.; Fernandez, M.; Jenner, K.; Lyon, D. Y.; de Knecht, J.; Mayer, P.; MacLeod, M.; Eisenreich, K.; Leonards, P.; Cesnaitis, R.; León-Paumen, M.; Embry, M.; Déglin, S. E. Improving the Environmental Risk Assessment of Substances of Unknown or Variable Composition, Complex Reaction Products, or Biological Materials. *Environ. Toxicol. Chem.* **2020**, *39* (11), 2097–2108.

(30) Burns, D. T.; Walker, M. J.; Mussell, C. Chondroitin Sulfate: A Critical Review of Generic and Specific Problems in Its Characterization and Determination—An Exemplar of a Material with an Unknown or Variable Composition (UVCB). *J. AOAC Int.* **2018**, *101* (1), 196–202.

(31) Olkowska, E.; Polkowska, Ž.; Namieśnik, J. Analytical Procedures for the Determination of Surfactants in Environmental Samples. *Talanta* **2012**, *88*, 1–13.

(32) Kassotis, C. D.; Tillitt, D. E.; Lin, C.-H.; McElroy, J. A.; Nagel, S. C. Endocrine-Disrupting Chemicals and Oil and Natural Gas Operations: Potential Environmental Contamination and Recommendations to Assess Complex Environmental Mixtures. *Environ. Health Persp.* **2016**, *124* (3), 256–264.

(33) Maxim, L. D.; Niebo, R.; McConnell, E. E. Bentonite Toxicology and Epidemiology - a Review. *Inhal. Toxicol.* **2016**, *28* (13), 591–617.

(34) Rotter, S.; Beronius, A.; Boobis, A. R.; Hanberg, A.; van Klaveren, J.; Luijten, M.; Machera, K.; Nikolopoulou, D.; van der Voet, H.; Zilliciacus, J.; Solecki, R. Overview on Legislation and Scientific Approaches for Risk Assessment of Combined Exposure to Multiple Chemicals: The Potential EuroMix Contribution. *Crit. Rev. Toxicol.* **2018**, *48* (9), 796–814.

(35) Kienzler, A.; Bopp, S. K.; van der Linden, S.; Berggren, E.; Worth, A. Regulatory Assessment of Chemical Mixtures: Requirements, Current Approaches and Future Perspectives. *Regul. Toxicol. Pharmacol.* **2016**, *80*, 321–334.

(36) Escher, B. I.; Stapleton, H. M.; Schymanski, E. L. Tracking Complex Mixtures of Chemicals in Our Changing Environment. *Science* **2020**, *367* (6476), 388–392.

(37) Kortenkamp, A.; Faust, M. Regulate to Reduce Chemical Mixture Risk. *Science* **2018**, *361* (6399), 224–226.

(38) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36.

(39) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (3), 244–255.

(40) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015**, *7* (1), 23.

(41) Hepler-Smith, E. Molecular Bureaucracy: Toxicological Information and Environmental Protection. *Environ. Hist.* **2019**, *24* (3), 534–560.

(42) United States Environmental Protection Agency. *About the TSCA Chemical Substance Inventory.* <https://www.epa.gov/tscainventory/about-tsc-chemical-substance-inventory> (accessed 2021-11-25).

(43) European Chemicals Agency. *Guidance for the Identification and Naming of the Substances under REACH and CLP;* European Chemicals Agency: 2017. DOI: 10.2823/538683.

(44) European Chemicals Agency. *How to Prepare an Inquiry Dossier;* European Chemicals Agency: 2016. DOI: 10.2823/327956.

- (45) German Federal Office for Chemicals. *40th Meeting of Competent Authorities for REACH and CLP - Information Point. Registration of UVCB Substances: Provisions for the Comprehensive Description of the Composition and of Sameness*. <https://circabc.europa.eu/ui/group/a0b483a2-4c05-4058-addf-2a4de71b9a98/library/d8c3e869-1bfd-41f5-b903-89387821958b/details> (accessed 2021-12-21).
- (46) United States Environmental Protection Agency. *Chemical Substances of Unknown or Variable Composition, Complex Reaction Products and Biological Materials (UVCB Substance) on the TSCA Inventory*. <https://www.epa.gov/tscainventory/chemical-substances-unknown-or-variable-composition-complex-reaction-products-and> (accessed 2021-10-23).
- (47) Wang, Z.; Wiesinger, H.; Groh, K. Time to Reveal Chemical Identities of Polymers and UVCBs. *Environ. Sci. Technol.* **2021**, *55*, 14473.
- (48) Uphoff, A.; Aparicio, A. M. *Key Substance Identity Concepts UVCB*. European Chemicals Agency. [https://echa.europa.eu/documents/10162/22816622/key\\_substance\\_identity\\_concepts\\_uvcb\\_en.pdf/20c02695-17e3-4ad0-9057-9c63685ddd54](https://echa.europa.eu/documents/10162/22816622/key_substance_identity_concepts_uvcb_en.pdf/20c02695-17e3-4ad0-9057-9c63685ddd54) (accessed 2021-09-01).
- (49) Mayfield, J. *CDK Depict*. <https://www.simolecule.com/cdkdepict/depict.html> (accessed 2021-09-06).
- (50) Yordanova, D. G.; Patterson, T. J.; North, C. M.; Camenzuli, L.; Chapkanov, A. S.; Pavlov, T. S.; Mekenyan, O. G. Selection of Representative Constituents for Unknown, Variable, Complex, or Biological Origin Substance Assessment Based on Hierarchical Clustering. *Environ. Toxicol. Chem.* **2021**, *40* (11), 3205–3218.
- (51) Dimitrov, S. D.; Georgieva, D. G.; Pavlov, T. S.; Karakolev, Y. H.; Karamertzanis, P. G.; Rasenberg, M.; Mekenyan, O. G. UVCB Substances: Methodology for Structural Description and Application to Fate and Hazard Assessment. *Environ. Toxicol. Chem.* **2015**, *34* (11), 2450–2462.
- (52) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M. The CompTox Chemistry Dashboard: A Community Data Resource for Environmental Chemistry. *J. Cheminform* **2017**, *9* (1), 61.
- (53) European Commission. *Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and the Committee of the Regions - Chemicals Strategy for Sustainability Towards a Toxic-Free Environment*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A667%3AFIN> (accessed 2021-09-05).
- (54) United States Environmental Protection Agency. *CompTox Portal*. <https://comptox.epa.gov/> (accessed 2021-09-05).
- (55) *Markush Tools*. ChemAxon. <https://chemaxon.com/products/markush-tools> (accessed 2021-09-05).
- (56) *MOLGEN*. <https://www.molgen.de/> (accessed 2021-09-05).
- (57) Gugisch, R.; Kerber, A.; Kohnert, A.; Laue, R.; Meringer, M.; Rücker, C.; Wassermann, A. Chapter 6 - MOLGEN 5.0, A Molecular Structure Generator. In *Advances in Mathematical Chemistry and Applications*; Basak, S. C., Restrepo, G., Villaveces, J. L., Eds.; Bentham Science Publishers: 2015; pp 1113–138.
- (58) Schymanski, E. *schymane/RChemMass*. <https://github.com/schymane/RChemMass> (accessed 2020-08-16).
- (59) Williams, A. J.; Grulke, C.; McEachran, A.; Schymanski, E. Markush Enumeration to Manage, Mesh and Manipulate Substances of Unknown or Variable Composition. Presented at the 254th Annual Meeting of the American Chemical Society, 2017. [figshare.com. https://doi.org/10.23645/epacomptox.6455630.v1](https://doi.org/10.23645/epacomptox.6455630.v1).
- (60) *SciFinder*. Chemical Abstracts Service. <https://scifinder.cas.org> (accessed 2021-09-05).
- (61) European Chemicals Agency. *Medium-chain chlorinated paraffins (MCCP) - Substance Information - ECHA*. <https://echa.europa.eu/substance-information/-/substanceinfo/100.323.845> (accessed 2022-02-24).
- (62) Chibwe, L.; Myers, A. L.; De Silva, A. O.; Reiner, E. J.; Jobst, K.; Muir, D.; Yuan, B. C12–30  $\alpha$ -Bromo-Chloro “Alkenes”: Characterization of a Poorly Identified Flame Retardant and Potential Environmental Implications. *Environ. Sci. Technol.* **2019**, *53* (18), 10835–10844.
- (63) Leonards, P. E. G. *Concept for a Generic Information Strategy for PBT and vPvB Assessment of UVCB Substances*; R-16/06; Vrije Universiteit Amsterdam: 2017; pp 1–29. [https://research.vu.nl/ws/portalfiles/portal/59833096/Report\\_UVCB\\_assessment\\_final.pdf](https://research.vu.nl/ws/portalfiles/portal/59833096/Report_UVCB_assessment_final.pdf) (accessed 2022-04-15).
- (64) Banner, C.; Forbes, S. *Concawe Investigation of the Value of Spectral Data for the Identification of Petroleum UVCB Substances (9/20)*. [https://www.concawe.eu/wp-content/uploads/Rpt\\_20-09.pdf](https://www.concawe.eu/wp-content/uploads/Rpt_20-09.pdf) (accessed 2021-10-18).
- (65) Qian, K. Molecular Characterization of Heavy Petroleum by Mass Spectrometry and Related Techniques. *Energy Fuels* **2021**, *35* (22), 18008–18018.
- (66) Forbes, S. *Concawe Substance Identification Group. Concawe Substance Identification Group Analytical Program Report (Abridged Version 5/19)*. <https://www.concawe.eu/wp-content/uploads/Concawe-Substance-Identification-Group-Analytical-Program-Report-Abridged-Version.pdf> (accessed 2021-10-18).
- (67) Roman-Hubers, A. T.; Cordova, A. C.; Aly, N. A.; McDonald, T. J.; Lloyd, D. T.; Wright, F. A.; Baker, E. S.; Chiu, W. A.; Rusyn, I. Data Processing Workflow to Identify Structurally Related Compounds in Petroleum Substances Using Ion Mobility Spectrometry-Mass Spectrometry. *Energy Fuels* **2021**, *35* (13), 10529–10539.
- (68) Grimm, F. A.; Russell, W. K.; Luo, Y.-S.; Iwata, Y.; Chiu, W. A.; Roy, T.; Boogaard, P. J.; Ketelslegers, H. B.; Rusyn, I. Grouping of Petroleum Substances as Example UVCBs by Ion Mobility-Mass Spectrometry to Enable Chemical Composition-Based Read-Across. *Environ. Sci. Technol.* **2017**, *51* (12), 7197–7207.
- (69) House, J. S.; Grimm, F. A.; Klaren, W. D.; Dalzell, A.; Kuchi, S.; Zhang, S.-D.; Lenz, K.; Boogaard, P. J.; Ketelslegers, H. B.; Gant, T. W.; Wright, F. A.; Rusyn, I. Grouping of UVCB Substances with New Approach Methodologies (NAMs) Data. *ALTEX* **2020**, *38* (1), 123–137.
- (70) Onel, M.; Beykal, B.; Ferguson, K.; Chiu, W. A.; McDonald, T. J.; Zhou, L.; House, J. S.; Wright, F. A.; Sheen, D. A.; Rusyn, I.; Pistikopoulos, E. N. Grouping of Complex Substances Using Analytical Chemistry Data: A Framework for Quantitative Evaluation and Visualization. *PLoS One* **2019**, *14* (10), No. e0223517.
- (71) Teixeira, B.; Marques, A.; Ramos, C.; Neng, N. R.; Nogueira, J. M. F.; Saraiva, J. A.; Nunes, M. L. Chemical Composition and Antibacterial and Antioxidant Properties of Commercial Essential Oils. *Ind. Crops Prod.* **2013**, *43*, 587–595.
- (72) Møller, M. T.; Birch, H.; Sjøholm, K. K.; Hammershøj, R.; Jenner, K.; Mayer, P. Biodegradation of an Essential Oil UVCB - Whole Substance Testing and Constituent Specific Analytics Yield Biodegradation Kinetics of Mixture Constituents. *Chemosphere* **2021**, *278*, 130409.
- (73) Chambre, D. R.; Moisa, C.; Lupitu, A.; Copolovici, L.; Pop, G.; Copolovici, D.-M. Chemical Composition, Antioxidant Capacity, and Thermal Behavior of Satureja Hortensis Essential Oil. *Sci. Rep* **2020**, *10* (1), 21322.
- (74) Yuan, B.; Tay, J. H.; Papadopoulou, E.; Haug, L. S.; Padilla-Sánchez, J. A.; de Wit, C. A. Complex Mixtures of Chlorinated Paraffins Found in Hand Wipes of a Norwegian Cohort. *Environ. Sci. Technol. Lett.* **2020**, *7* (3), 198–205.
- (75) Park, K. S.; Kim, Y. J.; Choe, E. K. Composition Characterization of Fatty Acid Zinc Salts by Chromatographic and NMR Spectroscopic Analyses on Their Fatty Acid Methyl Esters. *J. Anal. Methods Chem.* **2019**, *2019*, 7594767.
- (76) Oertel, A.; Maul, K.; Menz, J.; Kronsbein, A. L.; Sittner, D.; Springer, A.; Müller, A.-K.; Herbst, U.; Schlegel, K.; Schulte, A. *REACH Compliance: Data Availability in REACH Registrations Part 2: Evaluation of Data Waiving and Adaptations for Chemicals  $\geq$  1000 tpa (Series 64/2018)*. <https://www.umweltbundesamt.de/en/publikationen/reach-compliance-data-availability-in-reach> (accessed 2021-08-14).
- (77) De Graaff, R.; Forbes, S.; Gennart, J. P.; Gimeno Cortes, M. J.; Hovius, H.; King, D.; Kleise, H.; Martinez Martin, C.; Montanari, L.; Pinzuti, M.; Pollack, H.; Ruggieri, P.; Thomas, M.; Walton, A;



- Dmytrasz, B. REACH. *Analytical Characterisation of Petroleum UVCB Substances*; Report No. 7/12; Concawe: 2012. [https://www.concawe.eu/wp-content/uploads/2017/01/rpt\\_12-7-2012-05443-01-e.pdf](https://www.concawe.eu/wp-content/uploads/2017/01/rpt_12-7-2012-05443-01-e.pdf) (accessed 2020-11-03).
- (78) Verhaar, H. J. M.; Busser, F. J. M.; Hermens, J. L. M. Surrogate Parameter for the Baseline Toxicity Content of Contaminated Water: Simulating the Bioconcentration of Mixtures of Pollutants and Counting Molecules. *Environ. Sci. Technol.* **1995**, *29* (3), 726–734.
- (79) Fenner, K.; Scheringer, M. The Need for Chemical Simplification As a Logical Consequence of Ever-Increasing Chemical Pollution. *Environ. Sci. Technol.* **2021**, *55*, 14470.
- (80) Australian Government Department of Health. *Working out your hazards using read-across information*. <https://www.industrialchemicals.gov.au/help-and-guides/working-out-your-hazards-using-read-across-information> (accessed 2022-04-15).
- (81) Cousins, I. T.; DeWitt, J. C.; Glüge, J.; Goldenman, G.; Herzke, D.; Lohmann, R.; Miller, M.; Ng, C. A.; Scheringer, M.; Vierke, L.; Wang, Z. Strategies for Grouping Per- and Polyfluoroalkyl Substances (PFAS) to Protect Human and Environmental Health. *Environmental Science: Processes & Impacts* **2020**, *22* (7), 1444–1460.
- (82) Organization for Economic Co-operation and Development. *OECD Guidance for Characterising Oleochemical Substances for Assessment Purposes*. OECD Series on Testing and Assessment 193; OECD Publishing: 2017. DOI: 10.1787/9789264274778-en.
- (83) Organization for Economic Co-operation and Development. *OECD Guidance for Characterising Hydrocarbon Solvents for Assessment Purposes*; OECD Series on Testing and Assessment 230; OECD Publishing: 2017. DOI: 10.1787/9789264274808-en.
- (84) King, D. J.; Lyne, R. L.; Girling, A.; Peterson, D. R.; Stephenson, R.; Short, D. *Environmental Risk Assessment of Petroleum Substances: The Hydrocarbon Block Method*; Report No. 96/52; Concawe: 1996. [https://www.concawe.eu/wp-content/uploads/2017/01/rpt\\_96-52-2004-01719-01-e.pdf](https://www.concawe.eu/wp-content/uploads/2017/01/rpt_96-52-2004-01719-01-e.pdf). (accessed 2021-05-17).
- (85) Organization for Economic Co-operation and Development. *OECD Guidance on Grouping of Chemicals*, 2nd ed.; OECD Series on Testing and Assessment 194; OECD Publishing: 2017. DOI: 10.1787/9789264274679-en.
- (86) Carrillo, J.-C.; Adenuga, M. D.; Mckee, R. H. The Sub-Chronic Toxicity of Regular White Spirit in Rats. *Regul. Toxicol. Pharmacol.* **2014**, *70* (1), 222–230.
- (87) Catlin, N. R.; Collins, B. J.; Auerbach, S. S.; Ferguson, S. S.; Harnly, J. M.; Gennings, C.; Waidyanatha, S.; Rice, G. E.; Smith-Roe, S. L.; Witt, K. L.; Rider, C. V. How Similar Is Similar Enough? A Sufficient Similarity Case Study with Ginkgo Biloba Extract. *Food Chem. Toxicol.* **2018**, *118*, 328–339.
- (88) United States Environmental Protection Agency. *Guidance for Creating Generic Names for Confidential Chemical Substance Identity Reporting under the Toxic Substances Control Act (EPA 743B18001)*. <https://www.epa.gov/tsca-inventory/guidance-creating-generic-names-confidential-chemical-substance-identity-reporting> (accessed 2021-09-30).
- (89) Australian Government Department of Health. *Apply for protection of chemical name as confidential business information using AACN*. <https://www.industrialchemicals.gov.au/business/apply-confidentiality-data-and-information/apply-protection-chemical-name-confidential-business-information-using-aacn> (accessed 2022-04-15).
- (90) Ellis, G. Special Issue 40th ISEO: Toxicological Challenges for Essential Oils in REACH. *Flavour Fragr J.* **2010**, *25* (3), 138–144.
- (91) Sachdev, N. *ninasachdev/UVCB-MInChI*. <https://github.com/ninasachdev/UVCB-MInChI> (accessed 2022-01-01).
- (92) Clark, A. M.; McEwen, L. R.; Gedeck, P.; Bunin, B. A. Capturing Mixture Composition: An Open Machine-Readable Format for Representing Mixed Substances. *J. Cheminform* **2019**, *11* (1), 33.
- (93) Clark, A. M. *Online Mixtures Demo, with MInChI Generator*. Cheminformatics 2.0. <https://cheminf20.org/2020/05/11/online-mixtures-demo-with-minchi-generator/> (accessed 2020-12-28).
- (94) American Chemical Society. CAS REGISTRY <https://www.cas.org/cas-data/cas-registry> (accessed 2022-04-15).
- (95) Personal Care Products Council. INCI. <https://www.personalcarecouncil.org/resources/inci/> (accessed 2021-11-22).
- (96) Government of Canada. *Categorizing Substances on the Domestic Substances List*. <https://www.canada.ca/en/environment-climate-change/services/canadian-environmental-protection-act-registry/substances-list/domestic/domestic-list.html> (accessed 2022-01-01).
- (97) Government of Canada. *Chemicals Management Plan*. <https://www.canada.ca/en/health-canada/services/chemical-substances/chemicals-management-plan.html> (accessed 2021-10-23).
- (98) European Chemicals Agency. *How to Gather Information to Register a Multi-constituent or a UVCB substance - Toxicological Information*. [https://echa.europa.eu/documents/10162/23221373/example\\_multiconstituent\\_uvcb\\_en.pdf/74f2fdce-ba96-cbb2-fca8-f0fb6c53e5f4](https://echa.europa.eu/documents/10162/23221373/example_multiconstituent_uvcb_en.pdf/74f2fdce-ba96-cbb2-fca8-f0fb6c53e5f4) (accessed 2021-08-24).
- (99) Australian Government Department of Health. *Assessment Search, Australian Industrial Chemicals Introduction Scheme*. [https://www.industrialchemicals.gov.au/chemical-information/search-assessments-keywords?keywords=essential+oil&field\\_assessment\\_type=](https://www.industrialchemicals.gov.au/chemical-information/search-assessments-keywords?keywords=essential+oil&field_assessment_type=) (accessed 2022-04-15).
- (100) United Nations Economic Commission for Europe. *Globally Harmonized System of Classification and Labelling of Chemicals (GHS Rev. 9, 2021)*. <https://unece.org/transport/standards/transport/dangerous-goods/ghs-rev9-2021> (accessed 2021-10-23).
- (101) International Petroleum Industry Environmental Conservation Association. *The Application of Globally Harmonized System (GHS) Criteria to Petroleum Substances*. <https://www.ipieca.org/resources/good-practice/the-application-of-globally-harmonized-system-ghs-criteria-to-petroleum-substances/> (accessed 2021-10-04).
- (102) Verdonck, F.; Waeterschoot, H.; Van Sprang, P.; Vercaigne, I.; Delbeke, K.; Simons, C.; Verougstraete, V. MeClas: An Online Tool for Hazard Identification and Classification of Complex Inorganic Metal-Containing Materials. *Regul. Toxicol. Pharmacol.* **2017**, *89*, 232–239.
- (103) Clark, C. R.; McKee, R. H.; Freeman, J. J.; Swick, D.; Mahagaokar, S.; Pigram, G.; Roberts, L. G.; Smulders, C. J.; Beatty, P. W. A GHS-Consistent Approach to Health Hazard Classification of Petroleum Substances, a Class of UVCB Substances. *Regul. Toxicol. Pharmacol.* **2013**, *67* (3), 409–420.
- (104) Mckee, R. H.; Adenuga, M. D.; Carrillo, J.-C. Characterization of the Toxicological Hazards of Hydrocarbon Solvents. *Crit. Rev. Toxicol.* **2015**, *45* (4), 273–365.
- (105) Bergal, M.; Puginier, M.; Gerbeix, C.; Groux, H.; Roso, A.; Cottrez, F.; Milius, A. In Vitro Testing Strategy for Assessing the Skin Sensitizing Potential of “Difficult to Test” Cosmetic Ingredients. *Toxicol. in Vitro* **2020**, *65*, 104781.
- (106) European Chemicals Agency. *Registration Dossier - Juniper, Juniperus virginiana, ext.* <https://echa.europa.eu/registration-dossier/-/registered-dossier/20863/2/1> (accessed 2022-04-15).
- (107) European Chemicals Agency. *Guidance on Information Requirements and Chemical Safety Assessment: Chapter R.11: PBT and VpVB Assessment*; European Chemicals Agency: 2017. DOI: 10.2823/128621.
- (108) Redman, A. D.; Parkerton, T. F.; McGrath, J. A.; Di Toro, D. M. PETROTOX: An Aquatic Toxicity Model for Petroleum Substances. *Environ. Toxicol. Chem.* **2012**, *31* (11), 2498–2506.
- (109) de Bruijn, J.; Hansen, B. G.; Johansson, S.; Luotamo, M.; Munn, S. J.; Olsen, S. I.; Olsson, H.; Paya-Perez, A. B.; Pedersen, F.; Rasmussen, K.; Sokull-Kluttgen, B. *Technical Guidance Document on Risk Assessment. Part 1. Part 2. EUR 20418 EN. 2002. JRC23785*. <https://publications.jrc.ec.europa.eu/repository/handle/JRC23785> (accessed 2021-10-23).
- (110) Wassenaar, P. N. H.; Verbruggen, E. M. J. Persistence, Bioaccumulation and Toxicity-Assessment of Petroleum UVCBs: A Case Study on Alkylated Three-Ring PAHs. *Chemosphere* **2021**, *276*, 130113.
- (111) Prosser, C. M.; Redman, A. D.; Prince, R. C.; Paumen, M. L.; Letinski, D. J.; Butler, J. D. Evaluating Persistence of Petroleum Hydrocarbons in Aerobic Aqueous Media. *Chemosphere* **2016**, *155*, 542–549.
- (112) Rorije, E. *PBT Evaluation for UVCB Substances*. <https://rvs.rivm.nl/sites/default/files/2018-05/>

2017%2520presentatie%2520PBT-beoordeling%2520van%2520UVCB.pdf (accessed 2021-11-05).

(113) Åhlman, M.; Bergman, Å.; Darnerud, P. O.; Egestad, B.; Sjövall, J. Chlorinated Paraffins: Formation of Sulphur-Containing Metabolites of Polychlorohexadecane in Rats. *Xenobiotica* **1986**, *16* (3), 225–232.

(114) Brillet, F.; Cregut, M.; Durand, M. J.; Sweetlove, C.; Chenèble, J. C.; L'Haridon, J.; Thouand, G. Biodegradability Assessment of Complex Chemical Mixtures Using a Carbon Balance Approach. *Green Chem.* **2018**, *20* (5), 1031–1041.

(115) Brown, D. M.; Lyon, D.; Saunders, D. M. V.; Hughes, C. B.; Wheeler, J. R.; Shen, H.; Whale, G. Biodegradability Assessment of Complex, Hydrophobic Substances: Insights from Gas-to-Liquid (GTL) Fuel and Solvent Testing. *Sci. Total Environ.* **2020**, *727*, 138528.

(116) Birch, H.; Hammershøj, R.; Mayer, P. Determining Biodegradation Kinetics of Hydrocarbons at Low Concentrations: Covering 5 and 9 Orders of Magnitude of Kow and Kaw. *Environ. Sci. Technol.* **2018**, *52* (4), 2143–2151.

(117) Knudsmark Sjøholm, K.; Birch, H.; Hammershøj, R.; Saunders, D. M. V.; Dechesne, A.; Loibner, A. P.; Mayer, P. Determining the Temperature Dependency of Biodegradation Kinetics for 34 Hydrocarbons While Avoiding Chemical and Microbial Confounding Factors. *Environ. Sci. Technol.* **2021**, *55* (16), 11091–11101.

(118) Hammershøj, R.; Birch, H.; Redman, A. D.; Mayer, P. Mixture Effects on Biodegradation Kinetics of Hydrocarbons in Surface Water: Increasing Concentrations Inhibited Degradation Whereas Multiple Substrates Did Not. *Environ. Sci. Technol.* **2019**, *53* (6), 3087–3094.

(119) Hammershøj, R.; Sjøholm, K. K.; Birch, H.; Brandt, K. K.; Mayer, P. Biodegradation Kinetics Testing of Two Hydrophobic UVCBs - Potential for Substrate Toxicity Supports Testing at Low Concentrations. *Environ. Sci.: Process. Impacts* **2020**, *22*, 2172.

(120) Sühling, R.; Chen, C.-E.; McLachlan, M. S.; MacLeod, M. Bioconcentration of Cedarwood Oil Constituents in Rainbow Trout. *Environ. Sci.: Process. Impacts* **2021**, *23*, 689.

(121) Sühling, R.; Knudsmark Sjøholm, K.; Mayer, P.; MacLeod, M. Combining Headspace Solid-Phase Microextraction with Internal Benchmarking to Determine the Elimination Kinetics of Hydrophobic UVCBs. *Environ. Sci. Technol.* **2021**, *55*, 11125.

(122) Castro, M.; Sobek, A.; Yuan, B.; Breitholtz, M. Bioaccumulation Potential of CPs in Aquatic Organisms: Uptake and Depuration in *Daphnia Magna*. *Environ. Sci. Technol.* **2019**, *53* (16), 9533–9541.

(123) Mayer, P.; Reichenberg, F. Can Highly Hydrophobic Organic Substances Cause Aquatic Baseline Toxicity and Can They Contribute to Mixture Toxicity? *Environ. Toxicol. Chem.* **2006**, *25* (10), 2639–2644.

(124) Trac, L. N.; Schmidt, S. N.; Holmstrup, M.; Mayer, P. Headspace Passive Dosing of Volatile Hydrophobic Organic Chemicals from a Lipid Donor—Linking Their Toxicity to Well-Defined Exposure for an Improved Risk Assessment. *Environ. Sci. Technol.* **2019**, *53* (22), 13468–13476.

(125) Trac, L. N.; Sjøholm, K. K.; Birch, H.; Mayer, P. Passive Dosing of Petroleum and Essential Oil UVCBs—Whole Mixture Toxicity Testing at Controlled Exposure. *Environ. Sci. Technol.* **2021**, *55* (9), 6150–6159.

(126) Hammershøj, R.; Birch, H.; Sjøholm, K. K.; Mayer, P. Accelerated Passive Dosing of Hydrophobic Complex Mixtures—Controlling the Level and Composition in Aquatic Tests. *Environ. Sci. Technol.* **2020**, *54* (8), 4974–4983.

(127) Luo, Y.-S.; Ferguson, K. C.; Rusyn, I.; Chiu, W. A. In Vitro Bioavailability of the Hydrocarbon Fractions of Dimethyl Sulfoxide Extracts of Petroleum Substances. *Toxicol. Sci.* **2020**, *174* (2), 168–177.

(128) Swigert, J. P.; Lee, C.; Wong, D. C. L.; Podhasky, P. Aquatic Hazard and Biodegradability of Light and Middle Atmospheric Distillate Petroleum Streams. *Chemosphere* **2014**, *108*, 1–9.

(129) Carrillo, J.-C.; Adenuga, M. D.; Momin, F.; McKee, R. H. The Sub-Chronic Toxicity of a Naphthenic Hydrocarbon Solvent in Rats. *Regul. Toxicol. Pharmacol.* **2018**, *95*, 323–332.

(130) Wang, X.; Zhu, J.; Xue, Z.; Jin, X.; Jin, Y.; Fu, Z. The Environmental Distribution and Toxicity of Short-Chain Chlorinated Paraffins and Underlying Mechanisms: Implications for Further

Toxicological Investigation. *Science of The Total Environment* **2019**, *695*, 133834.

(131) Ren, X.; Zhang, H.; Geng, N.; Xing, L.; Zhao, Y.; Wang, F.; Chen, J. Developmental and Metabolic Responses of Zebrafish (*Danio Rerio*) Embryos and Larvae to Short-Chain Chlorinated Paraffins (SCCPs) Exposure. *Science of The Total Environment* **2018**, *622–623*, 214–221.

(132) Brooke, D. N.; Crookes, M. J. Case Study on Toxicological Interactions of Chlorinated Paraffins. Presented at the Seventh Meeting, Persistent Organic Pollutants Review Committee; UNEP/POPS/POPRC.7/INF/15; Geneva, Switzerland, 2011. <http://chm.pops.int/TheConvention/POPsReviewCommittee/Meetings/POPRC7/POPRC7Documents/tabid/2267/Default.aspx> (accessed 2022-04-15).

(133) European Chemicals Agency. *What Is Not a Substance?* <https://echa.europa.eu/support/substance-identification/what-is-not-a-substance> (accessed 2021-09-15).

(134) United States Environmental Protection Agency. *Toxic Substances Control Act Inventory Representation for Products Containing Two or More Substances: Formulated and Statutory Mixtures*. <https://www.epa.gov/sites/default/files/2015-05/documents/mixtures.pdf> (accessed 2021-09-15).

(135) de Bruijn, J. A Possible Approach to Safe-Guarding Risks of Non-intentional Mixtures: a View from ECHA. Presented at the Workshop on a Pragmatic Approach to Address the Risk from Combined Exposure to Non-intentional Mixtures of Chemicals—REACH as an Example, Leiden, The Netherlands, March 5–6, 2020 (last accessed 14 August 2021).

(136) Environment and Climate Change Canada. *Draft Screening Assessment - Resins and Rosins Group*. <https://www.canada.ca/en/environment-climate-change/services/evaluating-existing-substances/draft-screening-assessment-resins-rosins-group.html> (accessed 2021-11-23).

(137) Wheeler, J. R.; Lyon, D.; Di Paolo, C.; Grosso, A.; Crane, M. Challenges in the Regulatory Use of Water-Accommodated Fractions for Assessing Complex Substances. *Environ. Sci. Eur.* **2020**, *32* (1), 153.

(138) Cedergreen, N. Quantifying Synergy: A Systematic Review of Mixture Toxicity Studies within Environmental Toxicology. *PLoS One* **2014**, *9* (5), No. e96580.

(139) Martin, O.; Scholze, M.; Ermler, S.; McPhie, J.; Bopp, S. K.; Kienzler, A.; Parissis, N.; Kortenkamp, A. Ten Years of Research on Synergisms and Antagonisms in Chemical Mixtures: A Systematic Review and Quantitative Reappraisal of Mixture Studies. *Environ. Int.* **2021**, *146*, 106206.

(140) Rider, C. V.; Dinse, G. E.; Umbach, D. M.; Simmons, J. E.; Hertzberg, R. C. Predicting Mixture Toxicity with Models of Additivity. In *Chemical Mixtures and Combined Chemical and Nonchemical Stressors: Exposure, Toxicity, Analysis, and Risk*; Rider, C. V., Simmons, J. E., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp 235–270. DOI: 10.1007/978-3-319-56234-6\_9.

(141) Escher, B.; Braun, G.; Zarfl, C. Exploring the Concepts of Concentration Addition and Independent Action Using a Linear Low-Effect Mixture Model. *Environ. Toxicol. Chem.* **2020**, *39* (12), 2552–2559.

(142) Villeneuve, D. L.; Blankenship, A. L.; Giesy, J. P. Derivation and Application of Relative Potency Estimates Based on in Vitro Bioassay Results. *Environ. Toxicol. Chem.* **2000**, *19* (11), 2835–2843.

(143) Van den Berg, M.; Birnbaum, L. S.; Denison, M.; De Vito, M.; Farland, W.; Feeley, M.; Fiedler, H.; Hakansson, H.; Hanberg, A.; Haws, L.; Rose, M.; Safe, S.; Schrenk, D.; Tohyama, C.; Tritscher, A.; Tuomisto, J.; Tysklind, M.; Walker, N.; Peterson, R. E. The 2005 World Health Organization Reevaluation of Human and Mammalian Toxic Equivalency Factors for Dioxins and Dioxin-Like Compounds. *Toxicol. Sci.* **2006**, *93* (2), 223–241.

(144) Scientific Committees on Health and Environmental Risks, Emerging and Newly Identified Health Risks, and Consumer Safety. *Toxicity and Assessment of Chemical Mixtures*; European Union: 2012. DOI: 10.2772/21444.



- (145) Martin, O. V.; Evans, R. M.; Faust, M.; Kortenkamp, A. A Human Mixture Risk Assessment for Neurodevelopmental Toxicity Associated with Polybrominated Diphenyl Ethers Used as Flame Retardants. *Environ. Health Persp.* **2017**, *125* (8), 087016.
- (146) Warne, M. S. J.; Hawker, D. W. The Number of Components in a Mixture Determines Whether Synergistic and Antagonistic or Additive Toxicity Predominate: The Funnel Hypothesis. *Ecotoxicology and Environmental Safety* **1995**, *31* (1), 23–28.
- (147) van Wezel, A. P.; Opperhuizen, A. Narcosis Due to Environmental Pollutants in Aquatic Organisms: Residue-Based Toxicity, Mechanisms, and Membrane Burdens. *Critical Reviews in Toxicology* **1995**, *25* (3), 255–279.
- (148) McCarty, L. S.; Mackay, D.; Smith, A. D.; Ozburn, G. W.; Dixon, D. G. Residue-Based Interpretation of Toxicity and Bioconcentration QSARs from Aquatic Bioassays: Polar Narcotic Organics. *Ecotoxicology and Environmental Safety* **1993**, *25* (3), 253–270.
- (149) Escher, B. I.; Ashauer, R.; Dyer, S.; Hermens, J. L.; Lee, J.-H.; Leslie, H. A.; Mayer, P.; Meador, J. P.; Warne, M. S. Crucial Role of Mechanisms and Modes of Toxic Action for Understanding Tissue Residue Toxicity and Internal Effect Concentrations of Organic Chemicals. *Integrated Environmental Assessment and Management* **2011**, *7* (1), 28–49.
- (150) van Wezel, A. P.; de Vries, D. A. M.; Sijm, D. T. H. M.; Opperhuizen, A. Use of the Lethal Body Burden in the Evaluation of Mixture Toxicity. *Ecotoxicology and Environmental Safety* **1996**, *35* (3), 236–241.
- (151) McGrath, J. A.; Di Toro, D. M. Validation of the Target Lipid Model for Toxicity Assessment of Residual Petroleum Constituents: Monocyclic and Polycyclic Aromatic Hydrocarbons. *Environ. Toxicol. Chem.* **2009**, *28* (6), 1130–1148.
- (152) Di Toro, D. M.; McGrath, J. A. Technical Basis for Narcotic Chemicals and Polycyclic Aromatic Hydrocarbon Criteria. II. Mixtures and Sediments. *Environ. Toxicol. Chem.* **2000**, *19* (8), 1971–1982.
- (153) Environment and Climate Change Canada; Health Canada. *Screening Assessment, Anthraquinones Group*. <https://www.canada.ca/en/environment-climate-change/services/evaluating-existing-substances/final-screening-assessment-anthraquinones-group.html> (accessed 2021-09-22).
- (154) Environment and Climate Change Canada; Health Canada. *Screening Assessment Sector-Specific Inorganic Unknown or Variable Composition, Complex Reaction Products and Biological Materials Group*. <https://www.canada.ca/en/environment-climate-change/services/evaluating-existing-substances/screening-assessment-sector-specific-inorganic-uvcs-group.html> (accessed 2021-09-27).
- (155) Verdonck, F.; Iaccino, F.; Lacasse, K.; Oorts, K.; Van Assche, F.; Verougstraete, V.; Vetter, D. Risk Assessment of Exposure to Inorganic Substances of UVCBs (Unknown or Variable Composition, Complex Reaction Products, or Biological Materials) During Manufacturing (Recycling) of Metals. In *Risk Management of Complex Inorganic Materials*; Elsevier: **2018**; pp 191–205.
- (156) European Chemicals Agency. *Lead Alloy, Base, Pb, Sn, Dross Registration Dossier*. <https://echa.europa.eu/registration-dossier/-/registered-dossier/15040/5/1/?documentUUIID=2096c9b9-8a87-4032-802f-dd3bd72bc725> (accessed 2021-09-29).
- (157) Environment and Climate Change Canada. *Science Approach Document: Ecological Risk Classification of Organic Substances*. <https://www.ec.gc.ca/ese-ees/default.asp?lang=En&n=A96E2E98-1#toc042> (accessed 2021-09-29).
- (158) Redman, A. D.; Parkerton, T. F.; Comber, M. H.; Paumen, M. L.; Eadsforth, C. V.; Dmytrasz, B.; King, D.; Warren, C. S.; den Haan, K.; Djemel, N. PETRORISK: A Risk Assessment Framework for Petroleum Substances. *Integr. Environ. Assess. Manag.* **2014**, *10* (3), 437–448.
- (159) McKee, R. H.; Tibaldi, R.; Adenuga, M. D.; Carrillo, J.-C.; Margary, A. Assessment of the Potential Human Health Risks from Exposure to Complex Substances in Accordance with REACH Requirements. “White Spirit” as a Case Study. *Regul. Toxicol. Pharmacol.* **2018**, *92*, 439–457.
- (160) Bierkens, J.; Geerts, L. Environmental Hazard and Risk Characterisation of Petroleum Substances: A Guided “Walking Tour” of Petroleum Hydrocarbons. *Environ. Int.* **2014**, *66*, 182–193.
- (161) Rorije, E.; Verbruggen, E.; Knecht, J. A. *Service Request on a Critical Review of the Environmental and Physicochemical Methodologies Commonly Employed in the Environmental Risk Assessment of Petroleum Substances in the Context of REACH Registrations (Framework Contract No. ECHA/2008/2; Reference No. ECHA/2008/02/SR30)*. [https://echa.europa.eu/documents/10162/17221/review\\_environmental\\_physicochemical\\_methodol\\_en.pdf](https://echa.europa.eu/documents/10162/17221/review_environmental_physicochemical_methodol_en.pdf) (accessed 2022-04-13).
- (162) Environment and Climate Change Canada; Health Canada. *Screening Assessment, Substituted Diphenylamines*. <https://www.canada.ca/en/environment-climate-change/services/evaluating-existing-substances/screening-assessment-substituted-diphenylamines.html> (accessed 2021-11-23).
- (163) European Centre for Ecotoxicology and Toxicology of Chemicals. *Development of Guidance for Assessing the Impact of Mixtures of Chemicals in the Aquatic Environment (TR 111)*. <https://www.ecetoc.org/publication/tr-111-development-of-guidance-for-assessing-the-impact-of-mixtures-of-chemicals-in-the-aquatic-environment/> (accessed 2021-11-24).
- (164) Adolfsson-Erici, M.; Åkerman, G.; McLachlan, M. S. Internal Benchmarking Improves Precision and Reduces Animal Requirements for Determination of Fish Bioconcentration Factors. *Environ. Sci. Technol.* **2012**, *46* (15), 8205–8211.
- (165) Backhaus, T.; Faust, M. Predictive Environmental Risk Assessment of Chemical Mixtures: A Conceptual Framework. *Environ. Sci. Technol.* **2012**, *46* (5), 2564–2573.
- (166) Meek, M. E.; Boobis, A. R.; Crofton, K. M.; Heinemeyer, G.; Van Raaij, M.; Vickers, C. Risk Assessment of Combined Exposure to Multiple Chemicals: A WHO/IPCS Framework. *Regul. Toxicol. Pharmacol.* **2011**, *60* (2), S1–S14.
- (167) International Union of Pure and Applied Chemistry. *InChI Extension for Mixture Composition*. <https://iupac.org/project/2015-025-4-800> (accessed 2021-11-02).
- (168) Washington, J. W.; Rosal, C. G.; McCord, J. P.; Strynar, M. J.; Lindstrom, A. B.; Bergman, E. L.; Goodrow, S. M.; Tadesse, H. K.; Pilant, A. N.; Washington, B. J.; Davis, M. J.; Stuart, B. G.; Jenkins, T. M. Nontargeted Mass-Spectral Detection of Chloroperfluoropolyether Carboxylates in New Jersey Soils. *Science* **2020**, *368* (6495), 1103–1107.
- (169) Liu, Y.; D’Agostino, L. A.; Qu, G.; Jiang, G.; Martin, J. W. High-Resolution Mass Spectrometry (HRMS) Methods for Nontarget Discovery and Characterization of Poly- and per-Fluoroalkyl Substances (PFASs) in Environmental and Human Samples. *Trends Anal. Chem.* **2019**, *121*, 115420.
- (170) Meekel, N.; Vughs, D.; Béen, F.; Brunner, A. M. Online Prioritization of Toxic Compounds in Water Samples through Intelligent HRMS Data Acquisition. *Anal. Chem.* **2021**, *93* (12), 5071–5080.
- (171) Hohrenk, L. L.; Vosough, M.; Schmidt, T. C. Implementation of Chemometric Tools To Improve Data Mining and Prioritization in LC-HRMS for Nontarget Screening of Organic Micropollutants in Complex Water Matrixes. *Anal. Chem.* **2019**, *91* (14), 9213–9220.
- (172) Du, X.; Yuan, B.; Zhou, Y.; Benskin, J. P.; Qiu, Y.; Yin, G.; Zhao, J. Short-, Medium-, and Long-Chain Chlorinated Paraffins in Wildlife from Paddy Fields in the Yangtze River Delta. *Environ. Sci. Technol.* **2018**, *52* (3), 1072–1080.
- (173) Yuan, B.; Lysak, D. H.; Soong, R.; Haddad, A.; Hisatsune, A.; Moser, A.; Golotvin, S.; Argyropoulos, D.; Simpson, A. J.; Muir, D. C. G. Chlorines Are Not Evenly Substituted in Chlorinated Paraffins: A Predicted NMR Pattern Matching Framework for Isomeric Discrimination in Complex Contaminant Mixtures. *Environ. Sci. Technol. Lett.* **2020**, *7* (7), 496–503.
- (174) Yuan, B.; Benskin, J. P.; Chen, C.-E. L.; Bergman, A. Determination of Chlorinated Paraffins by Bromide-Anion Attachment Atmospheric-Pressure Chemical Ionization Mass Spectrometry. *Environ. Sci. Technol. Lett.* **2018**, *5* (6), 348–353.

(175) Reth, M.; Zencak, Z.; Oehme, M. New Quantification Procedure for the Analysis of Chlorinated Paraffins Using Electron Capture Negative Ionization Mass Spectrometry. *Journal of Chromatography A* **2005**, *1081* (2), 225–231.

(176) Rogers, J. D.; Thurman, E. M.; Ferrer, I.; Rosenblum, J. S.; Evans, M. V.; Mouser, P. J.; Ryan, J. N. Degradation of Polyethylene Glycols and Polypropylene Glycols in Microcosms Simulating a Spill of Produced Water in Shallow Groundwater. *Environ. Sci.: Process. Impacts* **2019**, *21* (2), 256–268.

(177) Dulio, V.; Koschorreck, J.; van Bavel, B.; van den Brink, P.; Hollender, J.; Munthe, J.; Schlabach, M.; Aalizadeh, R.; Agerstrand, M.; Ahrens, L.; Allan, I.; Alygizakis, N.; Barcelo, D.; Bohlin-Nizzetto, P.; Boutroup, S.; Brack, W.; Bressy, A.; Christensen, J. H.; Cirka, L.; Covaci, A.; Derksen, A.; Deviller, G.; Dingemans, M. M. L.; Engwall, M.; Fatta-Kassinos, D.; Gago-Ferrero, P.; Hernández, F.; Herzke, D.; Hilscherová, K.; Hollert, H.; Junghans, M.; Kasprzyk-Hordern, B.; Keiter, S.; Kools, S. A. E.; Krueve, A.; Lambropoulou, D.; Lamoree, M.; Leonards, P.; Lopez, B.; López de Alda, M.; Lundy, L.; Makovinská, J.; Marigómez, I.; Martin, J. W.; McHugh, B.; Miège, C.; O'Toole, S.; Perkola, N.; Polesello, S.; Posthuma, L.; Rodriguez-Mozaz, S.; Roessink, I.; Rostkowski, P.; Ruedel, H.; Samanipour, S.; Schulze, T.; Schymanski, E. L.; Sengl, M.; Tarábek, P.; Ten Hulscher, D.; Thomaidis, N.; Togola, A.; Valsecchi, S.; van Leeuwen, S.; von der Ohe, P.; Vorkamp, K.; Vrana, B.; Slobodnik, J. The NORMAN Association and the European Partnership for Chemicals Risk Assessment (PARC): Let's Cooperate! *Environ. Sci. Eur.* **2020**, *32* (1), 100.

(178) Wang, Z.; Hellweg, S. First Steps Toward Sustainable Circular Uses of Chemicals: Advancing the Assessment and Management Paradigm. *ACS Sustainable Chem. Eng.* **2021**, *9* (20), 6939–6951.

UVCBs remain a formidable challenge from multiple perspectives, mostly stemming from their varied compositions and unknown or ambiguous identities. A central problem when dealing with UVCBs is that their chemical representation is not uniform within databases and registries, which in turn makes it difficult to catalogue, communicate, search, and assess them in an efficient way. In this work, the open format *Mixture InChI* was proposed as a solution to the issue of UVCB chemical representation, including demonstrative examples that served as proofs of concept. Furthermore, three main approaches to UVCB assessment were reviewed, namely by whole substance, known constituents, or fractions. Aspects of the challenges they pose in their analytical detection and characterisation were discussed, including acknowledgement that multiple analytical techniques and strategies are imperative. Furthermore, certain mixture toxicity approaches were recommended, and regulatory priorities concerning UVCBs were highlighted. Overall, the notion that there is “no one solution that fits all” UVCB substances emerged, and grouping strategies of similar UVCBs that would allow group-by-group testing and management seem a viable way forward.

Nevertheless, analytical efforts to identify the individual components of UVCBs in environmental mixtures persist, and the potential of cheminformatics and computational approaches to support these pursuits is explored in the next chapter. More specifically, homologous series of chemicals have frequently been detected in the environment using HRMS, but bottlenecks in their data analysis hinder their structure elucidation. The next chapter explores the use of a cheminformatics algorithm to enhance database resources, with the ultimate goal of supporting the structure identification of HRMS signals forming homologous series.



## Chapter 5

# A Cheminformatics Algorithm for Improved Identification of Homologous Series in Environmental Mixtures

Homologous series describe groups of related chemicals that share a common core structure and a monomer that repeats to different degrees. They occur throughout multiple chemistry domains, for example as natural products produced by various organisms, as well as in synthetic chemistry contributing towards chemical product development. Numerous UVCBs are homologous series, for example “Medium chain chlorinated paraffins” (CASRN 85535-85-9) and “Alcohols, C9-C11” (CASRN 64641-46-9).

Multiple homologous series have been frequently and simultaneously detected in environmental samples.<sup>96-98</sup> These compounds are likely emitted through the consumption or use of various domestic chemical products that contain surfactants, such as soaps and detergents. In HRMS data, their signals manifest as characteristic patterns: the elution profile usually has peaks that evoke a normal distribution, and plotting their  $m/z$  against  $t_R$  typically gives a line with a constant linear slope that typically represents the mass of the repeating unit. However, assigning plausible chemical structures to these analytical signals is an ongoing challenge because of their quantity in any given sample, and though database matching by masses can be relatively trivial, reasonable chemical assignment based on the elution profiles may be elusive because chemicals within databases that form homologous series exist without explicit relation to one another. Additionally, potentially hundreds to thousands of plausible matching candidates by mass may exist, which confounds the ability to elucidate the possible interrelationships between them.

In this chapter, a cheminformatics algorithm was developed to classify homologous series within compound databases. The algorithm was openly implemented in Python using the RDKit and applied to chemical datasets from the fields of natural products,

exposomics, and environmental chemistry. Furthermore, it was validated against existing similar approaches of chemical classification such as the categorisation of poly- and perfluorinated alkyl substances (PFAS) that had been performed automatically by cheminformatic means, as well as manually by domain experts. Classified homologous series in environmental chemistry data are foreseen to support identification of these compounds in environmental samples.

## Publication D

### An Algorithm to Classify Homologous Series Within Compound Datasets

Lai, A.<sup>1</sup>, Schaub, J.<sup>2</sup>, Steinbeck, C.<sup>3</sup> & Schymanski, E. L.<sup>4</sup>

DOI: 10.1186/s13321-022-00663-y

Reprinted with permission from Biomed Central Ltd, part of Springer Nature.  
Article is open-access, distributed under CC-BY licence.

<b>Author Contributions</b> (Underlined numbers refer to PhD students)				
<b>Author No.</b>	<b><u>1</u></b>	<b><u>2</u></b>	<b>3</b>	<b>4</b>
Conceptual Research Design	x		x	x
Planning of Research Activities	x		x	x
Reviewing the Tools	x	x		
Data Analysis & Interpretation	x	x		
Manuscript Writing	x	x	x	x
Suggested Publication Equivalence Value	1.0			

RESEARCH

Open Access



# An algorithm to classify homologous series within compound datasets

Adelene Lai<sup>1,2\*</sup> , Jonas Schaub<sup>2</sup> , Christoph Steinbeck<sup>2</sup>  and Emma L. Schymanski<sup>1</sup> 

## Abstract

Homologous series are groups of related compounds that share the same core structure attached to a motif that repeats to different degrees. Compounds forming homologous series are of interest in multiple domains, including natural products, environmental chemistry, and drug design. However, many homologous compounds remain unannotated as such in compound datasets, which poses obstacles to understanding chemical diversity and their analytical identification via database matching. To overcome these challenges, an algorithm to detect homologous series within compound datasets was developed and implemented using the RDKit. The algorithm takes a list of molecules as SMILES strings and a monomer (i.e., repeating unit) encoded as SMARTS as its main inputs. In an iterative process, substructure matching of repeating units, molecule fragmentation, and core detection lead to homologous series classification through grouping of identical cores. Three open compound datasets from environmental chemistry (NORMAN Suspect List Exchange, NORMAN-SLE), exposomics (PubChemLite for Exposomics), and natural products (the COllECTION of Open NatUral producTs, COCONUT) were subject to homologous series classification using the algorithm. Over 2000, 12,000, and 5000 series with CH<sub>2</sub> repeating units were classified in the NORMAN-SLE, PubChemLite, and COCONUT respectively. Validation of classified series was performed using published homologous series and structure categories, including a comparison with a similar existing method for categorising PFAS compounds. The OngLai algorithm and its implementation for classifying homologues are openly available at: <https://github.com/adelenelai/onglai-classify-homologues>.

**Keywords:** RDKit, Fragmentation, Algorithm, Scaffolds, Homologous series, Polymers, Environmental chemistry, Natural products, Exposomics, Pattern recognition

## Introduction

Homologous series are groups of compounds that share the same core structure with varying attached repeating chemical subunits. These structurally-related compounds occur in many areas of chemistry and can be represented by Markush structures [1], as in the patent literature, or as general molecular formulae, for example C<sub>n</sub>F<sub>2n+1</sub>SO<sub>3</sub>H (Fig. 1). In drug design, homologation is used as a molecular modification strategy to construct series for lead

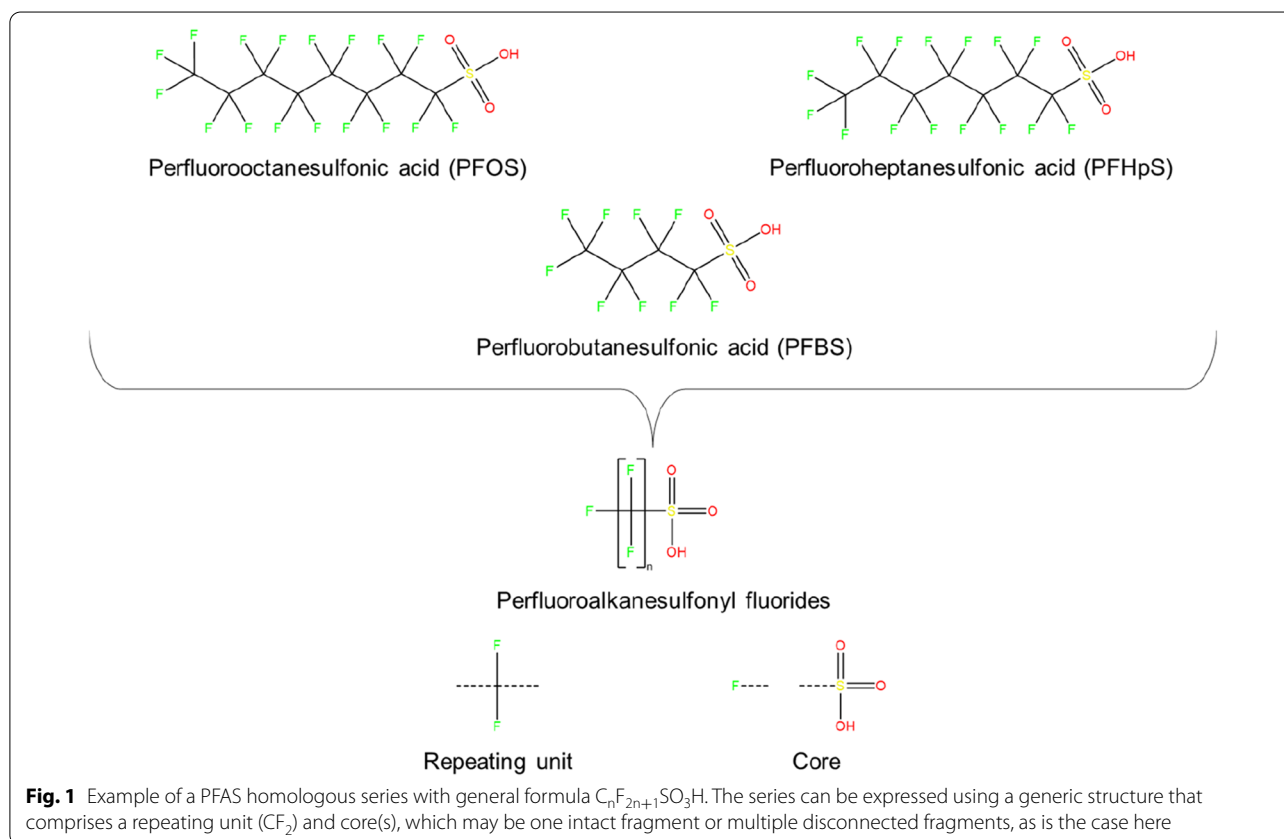
optimisation [2], while homologous series are prominent in pesticide synthesis [3], food [4], and material science [5], as well as formulation chemistry [6] for applications in myriad products such as cosmetics, surfactants, and pharmaceuticals. In nature, homologous series occur as natural products of multiple organisms including bacteria [7], fungi [8], marine sponges [9, 10], birds [11], bees [12], and avocados [13]. In the environment, synthetic compounds consisting of homologous series are considered anthropogenic pollutants, for example, surfactants that have been identified extensively in wastewater [14–17], and are classified as High Production Volume chemicals because of their widespread production and use. Other classes of environmental chemical pollutants containing

\*Correspondence: adelene.lai@uni.lu

<sup>1</sup> Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 Avenue du Swing, 4367 Belvaux, Luxembourg  
Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.



homologous series include the ‘forever chemicals’ i.e., per- and polyfluoroalkyl substances (PFAS) [18–21], as well as technical mixes of polymers such as chlorinated paraffins [22, 23], both of which have been identified extensively in the environment [24, 25], and can be considered as substances of Unknown or Variable composition, Complex reaction products, or Biological materials (UVCBs) [26].

Within compound datasets, having molecules grouped into homologous series can potentially advance several areas of chemistry, for example their analytical identification using liquid chromatography-high resolution mass spectrometry (LC-HRMS). As the structural similarity of homologous compounds can result in a trend in physicochemical properties, homologous series often exhibit characteristic comb-like elution patterns and constant  $m/z$ -retention time shifts in LC-HRMS data. Such signals are frequently detected in environmental samples, where the constant  $m/z$  difference between signals is indicative of the repeating unit’s mass and, in some cases, identity. Consequently, their identification is of high interest, especially since they form a relatively significant proportion of environmental unknowns [27] (also known as ‘non-target compounds’). Various data-mining routines [28–30] and screening tools [31] have been developed

to address this challenge, which usually involves trying to match spectral features with database entries by mass. However, interpreting the matches to find chemically related identifications i.e., homologous chemical series, remains extremely laborious for two reasons: (1) the sheer number of possible (interconnected) homologues in complex environmental samples, and (2) individual homologous compounds are not linked to each other within databases. Therefore, to address the latter, having homologous compounds classified into series within chemical databases would support environmental chemists in assigning related chemical structure identifications to unknown but likely homologous mass spectral features, series-by-series, where possible. Notably, this advantage extends to chemists seeking to discover novel natural products; if structures of the same homologous series within a combined structural and spectral database are annotated as such, their characteristic spectral similarities and trends can be identified, which could expedite the elucidation of previously unreported members of a given series and hence aid the dereplication of spectral data.

Another area of chemistry that would benefit from classified homologous series in datasets is property prediction. As homologous compounds are structurally

similar, structure–property relationships are typically predictable within a given series such that compounds usually share similar properties or show a trend [32], e.g., the Wiener index to predict the boiling points of alkanes [33], Kováts retention indices in gas chromatography to predict analyte retention relative to alkanes [34], or the effect of varying repeating unit chain length on insecticidal activity [13]. In this way, data gaps in physicochemical properties for homologous compounds can be filled using models based on series members that have property data.

Studies of chemical diversity/activity within a given chemical space may also benefit from homologous series classification; instead of focusing on homologous compounds that share repetitive structures and thus similar properties, focus can instead be refined on areas with interesting and varied properties. In other words, grouping together homologous compounds helps eliminate redundancy in the investigated chemical space, as related compounds can be considered group-wise instead of individually. This capability is likely pertinent to medicinal chemists interested in interrogating chemical spaces for diverse properties, or when developing screening decks [35]. In turn, concise representations of a particular chemical space or screening deck may be desired, which could be achieved by general formulae or Markush structures for homologous compounds.

Despite these potential advantages, most compound datasets do not contain homologous compounds classified into series. Instead, homologous compounds typically exist in databases as individual entities without explicit association to one another. To the human eye, homologous series are easily recognisable because of their structural similarity; especially when dealing with simple series and small numbers of chemical structures (10 s to 100 s), a trained chemist can easily classify homologous series by hand as it is a relatively simple, albeit time-consuming pattern recognition task. However, the sizes of today's compound databases regularly exceed hundreds of structures: as of August 2022, PubChem [36, 37] and ChemSpider [38, 39] contain over 110 million compounds each, while virtual screening libraries used for drug discovery are in the order of billions [40]. Such scale renders manual classification of homologous series impractical. Thus, automated methods using cheminformatic algorithms are needed.

The starting point for automated homologous series classification is the detection of appropriate cores i.e., the common fragment(s) shared by each member of a homologous series. As a series is defined by its core(s), correct core detection by cheminformatic means is as critical as it is challenging. Existing approaches for molecular substructure analysis, in this case to automatically detect

cores suitable for homologous series classification, fall into three main categories. The first and most instinctive approach is to consider potential cores as Maximum Common Substructures (MCS) [41, 42]. However, trying to find multiple possible MCS de novo amongst large sets of molecules (> 10,000) is computationally expensive and would likely require additional clustering post-processing steps to obtain the final homologous series. For this purpose, previous work such as Kruger et al.'s clustering approach for chemical series classification [43] has limited applicability because it would not generate core structures specific enough to determine homologous series correctly. An alternative related approach is to exploit pattern-mining algorithms, as homologous series classification can be considered as a task of frequent subgraph mining or graph-based substructure pattern mining [44]. However, these methods require a priori knowledge of a so-called minimum support value, defined as the percentage of all graphs in which a given subgraph must occur. In other words, users must know and specify as input how many series there should be within a given molecule collection, which is impossible to know upfront for most compound datasets. Alternatively, cores could be derived via graph representations of molecules leading to the generation of molecular frameworks as introduced by Bemis and Murcko [45]. However, a significant caveat therein is the required presence of ring systems, which cannot always be assumed.

To address this gap in automated homologous series classification, a free and open algorithm to detect homologous series within compound datasets was developed, which to the best of our knowledge, is the first of its kind. The algorithm was implemented in the RDKit as a Python package called OngLai (pronounced 'ong-lye'), and is openly and freely available on GitHub [46] (<https://github.com/adelenelai/onglai-classify-homologues>). (OngLai has a double meaning in Hokkien: literally, pineapple and figuratively, 'fortune is coming'.) The algorithm input includes a user-specified repeating unit, which forms the basis for the detection of cores that define series. The core fragments are detected without a priori knowledge of their structure, nor how many are present within a given dataset. This result is achieved through successive repeating unit substructure matching and molecule fragmentation steps. Identified homologous series are generated as output, with each compound assigned a number indicating series membership. For a given run of the algorithm, series membership is unique for each molecule as there is only one core fragment result possible once all repeating units have been removed. However, a molecule could in theory belong to multiple homologous series if multiple runs of the algorithm are performed

with different settings specified each time, e.g., different repeating unit.

OngLai was used to classify homologous series within three major chemical collections containing compounds from environmental chemistry, exposomics, and natural products. These collections were chosen to highlight the prevalence of homologous compounds in such varied research domains as well as to demonstrate the broad applicability of OngLai. The first of these three collections, the NORMAN Suspect List Exchange (NORMAN-SLE) [47], comprises synthetic chemicals suspected to be present in the environment such as pesticides, pharmaceuticals, surfactants, food-contact chemicals, and those used in industrial applications, like PFAS [48]. The NORMAN-SLE contains 99 so-called ‘suspect’ lists of chemicals hosted by the NORMAN Network, which are used for suspect screening mass spectrometry data generated from measuring environmental samples [47]. The second collection, PubChemLite for Exposomics (PubChemLite), is a subset of PubChem that aims to capture the chemical space relevant for exposomics [49], the study of exposures to chemicals over time. PubChemLite therefore contains chemicals associated with both metabolism and disease (e.g., ‘Biomolecular Interactions and Pathways’, ‘Associated Disorders and Diseases’ etc.), and environmental chemicals (e.g., ‘Agrochemicals’, ‘Drug and Medication Information’ etc.). Finally, the COCCollection of Open Natural prodUcTs (COCONUT) is a compilation of natural product compounds from over 50 open data resources and manually curated datasets from the literature [50, 51]. It is currently the largest open collection of natural products that is freely available online. Natural products consist of compounds produced by organisms such as bacteria, fungi, animals, and plants over the course of various life processes, and because of their potentially high bioactivity, natural products are of great interest for drug discovery. Selected homologous series classified by OngLai in these three collections are reported here.

Additionally, OngLai’s results were validated against published homologous series and PFAS structure categories from the 2018 OECD PFAS definition [52]. The latter is of particular interest to regulatory stakeholders, as PFAS categorisation remains a high-priority task in effort to catalogue and assess the environmental risks of these compounds. A comparison of OngLai to split-PFAS [53], an automated method based on SMARTS [54] matching developed to support PFAS categorisation efforts, was also performed. Previously, PFAS had been manually classified by experts for the 2018 OECD definition to provide common terminology for stakeholders to communicate, research, and regulate these compounds given their widespread uses and potential adverse

environmental and health effects. With an ever-growing number of PFAS compound registrations and detections in environmental samples, these so-called ‘forever chemicals’ and their categorisation remain of high priority to various stakeholders interested in their future registration, use, and regulation.

## Methods

### Algorithm and implementation

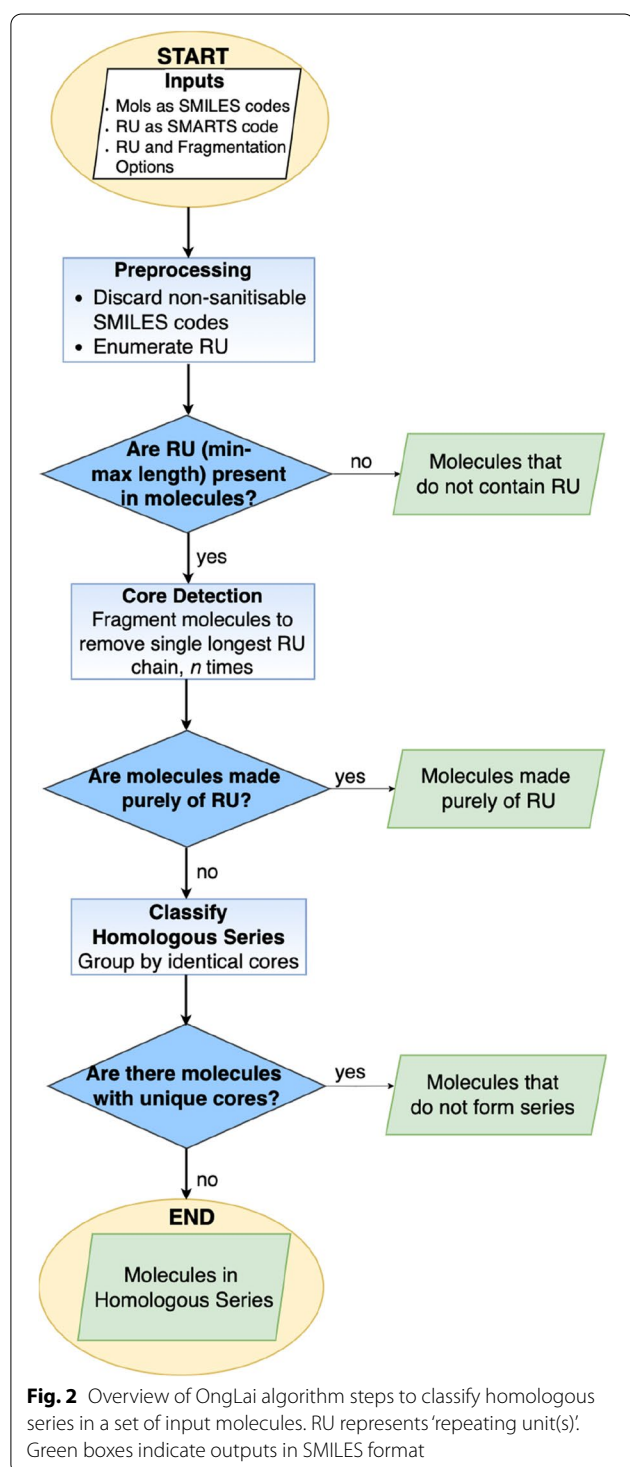
OngLai was developed and implemented using the RDKit (RDKit version 2021.09.4 [55, 56] and Python version 3.7 [57]) and is openly and freely available on GitHub (<https://github.com/adelenelai/onglai-classify-homologues>). OngLai is designed to be run in the command line; more information is available in the GitHub README file.

Within a set of input molecules given as SMILES strings, OngLai detects homologous series by first detecting cores. It does this by substructure matching chains of user-specified repeating units, then fragmenting the molecules a specified number of times to remove these chains. Molecules with the same remaining core fragments are then grouped together into what is considered a homologous series. The sequence of the algorithm’s steps is provided in Fig. 2 and described in more detail below.

OngLai requires two main inputs: the first is a CSV file with a minimum of two columns containing SMILES representations and molecule names (column names can be specified in the command line according to the dataset used; additional columns will be ignored). In a pre-processing step (Fig. 2), the SMILES codes are parsed and checked for validity i.e., whether they can be converted into sanitised molecule objects within the RDKit. Unparseable SMILES strings are discarded. Molecular sanitisation is a RDKit concept that ensures molecules are ‘reasonable’ i.e., can be represented by Lewis structures with complete octets, and that properties such as ring membership and hybridisation can be calculated for each atom [58].

The second input is a repeating unit of choice, expressed as a SMARTS string. For example, the repeating unit of a series of homologous molecules defined by a growing alkyl chain would be  $-CH_2-$ , represented as ‘[#6&H2]’ in SMARTS. The definition of a suitable repeating unit is crucial because it determines which cores, and therefore which homologous series, will be detected. Importantly, the starting and terminal atoms of this repeating unit SMARTS string should have open valences such that it is chemically feasible to create a linear chain by concatenating the SMARTS (Fig. 2, ‘Pre-processing’). Thus, the repeating unit SMARTS strings must be defined from connection point to connection





point. Example repeating unit SMARTS inputs are provided in Table 1.

Once pre-processing is complete, repeating unit chains are enumerated according to the first of two user-customisable settings: the minimum and maximum lengths

of the repeating unit chains (Table 2). This setting ultimately determines whether repeating units are considered present or absent in the input molecules (Fig. 2, first dark blue rhombus); the default minimum length of 3 is recommended to avoid detections of trivially short repeating unit chains that likely occur frequently in many molecules. Each of the enumerated repeating unit chains is searched within each molecule as potential substructure matches. The result (*HasSubstructMatch* = 1 or 0) is recorded as an element within a NumPy array, one array per input molecule. If the sum of the array elements is equal to zero, the molecule does not contain at least 1 repeating unit chain of the specified minimum length and is then eliminated from further analyses (Fig. 2, first green box). Having established that the remaining molecules contain repeating units, OngLai proceeds with core detection via molecule fragmentation to separate repeating unit chains from core structures. The default setting for the number of molecule fragmentation steps is 2 (Table 2) but can be customised if more than two repeating unit chains are expected to be present in the input molecules. The accuracy of core detection and homologous series classification would technically be unaffected by setting a higher number of fragmentation steps than is actually needed, albeit at the expense of longer computation times. Each time during fragmentation, only one—the longest—repeating unit chain is detected, then removed to ensure 'clean' core detection without leftover repeating unit fragments. Importantly, only one repeating unit chain is removed per fragmentation step, even in the case of symmetrical molecules or molecules that otherwise have multiple identical longest repeating unit matches (see Fig. 8 in "Discussion" for further details).

Molecule fragmentation is achieved using RDKit's *ReplaceCore* function, which introduces a dummy atom at each fragmentation site that is then replaced with a hydrogen atom. However, if the remaining molecule object for a given molecule is empty, it means the input molecule is made purely of repeating units and is reported as such (Fig. 2, second green box). Otherwise, the remaining fragment(s) is considered the core, which can consist of a single or multiple disconnected fragments.

In a final step, molecules are classified into homologous series; those with identical cores (same number and identity of fragments) are deemed members of the same series. Molecules with unique cores, i.e., cores that occur only once in the entire dataset, are considered 'molecules that do not form series' (Fig. 2, third green box). In this way, the results of the OngLai are entirely dataset-dependent, as input molecules and consequently their resulting cores are necessarily compared to each other in the homologous series detection process, meaning

**Table 1** Example repeating units and their SMARTS representations that are suitable for input to OngLai. The default repeating unit is alkyl (CH<sub>2</sub>)

Repeating unit pseudo-SMILES	Repeating unit chemical name	SMARTS (OngLai input)
CH <sub>2</sub>	Alkyl	[#6&H2]
CH <sub>2</sub> CH <sub>2</sub> O	Ethoxy	[#8]-[#6&H2]-[#6&H2]
CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> O	Propoxy	[#8]-[#6&H2]-[#6&H2]-[#6&H2]
CF <sub>2</sub>	Perfluoroalkyl	[#6](-[#9])(-[#9])
CF <sub>2</sub> O	Perfluorinated methyl ether	[#8]-[#6](-[#9])(-[#9])
CF <sub>2</sub> CF <sub>2</sub> O	Perfluorinated ethyl ether	[#8]-[#6](-[#9])(-[#9])-[#6](-[#9])(-[#9])
CH <sub>2</sub> C(CH <sub>3</sub> )=CCH <sub>2</sub>	Isoprene	[#6&H2]-[#6](-[#6&H3])=[#6]-[#6&H2]

**Table 2** User-customisable settings of OngLai to specify 'repeating unit options' in the command line

Setting	Format	Default
Minimum and maximum lengths of repeating unit chains	Integer	Min. = 3 Max. = 30
No. fragmentation steps	Integer	2

the co-presence or absence of possible series members determines series classification. A comparison of cores for equality is performed using sanitised canonical RDKit SMILES representations.

A CSV file is generated as output containing the following columns: 'SMILES' (only those sanitisable by RDKit), 'Name', and 'series\_no'. Series membership is encoded in the 'series\_no' field, as are the other aforementioned results (Fig. 2, green boxes) as shown in Table 3. Additionally, an overview of the classification results is provided as output, written to a TXT file called 'classification-results'.

### Datasets

OngLai was applied to three different datasets, NORMAN-SLE [47, 48] used in environmental chemistry, PubChemLite [49, 59] used in exposomics/metabolomics, and COCONUT [50, 51] in natural products research, respectively. All the datasets are openly available (see Additional file 1 Sect. 1.2, Declarations and References).

The NORMAN-SLE dataset used here is an aggregation of the suspect lists that were compiled by the NORMAN Network from various environmental chemistry researchers around the world. The exact dataset originated from the 'NORMAN Suspect List Exchange Classification' on PubChem's Classification Browser (downloaded 2022-03-21) [60, 61]. Using the PubChem Identifier Exchange Service [62], the molecules in NORMAN-SLE were mapped to their Parent CIDs (Operator Type: 'Parent CID') to remove salts, charged ions, and mixtures. Stereochemical information is preserved in this process if originally present. Conversion of 115,115 input compounds to Parent CIDs resulted in a final dataset of 98,116 'parent' compounds that were downloaded in CSV format via PubChem. The second dataset, PubChemLite for Exposomics (v.1.8.0), contains 392,465 molecules and was downloaded from Zenodo [49, 59] and used as-is. PubChemLite compounds have both neutral (InChIKey second and third blocks: UHFFFAOYSA-N) and non-neutral stereochemistry. During PubChemLite development, the stereochemical-neutral version was preferentially selected if available, otherwise a structure with stereochemistry was included; further details can be found in the original paper [49]. COCONUT, containing 407,270 molecules (v.11/2021 [50, 51]), was downloaded as SMILES (CDK Unique SMILES [63], i.e., representations without stereochemical information) and used as-is. The specific versions of these datasets used are archived on Zenodo [64]. Specific instructions for

**Table 3** Interpretation of 'series\_no' encoding as part of the output from homologous series detection. N+1 is the number of homologous series that were detected by OngLai in a given dataset

Series_no	Interpretation
0-N	Molecules that form homologous series
-1	Molecules with no repeating units matches of minimum chain length specified
-2	Molecules made purely of repeating units
-3	Molecules that have repeating units matches of minimum chain length specified but that do not form series (unique cores)

running the algorithm on these datasets are available in the GitHub README file <https://github.com/adelenelai/onglai-classify-homologues>.

#### Validation and comparison with existing methods

Validation of OngLai was performed in two ways, by comparing the homologous series it classified in NORMAN-SLE with (1) published homologous series, and (2) published structure categories.

Published homologous series are available in two suspect lists from the NORMAN Suspect List Exchange: *S7 EAWAGSURF* [65], and *S23 EIUBASURF* [66], which both contain surfactant compounds with CH<sub>2</sub> and CCO repeating units. Homologous series in these two compound lists are explicitly indicated by 'SurfactantCode' or 'Name' column entries, where members of a given series follow a sequential naming convention e.g., 'C10-LAS', 'C11-LAS', and 'C12-LAS' forming the 'Cx-LAS' series, or 'Amines, coco 10 EO', 'Amines, coco 11 EO', and 'Amines, coco 12 EO' forming the 'Amines, coco x EO' series (x = 10–12 in both examples). Validation was performed by comparing homologous series classified by OngLai in the NORMAN-SLE dataset with those published in these lists that were downloaded and used as-is.

Published 'Structure Categories' determined by experts for the 2018 OECD definition pertain to PFAS compounds containing CF<sub>2</sub> repeating units obtained from the NORMAN-SLE Classification Tree in PubChem under *S25 OECDPFAS* [52]. These lists of compounds were downloaded from PubChem per structure category via the Identifier Exchange Service and mapped to Parent CID as described above. Validation using these 'Structure Categories' proceeded as follows: molecules in a given homologous series classified by OngLai were inspected to see how many structure categories they belonged to, assuming that correctly classified series should have molecules belonging to the same single structure category.

To facilitate validation, a Python script was used to merge OngLai's CSV output (by InChIKey) with (1) the published homologous series and (2) published structure category CSV files respectively. Then, the merged data were manually inspected. The script and all CSV files resulting from this validation analysis are available in the Additional file 1: Sect. 3.

To compare OngLai to an existing method for categorising PFAS compounds called splitPFAS, OngLai was additionally applied to the 770 PFAS listed in the Supplementary Information file of Sha et al. [53] Homologous series with CF<sub>2</sub> repeating units detected by OngLai in NORMAN-SLE were compared with the categorisation results of splitPFAS. In the original paper, 770 PFAS were systematically divided into 4 categories with general formulae C<sub>n</sub>F<sub>2n+1</sub>-X-R: perfluoroalkanoyl (X = CO), sulfonyl

(X = SO<sub>2</sub>), n:1 fluorotelomer (X = CH<sub>2</sub>), and n:2 fluorotelomer (X = CH<sub>2</sub>CH<sub>2</sub>). For comparison purposes, compounds with the same X and same R groups but differing *n* are considered to form homologous series (henceforth referred to 'splitPFAS series'). Python code used to prepare and analyse the splitPFAS dataset and all results from the comparative analysis are available in Sect. 4 of Additional file 1.

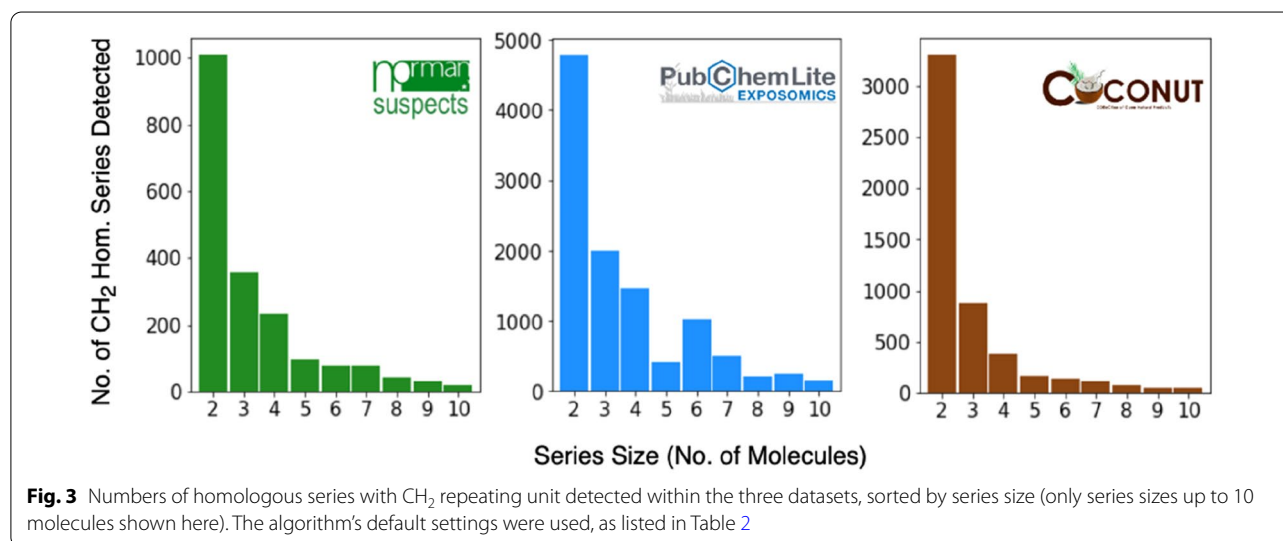
#### Results and discussion

OngLai was applied to 3 different datasets by running the Python script in the command line within a conda environment containing the RDKit. The script and all necessary modules are provided in the OngLai package on GitHub (see <https://github.com/adelenelai/onglai-classify-homologues> for the full list). A compute server with two Intel(R) Xeon(R) Silver 4114 CPUs and 64 GB of RAM was used in single-thread mode. OngLai's default settings (Table 2) were applied, including using '[#6&H2]' corresponding to CH<sub>2</sub> (alkyl) as the repeating unit SMARTS input (Table 1). Detection of homologous series by OngLai in NORMAN-SLE, PubChemLite, and COCONUT datasets using these parameters took approximately 2, 16, and 35 min respectively. Two further runs of the algorithm were performed on each dataset using '[#8]-[#6&H2]-[#6&H2]' and '[#6](-[#9])(-[#9])' as repeating unit SMARTS input, corresponding to CCO (ethoxy) and CF<sub>2</sub> (perfluoroalkyl) respectively; for validation, the homologous series detected in the NORMAN-SLE dataset were compared to the published lists as described above. Additionally, OngLai was also run on the 770 PFAS compounds used in the splitPFAS study for comparison.

This section is divided into two parts. First, an overview of the homologous series with CH<sub>2</sub> repeating units classified in the three datasets is provided, including an interpretation of OngLai's outputs, validation of the CH<sub>2</sub>, CCO and CF<sub>2</sub> series classified in NORMAN-SLE, and comparison with splitPFAS. Then, the second part focuses on the implementation and behaviour of OngLai's underlying algorithm, demonstrated in detail using selected examples of classified homologous series.

#### Homologous series classified in NORMAN-SLE, PubChemLite, and COCONUT

Thousands of homologous series with CH<sub>2</sub> repeating units were detected by OngLai: in total, 2098 in NORMAN-SLE, 12,105 in PubChemLite, and 5329 in COCONUT respectively. These series were detected using the default settings of the algorithm (Table 2). The size distributions of the homologous series classified are shown in Fig. 3, while Table 4 provides a summary of the overall results. Complete series classification results are available



**Fig. 3** Numbers of homologous series with CH<sub>2</sub> repeating unit detected within the three datasets, sorted by series size (only series sizes up to 10 molecules shown here). The algorithm's default settings were used, as listed in Table 2

**Table 4** Summary statistics of detected homologous series with CH<sub>2</sub> repeating units in the three datasets. The algorithm's default settings were used, as listed in Table 2. Full details and results are available in Additional file 1: Sect. 2

	NORMAN-SLE (n = 98,116)	PubChemLite (n = 392,465)	COCONUT (n = 407,270)
No. of homologous series detected	2098	12,105	5329
No. of molecules classified as members of homologous series	8775	82,476	18,528
No. of molecules consisting purely of CH <sub>2</sub> repeating units	0	0	0
No. of molecules containing CH <sub>2</sub> repeating units but not forming homologous series (unique cores)	10,778	35,111	36,864
No of molecules not containing CH <sub>2</sub> repeating units	78,559	274,861	351,527
No of molecules discarded from analysis (failed sanitation)	4	17	351

in Sect. 2 of the Additional file 1. Notably, most series detected comprise only 2 molecules, similar to chemical series classified within drug discovery projects [68]. Overall, there are more small series than there are large series, as evident in the series size distributions (Fig. 3), which may imply a high chemical diversity in the respective databases.

The proportion of molecules that were deemed members of CH<sub>2</sub> homologous series given the default settings used were 9% for NORMAN-SLE, 21% for PubChemLite, and 5% for COCONUT (Table 4). Approximately 10% of each dataset consists of molecules that contain CH<sub>2</sub> repeating units, but do not form homologous series, meaning the detected cores are unique within the respective dataset. The majority (70–86%) of all molecules in each dataset do not contain CH<sub>2</sub> repeating unit chains of minimum length 3 repeating units (Table 2, default algorithm setting), i.e., there were no substructure matches found in those molecules using the following SMARTS query: '[#6&H2]-[#6&H2]-[#6&H2]'; representing the

structure 'CH<sub>2</sub>CH<sub>2</sub>CH<sub>2</sub>' in pseudo-SMILES. Overall, less than 5% of molecules were discarded from the analysis because they were either not parseable by the RDKit due to valence model violations e.g., pentavalent carbons, or the SMILES strings were invalid (reported to the respective data maintainers).

Notably, zero molecules consisting purely of CH<sub>2</sub> repeating units were detected across the three datasets. Instinctively, one would think alkanes such as propane, butane, and pentane fall into this category, but they do not because the terminal carbon atoms in these alkanes are bonded to three H atoms and not exactly two, as specified in the SMARTS representing CH<sub>2</sub> repeating units (Table 1, '[#6&H2]'). Therefore, alkanes are considered to form their own homologous series by OngLai, with the terminal carbon atoms ultimately forming the core ('H<sub>3</sub>C. CH<sub>3</sub>' in pseudo SMILES). This result highlights how the specificity of the SMARTS repeating unit definition directly determines the homologous series



classified, which is further discussed in “Effect of repeating unit SMARTS specification on homologous series classified”.

Details of the CCO and CF<sub>2</sub> homologous series detected in the three datasets are available in Additional file 1: Sect. 2. Notably, 64 molecules in COCONUT were classified into 23 homologous series with CF<sub>2</sub> repeating units. These molecules do not appear to be natural products and should be removed in future curation exercises of natural product space. As these molecules have been classified into series, entire series of these non-natural-product-like molecules can be removed together instead of having to search and remove these molecules on an individual basis. These findings have been reported to the COCONUT database maintainers [69].

#### Validation of classified series

The validation of homologous series classified in NORMAN-SLE was performed in two ways: (1) by comparing classified series with published homologous series, and (2) by inspecting their homologous compound membership within published structure categories. All validation results described below are available in Sect. 3.3 of Additional file 1.

#### Validation with published homologous series

As shown in Table 5, the majority of CH<sub>2</sub> and CCO homologous series detected in NORMAN-SLE were in overall agreement with published homologous series in S7 and S23 (62%, 60%, 80%, 64% ‘Full Match’ respectively). Partial or mixed classifications arose due to various factors such as suboptimal algorithm settings for that particular series of molecules (e.g., the minimum repeating unit chain length of 3 was too long), or differences in stereochemistry specificity across molecules that would otherwise belong to the same series within

NORMAN-SLE. Less than 5% of homologous series were not identified by OngLai across all repeating units and published homologous series because of either of the two aforementioned factors. An example of published homologous molecules that were not classified by OngLai is the ‘Cx, sorbitan monoester, 20 EO’ series. This series is listed in S23 *EIUBASURF* as having two molecules (x=12 and 18). In the NORMAN-SLE dataset however, the C<sub>12</sub> species has no stereochemistry specified, but the C<sub>18</sub> species does, thus causing them to have different cores detected, resulting in the series not being classified by OngLai (Fig. 4; further discussion on stereochemistry below). In this sense, OngLai provides a more specific classification of homologous series than what is listed and indicated by the Name field in S23 *EIUBASURF*, as it distinguishes between levels of stereochemical information specificity that were not captured by the naming convention used in S23 *EIUBASURF*.

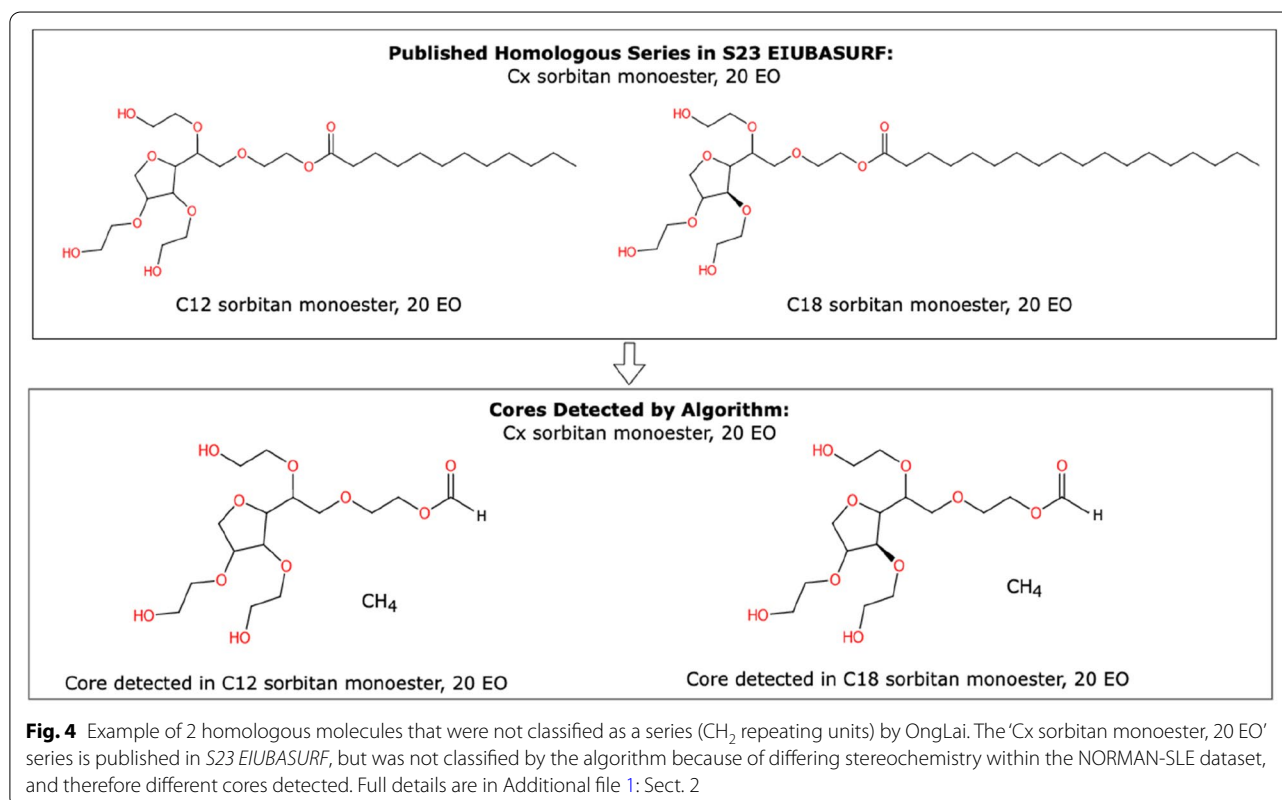
Importantly, validation using published homologous series in the S7 and S23 datasets was possible because of the naming convention used by the datasets’ curators. For example, in these datasets, compounds with the names C9-LAS, C10-LAS, C11-LAS, and C12-LAS clearly belong to the Cx-LAS series. The fact that homologous compounds in these datasets can be recognised just from their names without any inspection of their chemical structures supports the use of these lists as independent sources of information ideal for homologous series validation.

#### Validation with published structure categories

Similar results were obtained in the validation of classified homologous series with CF<sub>2</sub> repeating units using the OECD’s PFAS Structure Categories: 50% of the 600 homologous series detected contain molecules that belong to the same single Structure Category within the

**Table 5** Validation by comparing homologous series in NORMAN-SLE classified OngLai with published homologous series containing CH<sub>2</sub> and CCO repeating units. Series in S7 and S23 were manually compared to OngLai results. Full Match indicates a 1:1 relationship between published series and series classified by OngLai. Homologous series from NORMAN-SLE containing molecules that are not in the published homologous series list or vice versa, but that otherwise match, are also considered Full Matches (‘or as available’). Partial or Mixed Classification indicates either a 1:n relationship between published homologous series and homologous series classified by the algorithm, or that certain molecules were not classified together with the others in a given published series. Full details in Additional file 1: Sect. 3.3

List containing published homologous series	Repeating unit	No. of published homologous series				
		Full match (or as available)	Partial or mixed classification	Not classified by OngLai	Present in list, absent in NORMAN-SLE	Total
S7 EAWAGSURF	CH <sub>2</sub>	8	5	0	0	13
	CCO	6	4	0	0	10
S23 EIUBASURF	CH <sub>2</sub>	105	17	6	4	132
	CCO	62	35	0	0	97



respective series (Table 6). The remainder corresponds to homologous series containing molecules belonging to more than one Structure Category (10% of all series classified), no Structure Category (22.5%), or a mixture thereof (17.5%) within the same series.

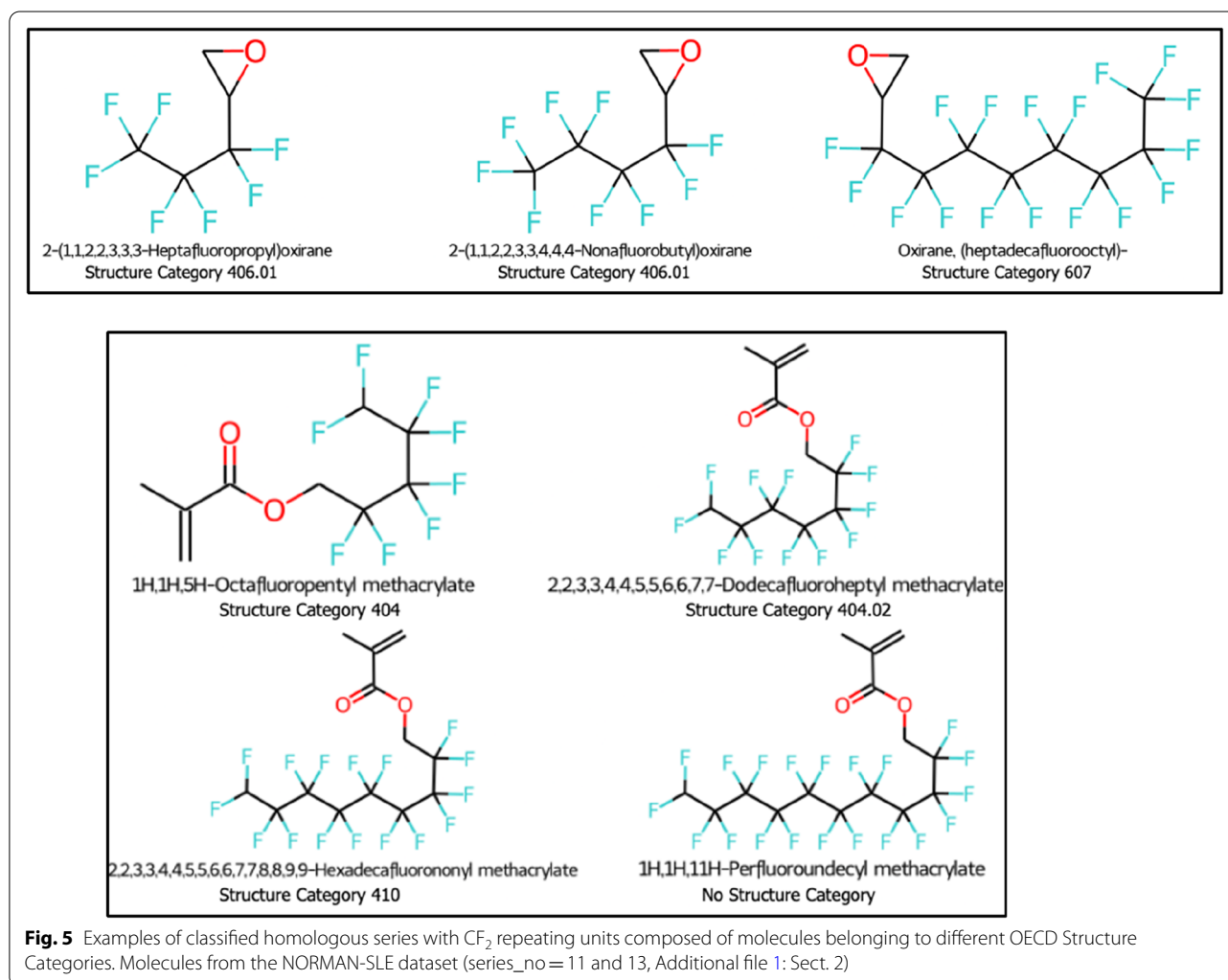
Two examples of molecules grouped into the same series having different OECD Structure Categories are shown in Fig. 5. The molecules in the first series (Fig. 5, top panel) belong to two different Structure Categories:

**Table 6** Comparison of published Structure Categories for PFAS compounds containing  $\text{CF}_2$  repeating units with homologous series classified by OngLai in the NORMAN-SLE dataset. Structure categories are published in the 2018 OECD PFAS report [52, 67]

	Series with 1 structure category	Series with > 1 structure category	Series with no structure category	Series with combination of no structure category and $\geq 1$ structure category	Total series classified by OngLai
No. of $\text{CF}_2$ homologous series	301	59	135	105	600

Category 406.01 corresponding to fluorotelomer epoxides ( $\text{C}_n\text{F}_{2n+1}\text{I} + \text{CH}_2 = \text{CHCH}_2\text{OH} \rightarrow \text{C}_n\text{F}_{2n+1}\text{-CH}_2\text{CH}(\text{I})\text{CH}_2\text{OH} \rightarrow \text{C}_n\text{F}_{2n+1}\text{-CH}_2(\text{CHCH}_2\text{O})$ ); and Category 607 corresponding to perfluoroalkyl epoxides & derivatives ( $\text{C}_n\text{F}_{2n+1}$ -epoxides). Another example (Fig. 5, bottom panel) has molecules in the same classified series that do not belong to any Structure Category *and* a combination of Structure Categories 404–n:1 fluorotelomer-based non-polymers ( $\text{C}_n\text{F}_{2n+1}\text{-CH}_2\text{-R}$ ); 404.02–n:1 FT (meth)acrylate ( $\text{CH}_2\text{-OC}(=\text{O})\text{CH}=\text{CH}_2$ ); and 410–n:1 FT (meth)acrylate ( $\text{CH}_2\text{-OC}(=\text{O})\text{CH}=\text{CH}_2$ ). The last molecule in the series does not belong to any OECD Structure Category because it is absent from the original *S25 OECDPFAS* list, but was present in the NORMAN-SLE because it originated from other lists (e.g., *S46* and *S71*) that make up the PFAS within NORMAN-SLE.

These mixed results are attributable to the broader definitions of Structure Categories compared to homologous series; the former often contain a mixture of homologous and non-homologous molecules. Per the 2018 OECD definition, a Structure Category can represent various properties, such as sharing a common general formula, varying functional groups, and/or being derivatives of the same compound e.g., 'category 101: perfluoroalkyl carbonyl halides ( $\text{C}_n\text{F}_{2n+1}\text{-C}(=\text{O})\text{R}$ ,  $\text{R}=\text{F}/\text{Cl}/\text{Br}/\text{I}$ )' and 'category 202: perfluoroalkane sulfonic acids (PFASs),



their salts and esters ( $\text{R}=\text{OH}$ ,  $\text{ONa}$ ,  $\text{OCH}_3$ , etc.): These relatively broader categories likely reflect some of the challenges of assigning Structure Categories to numerous PFAS in a manual fashion, as was done for the 2018 OECD PFAS definition. As manual assignment is prone to typographical errors, wrong assignments, or inconsistent assignments, cheminformatic-based tools for automated assignment of Structure Categories are highly desirable and warranted [53, 70, 71].

Overall, as approximately 50% of  $\text{CF}_2$  series classified by OngLai in the NORMAN-SLE dataset contain molecules belonging to the same OECD Structure Category, there appears to be reasonable consistency in the 2018 OECD manual categorisation of PFAS. Given that the homologous series classified by OngLai have stricter definitions in terms of chemical structure similarity, OngLai's results could support or inform future OECD efforts to subcategorise PFAS.

#### Comparison with existing method for categorising PFAS: splitPFAS

OngLai was applied using the same compute server described above to the 770 PFAS compounds that were

**Table 7** Summary statistics of detected homologous series with  $\text{CF}_2$  repeating units in the splitPFAS dataset

	splitPFAS dataset (n = 770)
No. of series detected	132
No. of molecules classified as members of homologous series	540
No. of molecules consisting purely of $\text{CF}_2$ repeating units	0
No. of molecules containing $\text{CF}_2$ repeating units but not forming homologous series	196
No. of molecules not containing $\text{CF}_2$ repeating units	34
No. of molecules discarded from analysis (failed sanitation)	0



originally categorised by splitPFAS. In approximately 2 min, 132 homologous series with  $\text{CF}_2$  repeating units were classified (Table 7). These results were compared with those of the splitPFAS tool (XLSX file in Supplementary Information of Sha et al. [53]). For comparison here, molecules in a given PFAS category out of the four outlined by Sha et al. that share identical R groups are assumed to be homologous series because they have the same general formula (same X and R groups in  $\text{C}_n\text{F}_{2n+1}\text{-X-R}$ ). These series will henceforth be referred to as 'splitPFAS series'. There were 124 of such splitPFAS series found in Sha et al.'s work; OngLai detected 132 homologous series (full details in Sect. 4 of Additional file 1).

Comparison of the series classified by OngLai and splitPFAS series generally shows good agreement between the two methods in terms of their matching results. However, there are some differences in the number of series and composition of certain series which can partly be attributed to the fact that some PFAS were not categorised by splitPFAS, but were classified as homologous series by OngLai. The reason for this result is because within splitPFAS outputs, no X groups were detected for these molecules by splitPFAS. Consequently, in the results XLSX file, these molecules have 'NA' in their 'SplitSMARTS (X)' column, attributed to 'No splittable bond found for the input molecule'. Associated error codes provided as splitPFAS output explain the various underlying reasons, for example '1—the perfluoroalkyl chain was branched or cyclic', or '4—the R group was a single F atom'. There were 11 homologous series classified by the algorithm containing such molecules (examples in Fig. 6).

Another reason for the difference in the results produced by splitPFAS and OngLai is that some PFAS do not actually conform to the general formula  $\text{C}_n\text{F}_{2n+1}\text{-X-R}$  prescribed by Sha et al. For example, all the molecules shown in Fig. 7 have the same X groups and R groups in the general formula prescribed by Sha et al. ( $\text{C}_n\text{F}_{2n+1}\text{-X-R}$ ), as indicated in the splitPFAS results (XLSX file, Fluorotelomer tab), where  $\text{X}=[\text{CH}_2]$  and  $\text{R}=\text{CC}(=\text{C})\text{C}(=\text{O})\text{O}$  (methacrylic acid). Therefore, they technically belong to the same splitPFAS series according to the assumption made for this comparison exercise. Evidently however, the molecules in the top panel of Fig. 7 actually have the general formula  $\text{C}_n\text{F}_{2n}\text{-X-R}$  because the terminal carbon is bonded to two fluorine atoms and one hydrogen atom instead of three fluorine atoms, as in the bottom panel. In this case, OngLai distinguished this fact; the core detected for the series in the top panel of Fig. 7 is methacrylate, while that for the series in the bottom panel consists of two disconnected fragments: methacrylate and a fluorine atom. As shown in this example, OngLai was able to distinguish and thus group different PFAS

into homologous series with higher granularity than splitPFAS.

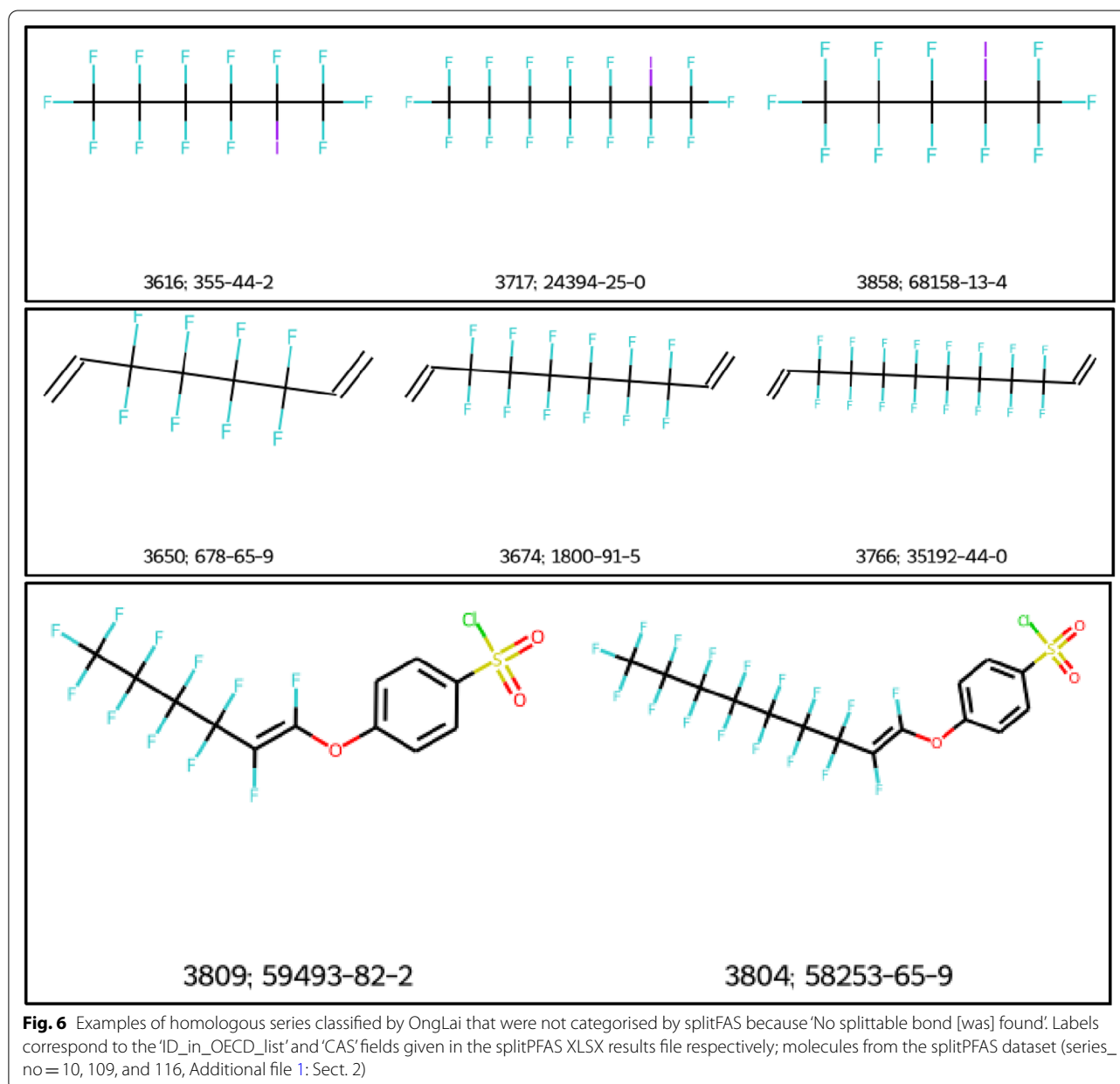
Overall, the categorisation results of splitPFAS are very similar to the results of the presented homologous series classification algorithm (full results available in Additional file 1: Sect. 4). This outcome indicates that the assumption made for the purpose of this comparison—that compounds having the same X and R groups in the general formula  $\text{C}_n\text{F}_{2n+1}\text{-X-R}$  are indeed homologous—was reasonable. However, in some cases, OngLai demonstrated more flexibility in handling different PFAS structures than splitPFAS because the latter has more hard-coded elements in its cheminformatics processing of input structures than OngLai does. For example, splitPFAS has specific SMARTS corresponding to the 4 PFAS categories specified, which likely explains why no splittable bonds could be detected in some cases. That said, it is important to bear in mind that splitPFAS was designed with a different intention than OngLai; splitPFAS is not dedicated to homologous series classification, therefore it cannot be directly compared. Nevertheless, this comparison shows that OngLai could be used to support PFAS categorisation efforts by e.g., providing further subcategorisation.

#### Implementation of OngLai

In this section, important features of the OngLai algorithm and its implementation, independent of the datasets it is applied to, are discussed using demonstrative examples of  $\text{CH}_2$  series classified across NORMAN-SLE, PubChemLite, and COCONUT.

#### Molecular fragmentation—removing one substructure match at a time

In cheminformatics, removing one substructure match at a time instead of multiple simultaneously in a given molecule is not a trivial task, yet here, it is crucial for preserving the accuracy of the core detected and thus correct classification of homologous series. In the RDKit, the most intuitive choice to achieve substructure removal is *DeleteSubstructs*, but this function removes all repeating units matched at a time in one go, which is undesirable. Therefore, *ReplaceCore* is used instead and shown in comparison to *DeleteSubstructs* in Fig. 8. To date, the RDKit community has explored two further alternatives to remove one substructure at a time [72], but these methods are not suitable here because (1) there is no way to remove entire substructures from *RWMol* objects, only atoms and bonds, and (2) encoding the substructure to be removed as a chemical reaction is impractical, as a new Reaction SMARTS query would have to be encoded for each input molecule depending on its specific structure. In this sense, *ReplaceCore*, typically used

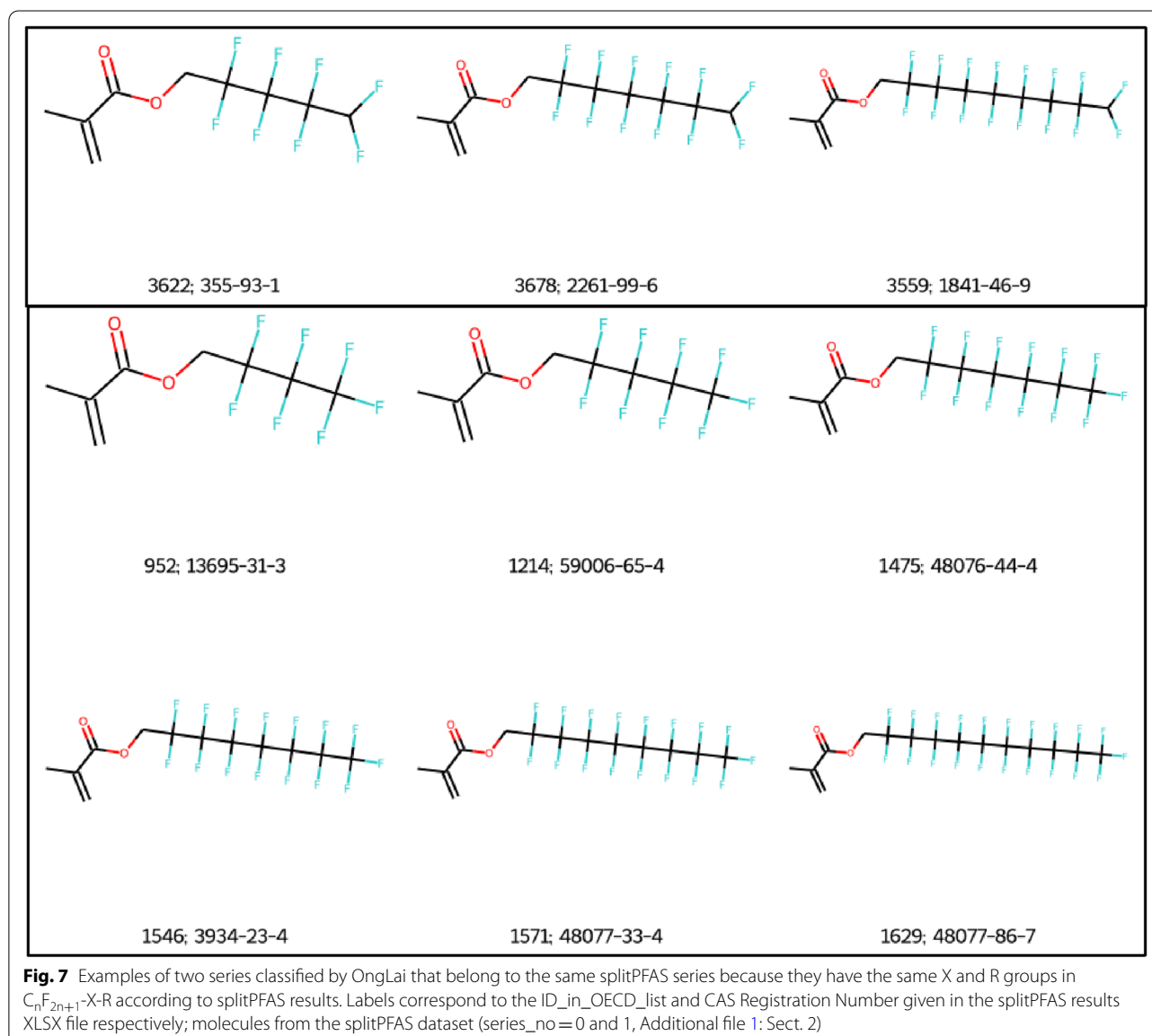


for common cheminformatic tasks like R-group decomposition or constructing Structure–Activity Relationship tables, was applied here in a novel and perhaps unorthodox, but effective manner to remove substructures.

#### Effect of repeating unit SMARTS specification on homologous series classified

As described in a previous example in this section, the repeating unit SMARTS definition directly influences the homologous series classified, for example, by explicitly defining the exact number of connected hydrogen atoms. Other properties of atoms defined in the SMARTS string

also play an important role: in the default repeating unit SMARTS used, '[#6&H2]'; the carbon atom is bonded to exactly two hydrogen atoms, regardless of that carbon's ring membership. Therefore, repeating units forming rings would also be positive matches just like repeating units in linear chains, as shown in Fig. 9, where the CH<sub>2</sub> moieties in the pyrrolidine ring of 1-(4-bromobutyl)pyrrolidine hydrobromide, in addition to those in the linear chain, matched the repeating units SMARTS '[#6&H2]'. Thus, these matches were subsequently removed during molecule fragmentation in the core detection process. The resultant core common to all these three molecules



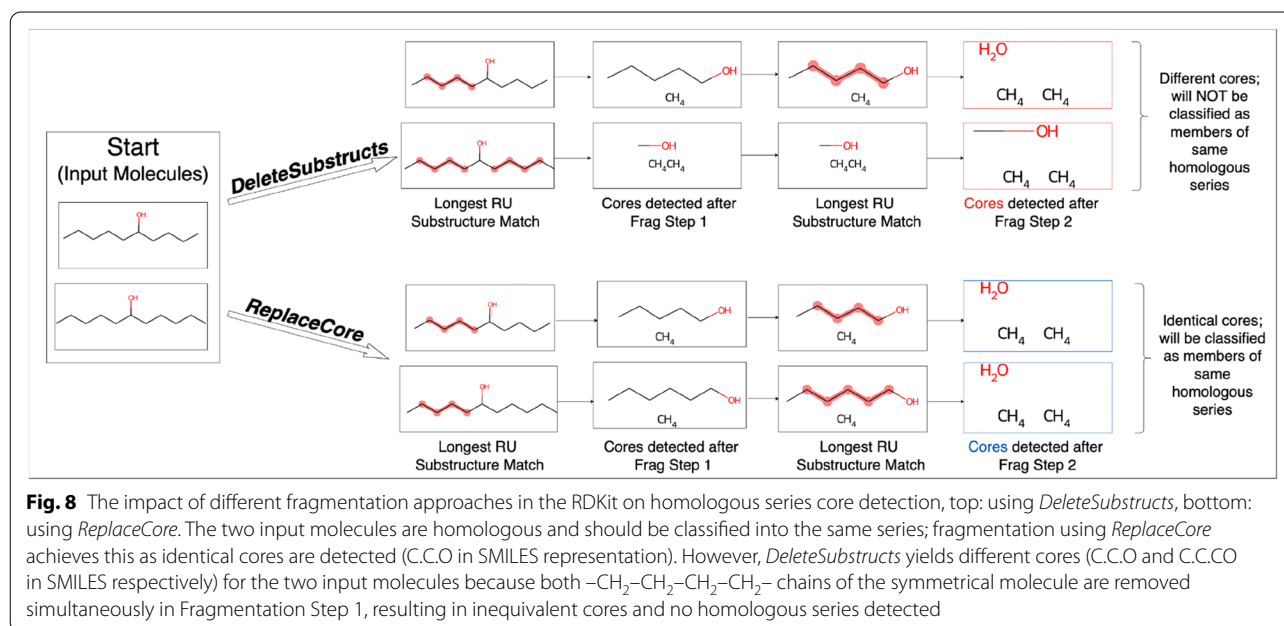
thus consists of two disconnected atoms, one bromine and one nitrogen ('Br.N' in SMILES).

However, if a more specific repeating unit SMARTS query specifying ring membership is used, the first two molecules could be distinguished from 1-(4-bromobutyl)pyrrolidine. Using the repeating unit '[#6;!R&H2]' (carbon atom that is not a member of a ring bonded to exactly two hydrogen atoms) yields two different cores for the three molecules in Fig. 9: while the core detected for the first two molecules remains the same as before, that for 1-(4-bromobutyl)pyrrolidine consists of the intact pyrrolidine ring and a single Br atom, represented in SMILES as 'Br.C1CCN(C1)'. Thus, 1-(4-bromobutyl)pyrrolidine would not be included in the same homologous series as

the first two molecules in Fig. 9 which underscores the importance of repeating units SMARTS specification in the resulting homologous series classified. In other words, users should be careful when specifying their repeating units SMARTS to achieve the desired results.

#### Effect of maximum length of repeating unit chains specified

The maximum length of repeating unit chains to be enumerated for substructure matching and removal is user-customisable, with the default value set to 30 repeating units (Table 2). This default value was used in the present analysis to avoid prolonged computation times that result from having a larger maximum value. It was also assumed



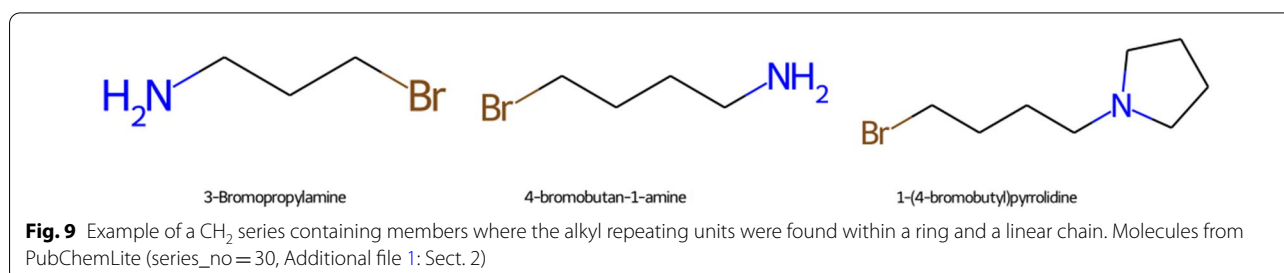
that this would be sufficient to cover all possible cases of repeating units in the molecules analysed. This assumption held true for the NORMAN-SLE and PubChemLite datasets, but not COCONUT, where some molecules were misclassified due to this default value (see Fig. 10).

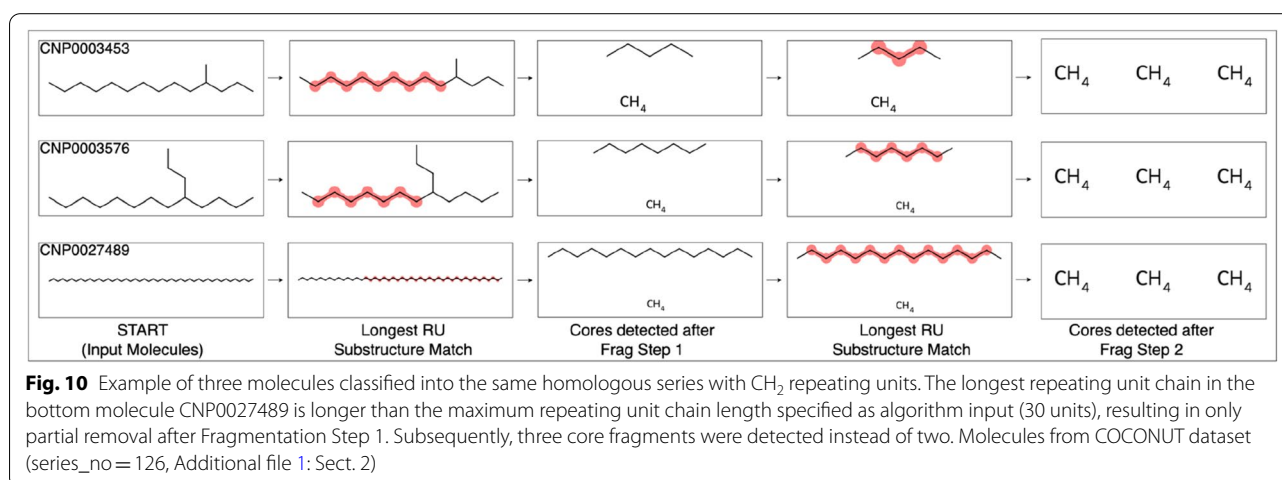
In the classified homologous series shown, the linear alkane CNP0027489 (molecular formula  $\text{C}_{46}\text{H}_{94}$ ) should have been classified together with other linear alkanes having core (' $\text{CH}_3$ .  $\text{H}_3\text{C}'$  in pseudo SMILES). However, because the longest repeating unit chain in CNP0027489 is  $\text{C}_{44}\text{H}_{88}$  (corresponding to a maximum repeating unit length of 44) and not  $\text{C}_{30}\text{H}_{60}$  (a maximum repeating unit length of 30), the resulting core after two fragmentation steps contains three  $\text{CH}_3$  fragments instead of two, causing it to be classified together with branched alkanes having the same core. In this case, correct classification would be achieved if the maximum value was set to 44 or higher, albeit at the expense of significantly longer computational times.

#### Effect of number of fragmentation steps

The 'No. Fragmentation Steps' setting (Table 2) affects the extent of fragmentation of the input molecule and as a result, the cores detected. Therefore, the cores detected can vary in structure depending on the number of fragmentation steps specified, especially in cases where (1) there are multiple repeating unit chains within a given molecule, (2) the repeating unit chains are of different lengths, and/or (3) the repeating unit chains are bonded to the same atom.

Figure 11 shows the impact of varying the number of fragmentation steps on three input molecules belonging to the same homologous series 'Cx-SPADCs', published in *S7 EAWAGSURF*. Starting with the input molecules in the left-most column, had 'No. Fragmentation Steps' been set to 1, the final cores detected would have been those shown in the red boxes. However, as none of these cores are identical to each other, these three molecules would not be classified into the same homologous series.



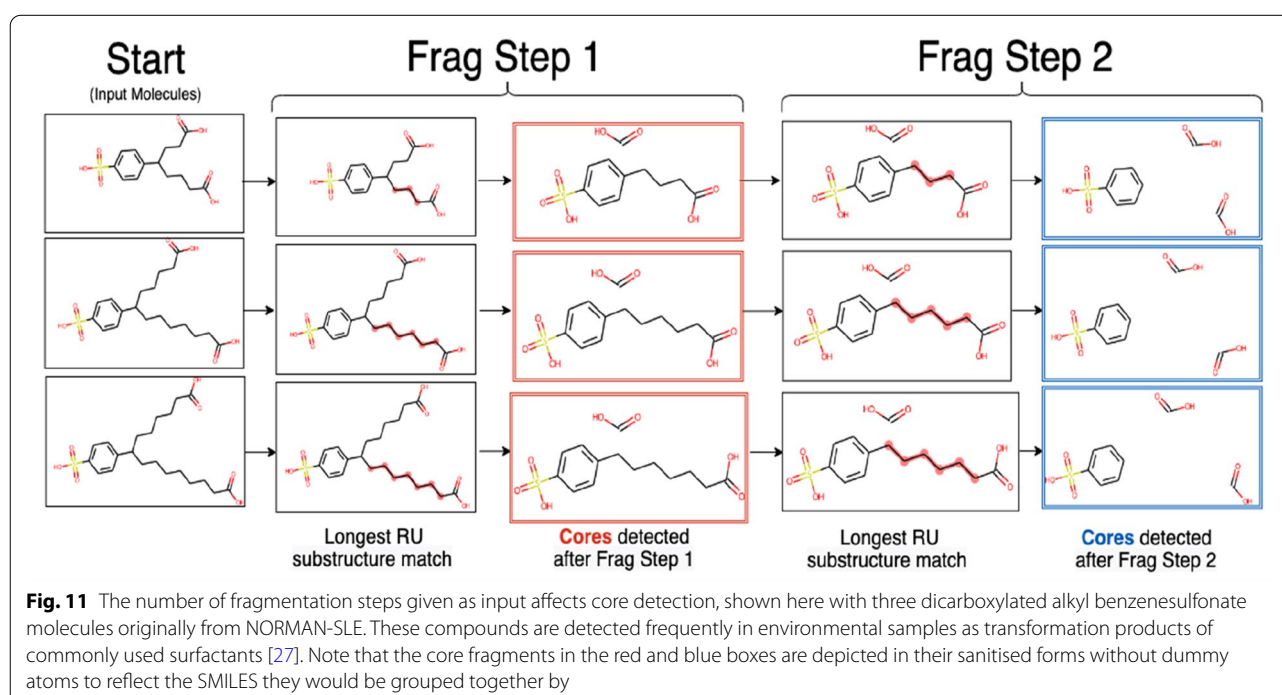


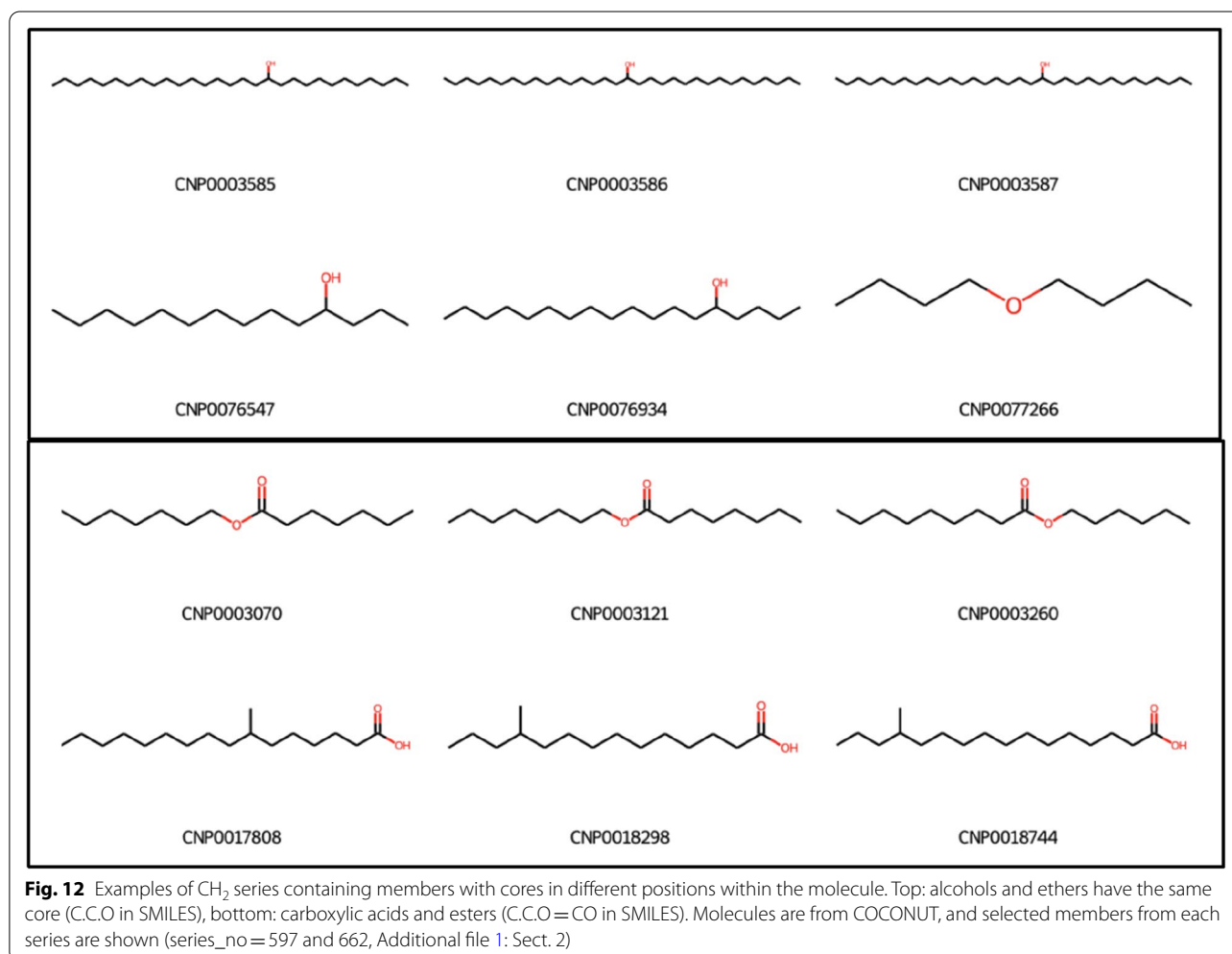
In contrast, a second fragmentation step yields identical cores for the three input molecules (Fig. 11 blue boxes) that would result in the three input molecules being grouped together into the same series. Thus, the number of fragmentation steps selected is crucial for appropriate core detection and homologous series classification.

#### Effect of sanitisation on core detection

The position of core fragment(s) within input molecules is irrelevant for OngLai. In other words, molecules containing the same core fragments, albeit in different positions within the molecule relative to the repeating units,

are classified into the same homologous series. Concrete examples are shown in Fig. 12, where molecules containing either alcohol or ether functional groups are considered homologous (Fig. 12, top panel). A second example shows molecules containing either a carboxylic acid or ester moiety belonging to the same classified series (Fig. 12, bottom panel). Here, whether the core is in a terminal or central position within the molecule is not considered in core detection because its atomic neighbourhood is not taken into account. Consequently, the number of repeating unit chains attached to the core is also not considered, meaning the core could be attached





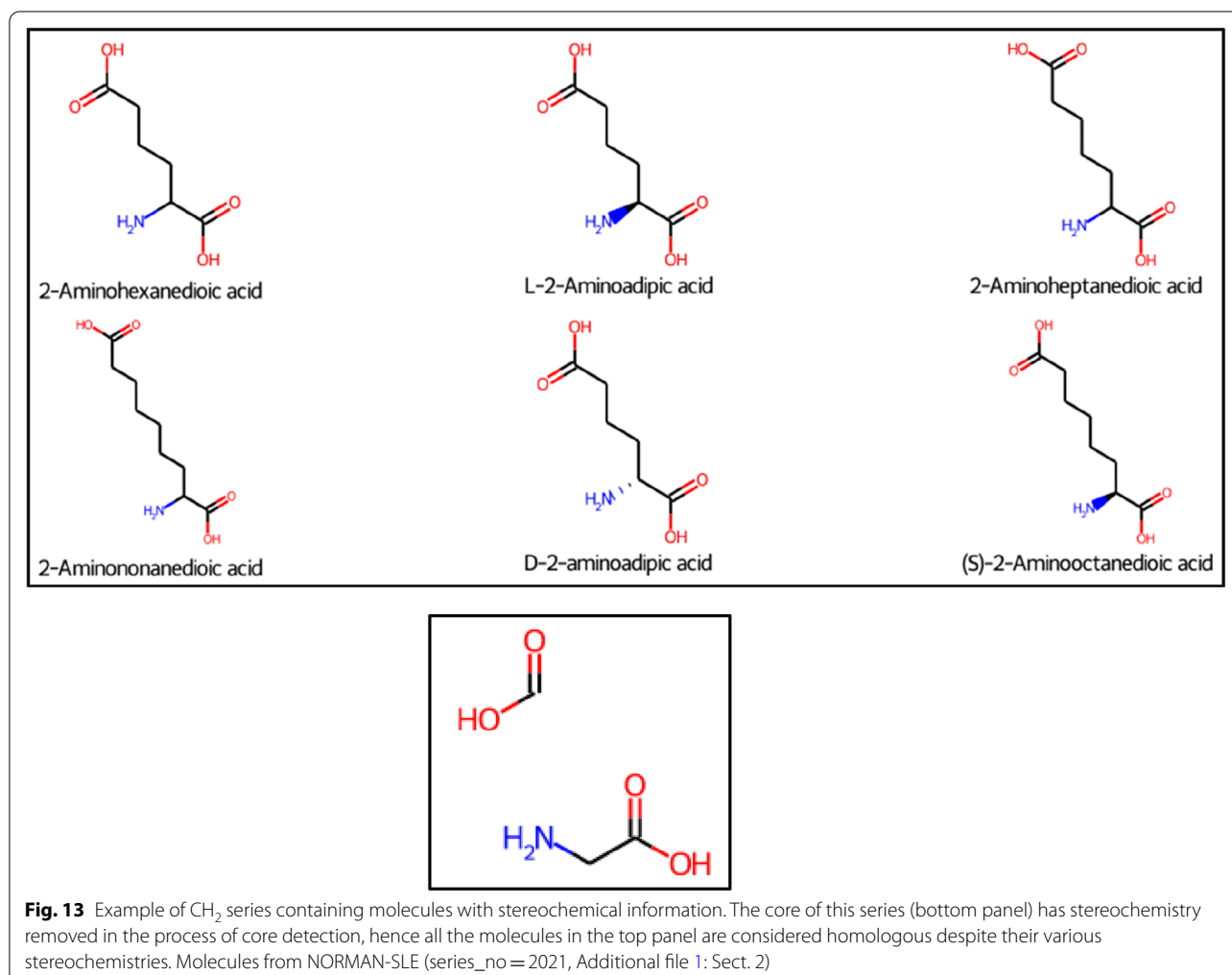
to carbons of varying connectivity degrees across the different members of a homologous series. For example, the 'O' fragment in the ether core of molecule [CNP0077266](#) is attached to two primary carbon atoms (Fig. 12, top panel), while the 'O' fragments in the other molecules of the same series shown are attached to one secondary carbon atom each. Depending on user preference, grouping together molecules with varying core fragment position in the same homologous series may be desirable, but it is possible that future augmentations of OngLai could address the consideration of the number of repeating unit chains attached to the core, or atomic neighbourhood of the core in general.

#### Effect of stereochemical information

Stereochemical information can play a discriminatory role in homologous series detection, depending on where it is specified relative to the core fragment(s) and molecular fragmentation site(s). If bonds with no

stereochemistry specified connecting repeating units and core fragments are fragmented, but stereochemical information is present elsewhere in the molecule, the latter is preserved and taken into consideration during the process of homologous series detection via grouping molecules with identical cores. For example, as shown in Fig. 4, the 'C<sub>18</sub> sorbitan monoester' input molecule contains a bond pointing outwards, as does its core. However, the 'C<sub>12</sub> sorbitan monoester' and its core have planar bonds throughout, so the C<sub>12</sub> and C<sub>18</sub> species are not considered homologous by OngLai. In contrast, the molecules in Fig. 13 are classified as homologous despite their different stereochemistries, because the amino acid core fragment common to all 6 molecules (Fig. 13, bottom panel) was originally adjacent to the fragmented bond and therefore experienced stereochemistry neutralisation in the process of core detection (addition of dummy atom, then conversion to hydrogen atom). Thus, molecules with different stereochemistries may be





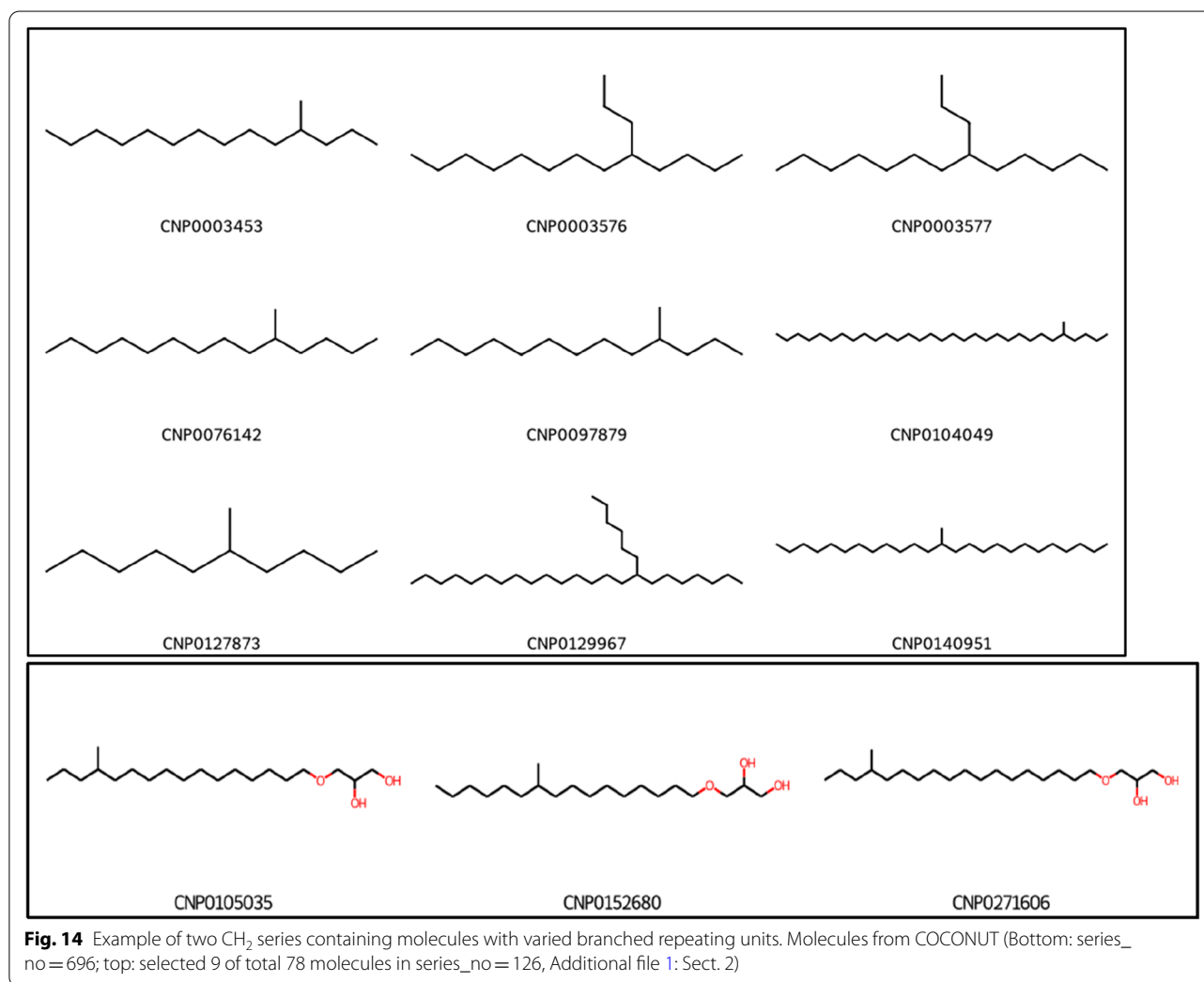
grouped into the same series if fragmentation happens on bonds or adjacent bonds that originally have stereochemistry specified, as this information is removed during core detection. This behaviour is desirable in the specific case of annotating databases to support the identification of chemicals in environmental samples using mass spectrometry (which was the original motivation of OngLai), where stereochemistry differences are less relevant for compound identification. By grouping together all homologous compounds regardless of their stereochemistry differences, the remaining ‘unannotated’ chemical space that should be considered for unknown identification would be smaller, which could make unknown identification easier and more efficient. Overall, however, the desirability of this behaviour would depend on the individual user’s ultimate goal and intended application of classifying homologous series.

Regarding stereochemistry in the datasets used relative to their preparation as described in “Methods”, only

the molecules in COCONUT have no stereochemistry encoded, whereas molecules in NORMAN-SLE and PubChemLite have mixed stereochemical information availability. To investigate the influence of stereochemistry on homologous series detected further, future efforts could include applying OngLai to the version of COCONUT containing all stereoisomers.

#### Molecules with branched repeating units classified as series

Molecules with branched repeating units, irrespective of the length of the branch and branching site, are classified into the same series since OngLai does not consider the atomic neighbourhood of the matched repeating units it removes during core detection (Fig. 14). Rather, it simply detects the longest repeating unit chain and removes it in the process of series classification. In certain applications, this insensitivity could be advantageous, for example when characterising chemical space or preparing



diversity decks in high-throughput chemical screening [35, 73], as grouping together such highly similar analogues could result in reduced redundancy and better representation of the molecules within a given chemical series. However, it is also possible that this insensitivity to the site and extent of branching could be addressed in future augmentations of the algorithm by e.g., introducing filters for molecules that have repeating unit chains of the same lengths.

#### Structural isomers classified as series

As explained above, the atomic neighbourhood of repeating units is not considered when repeating units are being matched for substructure removal in core detection. Thus, being insensitive to atomic neighbourhoods results in ring substitution isomers (meta-, para-, and ortho-) being classified as members of the same series, irrespective of the attachment position of the repeating unit chain (Fig. 15). If desired, such occurrences could be

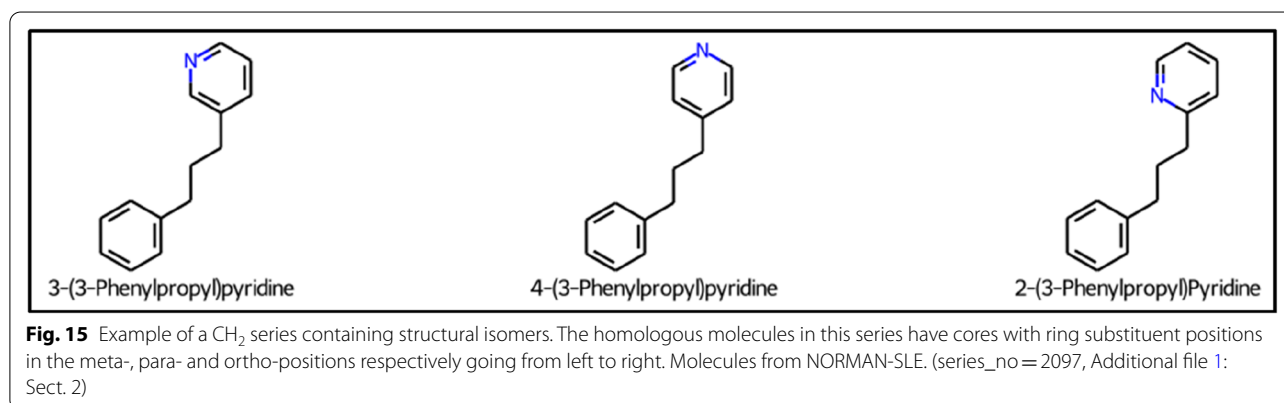
identified and filtered or grouped together on the basis of formula or mass in a post-processing step.

#### Future work

The present work introduces OngLai, an algorithm to classify homologous series within compound datasets. Since this topic has been relatively unexplored, three areas of further research could be interesting to pursue based on the work presented here. Additionally, integration of this homologous series classification functionality into existing tools such as the 'Contrib' directory of the RDKit and the R package 'patRoon' [74] to further enhance the utility of this algorithm have already been discussed with the relevant software maintainers.

#### Algorithm

Consideration of the atomic neighbourhood of the core fragment(s) during core detection is a potential feature to implement in the next version of OngLai. As highlighted



in the “Discussion” section, doing so could improve the accuracy of core detection and thereby generate homologous series containing molecules that have less variability with respect to branching, structural isomerism, and position of the core in the molecules. Atomic neighbourhood consideration could be achieved by attaching R-groups onto repeating unit chains at the fragmentation site, then integrating this information when grouping identical cores together in the final step of homologous series classification.

#### Further results analysis

Additional automated analyses can be performed with homologous series structures after their classification in a given dataset. A first functionality could be to order the series by the number of chains and the number of repeating units in their chains within one identified homologous series. Alternatively, homologous series could be grouped together based on multiple characteristics or properties such as having the same type of repeating units or similar core fragments, e.g., homologous series with core fragments that represent different ortho/meta/para variants of the same structure could be grouped. At higher levels, classified series could be grouped according to similarities between their core or repeating units, based on a defined similarity measure. Basically, any known chemical clustering algorithm can be applied to representative structures of each homologous series group here. This grouping and ordering for different characteristics at different levels can result in a homologous series hierarchy for the given dataset, similar to a scaffold tree [75], which could allow for an intuitive, multi-layer visualisation of homologous series diversity in a given dataset. In terms of mass spectral data processing, specific groups of homologues of interest could also be used either as potential suspect lists or database files during non-target LC-HRMS data processing.

#### Alternative cheminformatics approaches to classify homologous series

Currently, repeating unit structures have to be provided as algorithm input in the form of SMARTS, which requires a priori knowledge of the identity of repeating units and familiarity with SMARTS syntax. On one hand, this requirement makes OngLai highly suited to its original intended application, which is to aid in the identification of unknown but related features in mass spectra. In this case, repeating units are typically known from the outset, as their structures can be deduced from the constant  $m/z$  differences between HRMS features (e.g.,  $m/z = 14.0157$  difference between features likely indicates that the repeating unit is CH<sub>2</sub>). However, from a pure cheminformatics perspective, homologous series classification should ideally be achievable without prior knowledge of repeating unit identity. Developing and implementing such an approach poses a complex but relevant problem, which could be addressed using maximum common substructure (MCS) detection functionality [41, 42] in an all-versus-all approach. That said, the necessity to determine the MCS of every molecule with every other molecule in the given dataset is potentially problematic due to the exponential scaling of required computation that is exacerbated when dealing with large chemical structures like polymers or certain natural products. Common cheminformatics methods like pre-screening and filtering repeating unit-less molecules to overcome these time-consuming MCS functionalities could be explored. Alternatively, parallelisation would be applicable here because the MCS of one molecule pair can be determined separately from the other pairs.

Another idea to approach the problem of homologous series detection is to employ spherical substructures of molecules, also called atomic environments, as used in molecular signatures [76], Morgan fingerprints [77, 78], or HOSE codes [79]. The first step would be to generate

spherical substructures of different heights for every atom in a molecule, where a substructure of height 0 contains only the centre atom itself, the substructure of height 1 contains the centre atom and its direct neighbours, etc. For each height, the number of unique spherical substructures can be tracked. If there is a repeating unit in the molecular structure, there should be a detectable minimum in the diversity of a molecule's spherical substructures for the height equal to the size of the repeating unit. This approach would have the advantage that it is dataset-independent, unlike the current or MCS approach, but would require many specific rules or heuristics for corner cases and a very fine tuning of the parameters for the detection of the assumed height that matches the repeating unit size, if a generally applicable parameter set can be identified at all.

A less complex application of spherical substructure approaches might also be used to detect repeating unit chains with an a priori definition of the repeating substructure that is searched for, as in this work. Instead of SMARTS-based matching as used here, spherical substructures of a matching height for one molecule would be generated and matched with the pre-defined repeating units to identify homologous compounds by their chains. The set height of the included atom neighbours could then be gradually increased to include the neighbouring repeating units until the structure no longer fits the predefined repeating unit structure. This way, a repeating unit chain could be detected directly as a coherent substructure. A disadvantage of the approach would be that spherical substructure notations like HOSE codes are more complex to define by hand and provide less options than SMARTS definitions, since they were not originally developed for substructure matching.

Beyond the classical methods of structural cheminformatics, further alternative approaches could employ machine learning. For example, one could define the problem as a classification task by training a model to recognise homologous vs. non-homologous molecules based on their SMILES strings or even structure depictions. In both data structures, repetitive repeating unit patterns should be detectable in a straightforward manner. A more complex alternative would be to extract the core and (in a generalised model) repeating unit structures, e.g., as SMILES strings. Current successes in similar applications are encouraging [80] but available training data would be a limiting factor, as the numbers of homologous structures detected in relevant datasets reported above and of published homologous series e.g., in specialised databases, appear too low for most machine learning tasks. However, defining core structures with chain attachment points and multiple repeating units structures may allow training data to be

synthetically generated through recombination and enumeration to form diverse homologous series structures.

## Conclusions

OngLai is an open source algorithm implemented in RDKit that classifies homologous series within compound datasets based on two inputs: a CSV file containing compound SMILES representations and a repeating unit represented by a SMARTS string. Using the SMARTS definition of the repeating unit, OngLai first detects suitable cores by molecule fragmentation prior to series classification. Homologous series classification was demonstrated by applying OngLai to three open datasets: NORMAN-SLE, PubChemLite for Exposomics, and COCONUT. Thousands of homologous series with CH<sub>2</sub> repeating units were detected within these datasets using the default algorithm settings. The results were validated using published homologous series and structure categories for surfactant and PFAS examples, and compared with the splitPFAS method for categorising PFAS. Both validation and comparison generally showed good agreement, with OngLai proving to be more granular in its detection of homologous series in some scenarios.

Overall, homologous series classification bears several advantages. Firstly, the detection of homologous series in datasets such as NORMAN-SLE and PubChemLite may support their identification using (LC-)HRMS. Homologous mass spectral features are frequently detected at high intensities in environmental samples and may form a large proportion of measured features that typically remain unknown (but are suspected to be compounds in chemical consumer products that are heavily produced and used, like surfactants). OngLai's results could support the characterisation of these unknowns by providing researchers with classified homologous series within datasets, so they can perform more effective database matching of homologous features detected in their samples in a group-wise fashion. Such steps would contribute to tackling the problem of identifying and characterising UVCBs in the environment and further our understanding of the effects of chemical exposure and its impacts on the environment and/or disease, with the ultimate goal of protecting human health and the environment [26].

Secondly, the characterisation of chemical spaces is enhanced by identifying similar or related compounds that could be considered as a group. As OngLai essentially performs a type of clustering by grouping together homologous compounds, applying it to large screening datasets is a viable method for analysing large chemical spaces of interest and supporting the design of diverse molecule screening decks, which are of interest in drug discovery [70, 71]. An additional benefit accrued from chemical space characterisation via homologous series

detection is that classified series can contribute to more efficient dataset curation, as mentioned with respect to polyfluorinated compounds found in the COCONUT dataset.

OngLai is freely and openly available on <https://github.com/adelenelai/onglai-classify-homologues>. Users are invited to apply OngLai on chemical datasets of interest, possibly as a first data exploration step, to uncover homologous compounds in their datasets, which may lead to insights on potential chemical groups, open new avenues for property prediction, and/or facilitate analytical detection. OngLai can also be used as a means for chemical grouping or to validate existing approaches, which may be of particular interest to e.g., regulatory stakeholders in environmental chemistry [81].

#### Abbreviations

COCONUT: COllECTION of Open Natural ProDUcTs; HOSE: Hierarchically ordered spherical description of environment; LC-HRMS: Liquid chromatography-high resolution mass spectrometry; MCS: Maximum common substructure; NORMAN-SLE: NORMAN Suspect List Exchange; PFAS: Per- and Polyfluoroalkyl substances; PubChemLite: PubChemLite for Exposomics; SMILES: Simplified Molecular Input Line Entry System; SMARTS: SMILES ARbitrary Target Specification; UVCB: Substances of Unknown or Variable composition, Complex reaction products, or Biological materials.

#### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13321-022-00663-y>.

**Additional file 1.** File containing links to code, datasets, and complete results described in the manuscript.

#### Acknowledgements

The authors gratefully acknowledge Steffen Neumann for helpful discussions especially during project conception; Charles Tapley-Hoyt for contributions to Python packaging; Kohulan Rajan and Mahnoor Zulfiqar for assistance with the compute server; Anjana Elapavalore for code testing; the developers and community of RDKit (in particular Greg Landrum and Paolo Tosco) and datamol; and Zhanyun Wang, Christos Nicolaou, and Maximilian Beckers for helpful discussions. We also thank the two anonymous reviewers for their helpful comments.

#### Author contributions

A.L. developed the methodology; wrote, tested, and applied the software; performed the data curation, analysis and validation; wrote, edited, and reviewed the manuscript; and prepared all tables and Figs. 2 to 14. J.S. developed the methodology; wrote, edited, and reviewed the manuscript; and prepared Fig. 1. C.S. conceptualised the project; provided resources; edited and reviewed the manuscript; provided supervision, project administration; and acquired funding. E.L.S. conceptualised the project; provided resources; edited and reviewed the manuscript; provided supervision, project administration; and acquired funding. All authors reviewed, read and approved the final manuscript.

#### Authors' Information

A.L. is a Cotutelle (dual) doctoral candidate in both the research groups of E.L.S. (Environmental Cheminformatics at the Luxembourg Centre for Systems Biomedicine, University of Luxembourg) and C.S. (Cheminformatics and Computational Metabolomics group at the Friedrich Schiller University in Jena, Germany). In her research, she applies cheminformatics and chemical data science to address problems in environmental chemistry, ranging from

environmental monitoring to database curation and chemicals management. J.S. is a doctoral candidate in the Cheminformatics and Computational Metabolomics research group of C.S. at the Friedrich Schiller University in Jena, Germany. His research focuses on cheminformatics, natural products, chemical spaces, open software development, and rule-based algorithms for the extraction of specific substructures from molecular structures (in silico fragmentation). C.S. is a Professor for Analytical Chemistry, Cheminformatics and Chemometrics as well as Vice President for Digitalisation at the Friedrich Schiller University in Jena, Germany. The Steinbeck group's research is dedicated to computational natural products research, the elucidation of metabolomes by means of computer-assisted structure elucidation and other prediction methods, the application of artificial intelligence, in particular, deep-learning methods, as well as algorithm development in cheminformatics. E.L.S. is Associate Professor and head of the Environmental Cheminformatics (ECI) group at the Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg. Her research combines cheminformatics and computational (high resolution) mass spectrometry approaches to elucidate the unknowns in complex samples and relate these to environmental causes of disease. She is involved in and organizes several European and worldwide activities to improve the exchange of data, information and ideas between scientists, including NORMAN-SLE, MassBank, MetFrag and PubChemLite for Exposomics.

#### Funding

Open Access funding enabled and organized by Projekt DEAL. AL and ELS acknowledge funding from the Luxembourg National Research Fund (FNR) for Project A18/BM/12341006. JS and CS acknowledge funding from the Carl-Zeiss-Foundation.

#### Availability of data and materials

OngLai homologue detection algorithm source code: Apache 2.0 Licence; <https://github.com/adelenelai/onglai-classify-homologues>, Software Requirements: Python 3.7 or higher, RDKit v2021.09.4 or higher, datamol v.0.7.3 or higher. Specific versions of the datasets used (NORMAN-SLE, PubChemLite for Exposomics, and COCONUT), as well as complete results, Python scripts and supporting files are freely available within the Supplementary Information archive on Zenodo: <https://doi.org/10.5281/zenodo.7035020>. The Additional file (.doc) contains details of the above archive.

#### Declarations

#### Competing interests

The authors declare that they do not have any competing interests.

#### Author details

<sup>1</sup>Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 Avenue du Swing, 4367 Belvaux, Luxembourg. <sup>2</sup>Institute for Inorganic and Analytical Chemistry, Friedrich Schiller University Jena, Lessing Strasse 8, 07743 Jena, Germany.

Received: 31 August 2022 Accepted: 27 November 2022

Published online: 13 December 2022

#### References

1. Markush EA (1924) Pyrazolone Dye and Process of Making the Same. USA101506316, August 26, 1924. <https://pdfpiw.uspto.gov/piw?PageNum=USA101506316&docid=01506316&IDKey=83E682D73B35&HomeUrl=http%3A%2F%2Fpatft.uspto.gov%2Fnetacgi%2Ffnph-Parser%3FSection=DPTO%2526Sect2%3DHITOFF%2526p%3D1%2526u%3D%2Fnetacgi%2FPTO%2Fsrchnum.html%2526r%3D1%2526f%3DG%2526l%3D50%2526d%3DPALL%2526s1%3D1506316.PN.%2526OS%3D%2526RS%3D>. Accessed 25 Mar 2022
2. Lima LM, Alves MA, Amaral DN (2019) Homologation: a versatile molecular modification strategy to drug discovery. *Curr Top Med Chem*. 19:1734–1750. <https://doi.org/10.2174/1568026619666190808145235>
3. Niemczak M, Rzemieniecki T, Sobiech Ł, Skrzypczak G, Praczyk T, Pernak J (2019) Influence of the alkyl chain length on the physicochemical



- properties and biological activity in a homologous series of dichloroprop-based herbicidal ionic liquids. *J Mol Liq* 276:431–440. <https://doi.org/10.1016/j.molliq.2018.12.013>
- Zhu J-P, Liang M-Y, Ma Y-R, White LV, Banwell MG, Teng Y, Lan P (2022) Enzymatic synthesis of an homologous series of long- and very long-chain sucrose esters and evaluation of their emulsifying and biological properties. *Food Hydrocoll* 124:107149. <https://doi.org/10.1016/j.foodhyd.2021.107149>
  - Wolf SE, Liu T, Govind S, Zhao H, Huang G, Zhang A, Wu Y, Chin J, Cheng K, Salami-Ranjbaran E, Gao F, Gao G, Jin Y, Pu Y, Toledo TG, Ablajan K, Walsh PJ, Fakhraai Z (2021) Design of a homologous series of molecular glassformers. *J Chem Phys* 155(22):224503. <https://doi.org/10.1063/5.0066410>
  - Samarkina DA, Gabdrakhmanov DR, Lukashenko SS, Nizameev IR, Kadirov MK, Zakharova LY (2019) Homologous series of amphiphiles bearing imidazolium head group complexation with bovine serum albumin. *J Mol Liq* 275:232–240. <https://doi.org/10.1016/j.molliq.2018.11.082>
  - Carballeira NM, Miranda C, Lozano CM, Nechev JT, Ivanova A, Stefanov K, Ilieva M, Tzvetkova I (2001) Characterization of novel methyl-branched chain fatty acids from a halophilic bacillus species. *J Nat Prod* 64(2):256–259. <https://doi.org/10.1021/np000494d>
  - Schlingmann G, Roll DM (2007) Homolog separation, a necessity for the proper identification of fungal metabolites. *J Chromatogr A* 1156(1):264–270. <https://doi.org/10.1016/j.chroma.2006.11.098>
  - Rama Rao M, Faulkner DJ (2002) Isotactic Polymethoxydienes from the philippines sponge *Myriastrea clavosa*. *J Nat Prod* 65(8):1201–1203. <https://doi.org/10.1021/np020040b>
  - Ross SA, Weete JD, Schinazi RF, Wirtz SS, Tharnish P, Scheuer PJ, Hamann MT (2000) Mololipids, a new series of anti-HIV bromotyramine-derived compounds from a sponge of the order *Verongida*. *J Nat Prod* 63(4):501–503. <https://doi.org/10.1021/np980414u>
  - Rijpstra WIC, Reneerkens J, Piersma T, Damsté JSS (2007) Structural identification of the  $\beta$ -hydroxy fatty acid-based diester preen gland waxes of shorebirds. *J Nat Prod* 70(11):1804–1807. <https://doi.org/10.1021/np070254z>
  - Bloor S, Catchpole O, Mitchell K, Webby R, Davis P (2019) Antiproliferative acylated glycerols from New Zealand Propolis. *J Nat Prod* 82(9):2359–2367. <https://doi.org/10.1021/acs.jnatprod.8b00562>
  - Rodriguez-Saona CR, Maynard DF, Phillips S, Trumble JT (1999) Alkyl-furans: effects of alkyl side-chain length on insecticidal activity. *J Nat Prod* 62(1):191–193. <https://doi.org/10.1021/np980340m>
  - Nikolopoulou V, Aalizadeh R, Nika M-C, Thomaidis NS (2022) TrendProbe: time profile analysis of emerging contaminants by LC-HRMS non-target screening and deep learning convolutional neural network. *J Hazard Mater* 428:128194. <https://doi.org/10.1016/j.jhazmat.2021.128194>
  - Schinkel L, Lara-Martín PA, Giger W, Hollender J, Berg M (2022) Synthetic surfactants in Swiss sewage sludges: analytical challenges, concentrations and per capita loads. *Sci Total Environ* 808:151361. <https://doi.org/10.1016/j.scitotenv.2021.151361>
  - Mairinger T, Loos M, Hollender J (2021) Characterization of water-soluble synthetic polymeric substances in wastewater using LC-HRMS/MS. *Water Res* 190:116745. <https://doi.org/10.1016/j.watres.2020.116745>
  - Krauss M, Hug C, Bloch R, Schulze T, Brack W (2019) Prioritising site-specific micropollutants in surface water from LC-HRMS non-target screening data using a rarity score. *Environ Sci Eur* 31(1):45. <https://doi.org/10.1186/s12302-019-0231-z>
  - Jacob P, Barzen-Hanson KA, Helbling DE (2021) Target and nontarget analysis of per- and polyfluoroalkyl substances in wastewater from electronics fabrication facilities. *Environ Sci Technol* 55(4):2346–2356. <https://doi.org/10.1021/acs.est.0c06690>
  - Dimzon IK, Trier X, Frömler T, Helmus R, Knepper TP, de Voogt P (2016) High resolution mass spectrometry of polyfluorinated polyether-based formulation. *J Am Soc Mass Spectrom* 27(2):309–318. <https://doi.org/10.1007/s13361-015-1269-9>
  - Jia S, Marques Dos Santos M, Li C, Snyder SA (2022) Recent advances in mass spectrometry analytical techniques for per- and polyfluoroalkyl substances (PFAS). *Anal Bioanal Chem*. <https://doi.org/10.1007/s00216-022-03905-y>
  - Glüge J, Scheringer M, Cousins IT, DeWitt JC, Goldenman G, Herzke D, Lohmann R, Ng CA, Trier X, Wang Z (2020) An overview of the uses of per- and polyfluoroalkyl substances (PFAS). *Environ Sci Process Impacts* 22(12):2345–2373. <https://doi.org/10.1039/D0EM00291G>
  - Oellig C, Hammel Y-A (2019) Screening for chlorinated paraffins in vegetable oils and oil-based dietary supplements by planar solid phase extraction. *J Chromatogr A* 1606:460380. <https://doi.org/10.1016/j.chroma.2019.460380>
  - Glüge J, Schinkel L, Hungerbühler K, Cariou R, Bogdal C (2018) Environmental risks of medium-chain chlorinated paraffins (MCCPs): a review. *Environ Sci Technol* 52(12):6743–6760. <https://doi.org/10.1021/acs.est.7b06459>
  - Du X, Yuan B, Zhou Y, Benskin JP, Qiu Y, Yin G, Zhao J (2018) Short-, medium-, and long-chain chlorinated paraffins in wildlife from paddy fields in the Yangtze River Delta. *Environ Sci Technol* 52(3):1072–1080. <https://doi.org/10.1021/acs.est.7b05595>
  - Washington JW, Jenkins TM, Weber EJ (2015) Identification of unsaturated and 2H polyfluorocarboxylate homologous series and their detection in environmental samples and as polymer degradation products. *Environ Sci Technol* 49(22):13256–13263. <https://doi.org/10.1021/acs.est.5b03379>
  - Lai A, Clark AM, Escher BI, Fernandez M, McEwen LR, Tian Z, Wang Z, Schymanski EL (2022) The next frontier of environmental unknowns: substances of unknown or variable composition, complex reaction products, or biological materials (UVCBs). *Environ Sci Technol* 56(12):7448–7466. <https://doi.org/10.1021/acs.est.2c00321>
  - Schymanski EL, Singer HP, Longrée P, Loos M, Ruff M, Stravs MA, Ripollés Vidal C, Hollender J (2014) Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. *Environ Sci Technol* 48(3):1811–1818. <https://doi.org/10.1021/es40444374>
  - Carlson JE, Gasson JR, Barth T, Eide I (2012) Extracting homologous series from mass spectrometry data by projection on predefined vectors. *Chemom Intell Lab Syst* 114:36–43. <https://doi.org/10.1016/j.chemolab.2012.02.007>
  - Loos M, Singer H (2017) Nontargeted homologue series extraction from hyphenated high resolution mass spectrometry data. *J Cheminform*. <https://doi.org/10.1186/s13321-017-0197-z>
  - Mildau K, van der Hoof JJJ, Flasch M, Warth B, Abiead YE, Koellensperger G, Zanghellini J, Büschl C (2022) Homologue series detection and management in LC-MS data with homologuediscoverer. *bioRxiv*. <https://doi.org/10.1101/2022.07.20.500749>
  - Schymanski E (2020) *schymane/RChemMass*. <https://github.com/schymane/RChemMass>. Accessed 16 Aug 2020
  - St. Cholakov G, Stateva RP, Brauner N, Shacham M (2008) Estimation of properties of homologous series with targeted quantitative structure–property relationships. *J Chem Eng Data* 53(11):2510–2520. <https://doi.org/10.1021/je800272x>
  - Wiener H (1947) Structural determination of paraffin boiling points. *J Am Chem Soc* 69(1):17–20. <https://doi.org/10.1021/ja01193a005>
  - Kováts E (1958) Gas-chromatographische charakterisierung organischer verbindungen. Teil 1: retentionsindices aliphatischer halogenide, alkohole, aldehyde und ketone. *Helv Chim Acta* 41(7):1915–1932. <https://doi.org/10.1002/hlca.19580410703>
  - Schuffenhauer A, Schneider N, Hintermann S, Auld D, Blank J, Cotesta S, Engeloch C, Fechner N, Gaul C, Giovannoni J, Jansen J, Joslin J, Krastel P, Lounkine E, Manchester J, Monovich LG, Pelliccioli AP, Schwarze M, Shultz MD, Stiefel N, Baeschlin DK (2020) Evolution of Novartis' small molecule screening deck design. *J Med Chem* 63(23):14425–14447. <https://doi.org/10.1021/acs.jmedchem.0c01332>
  - PubChem. PubChem. <https://pubchem.ncbi.nlm.nih.gov/>. Accessed 02 Aug 2022
  - Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE (2021) PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res* 49(D1):D1388–D1395. <https://doi.org/10.1093/nar/gkaa971>
  - ChemSpider | Search and share chemistry. <https://www.chemspider.com/>. Accessed 2 Aug 2022
  - Pence HE, Williams A (2010) ChemSpider: an online chemical information resource. *J Chem Educ* 87(11):1123–1124. <https://doi.org/10.1021/ed100697w>
  - Warr W (2021) Report on an NIH workshop on ultralarge chemistry databases. <https://doi.org/10.26434/chemrxiv.14554803.v1>



41. Ehrlich H-C, Rarey M (2011) Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review. *WIREs Comput Mol Sci* 1(1):68–79. <https://doi.org/10.1002/wcms.5>
42. Raymond JW, Willett P (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J Comput Aided Mol Des* 16(7):521–533. <https://doi.org/10.1023/A:1021271615909>
43. Kruger F, Fechner N, Stiefl N (2020) Automated identification of chemical series: classifying like a medicinal chemist. *J Chem Inf Model* 60(6):2888–2902. <https://doi.org/10.1021/acs.jcim.0c00204>
44. Fournier-Viger P, Lin JC-W (2017) A survey of sequential pattern mining. *Data Sci Pattern Recognit* 1(1):54–77
45. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 39(15):2887–2893. <https://doi.org/10.1021/jm9602928>
46. Lai A. GitHub repository: an algorithm to classify homologous series. <https://github.com/adelenelai/onglai-classify-homologues>. Accessed 31 Aug 2022
47. Mohammed Taha H, Aalizadeh R, Alygizakis N, Antignac J-P, Arp HPH, Bade R, Baker N, Belova L, Bijlsma L, Bolton EE, Brack W, Celma A, Chen W-L, Cheng T, Chirsir P, Ćirka L, D'Agostino LA, DjoumbouFeunang Y, Dulio V, Fischer S, Gago-Ferrero P, Galani A, Geueke B, Glowacka N, Glüge J, Groh K, Grosse S, Haglund P, Hakkinen PJ, Hale SE, Hernandez F, Janssen EM-L, Jonkers T, Kiefer K, Kirchner M, Koschorreck J, Krauss M, Krier J, Lamoree MH, Letzel M, Letzel T, Li Q, Little J, Liu Y, Lunderberg DM, Martin JW, McEachran AD, McLean JA, Meier C, Meijer J, Menger F, Merino C, Muncke J, Muschket M, Neumann M, Neveu V, Ng K, Oberacher H, O'Brien J, Oswald P, Oswaldova M, Picache JA, Postigo C, Ramirez N, Reemtsma T, Renaud J, Rostkowski P, Rüdell H, Salek RM, Samanipour S, Scheringer M, Schliebner I, Schulz W, Schulze T, Sengl M, Shoemaker BA, Sims K, Singer H, Singh RR, Sumarah M, Thiessen PA, Thomas KV, Torres S, Trier X, van Wezel AP, Vermeulen RCH, Vlaanderen JJ, von der Ohe PC, Wang Z, Williams AJ, Willighagen EL, Wishart DS, Zhang J, Thomaidis NS, Hollender J, Slobodnik J, Schymanski EL (2022) The NORMAN Suspect List Exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry. *Environ Sci Eur* 34(1):104. <https://doi.org/10.1186/s12302-022-00680-6>
48. Dulio V, Koschorreck J, van Bavel B, van den Brink P, Hollender J, Munthe J, Schlabach M, Aalizadeh R, Agerstrand M, Ahrens L, Allan I, Alygizakis N, Barcelo' D, Bohlin-Nizzetto P, Boutroux S, Brack W, Bressy A, Christensen JH, Ćirka L, Covaci A, Derksen A, Deviller G, Dingemans MML, Engwall M, Fatta-Kassinos D, Gago-Ferrero P, Hernández F, Herzke D, Hilscherová K, Hollert H, Junghans M, Kasprzyk-Hordern B, Keiter S, Kools SAE, Krueve A, Lambropoulou D, Lamoree M, Leonards P, Lopez B, Lópezde Alda M, Lundy L, Makovinská J, Marióomez I, Martin JW, McHugh B, Miège C, O'Toole S, Perkola N, Polesello S, Posthuma L, Rodriguez-Mozaz S, Roesink I, Rostkowski P, Ruedel H, Samanipour S, Schulze T, Schymanski EL, Sengl M, Tarábek P, Ten Hulscher D, Thomaidis N, Togola A, Valsecchi S, van Leeuwen S, von der Ohe P, Vorkamp K, Vrana B, Slobodnik, J (2020) The NORMAN Association and the European Partnership for Chemicals Risk Assessment (PARC): Let's Cooperate! *Environ Sci Eur* 32(1), 100. <https://doi.org/10.1186/s12302-020-00375-w>
49. Schymanski EL, Kondić T, Neumann S, Thiessen PA, Zhang J, Bolton EE (2021) Empowering large chemical knowledge bases for exposomics: PubChemLite Meets MetFrag. *J Cheminform* 13(1):19. <https://doi.org/10.1186/s13321-021-00489-0>
50. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C (2021) COCONUT online: collection of open natural products database. *J Cheminform* 13(1):2. <https://doi.org/10.1186/s13321-020-00478-9>
51. COCONUT: natural products online. <https://coconut.naturalproducts.net/download>. Accessed 4 Apr 2022
52. Organization for Economic Co-operation and Development (2018) Toward a new comprehensive global database of per- and polyfluoroalkyl substances (PFASs): summary report on updating the OECD 2007 list of per- and polyfluoroalkyl substances (PFASs); Series on Risk Management No. 39 ENV/JM/MONO(2018)7; p 24. [https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV-JM-MONO\(2018\)7&doclanguage=en](https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV-JM-MONO(2018)7&doclanguage=en)
53. Sha B, Schymanski EL, Ruttkies C, Cousins IT, Wang Z (2019) Exploring open cheminformatics approaches for categorizing per- and polyfluoroalkyl substances (PFASs). *Environ Sci Process Impacts* 21(11):1835–1851. <https://doi.org/10.1039/C9EM00321E>
54. Daylight Theory: SMARTS—a language for describing molecular patterns. <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>. Accessed 10 Jun 2022
55. RDKit. <https://www.rdkit.org/>. Accessed 31 Aug 2022
56. Landrum G. RDKit Release 2021\_09\_4 (Q3 2021). [https://github.com/rdkit/rdkit/releases/tag/Release\\_2021\\_09\\_4](https://github.com/rdkit/rdkit/releases/tag/Release_2021_09_4). Accessed 31 Aug 2022
57. Python Release Python 3.7.0. Python.org. <https://www.python.org/downloads/release/python-370/>. Accessed 31 Aug 2022
58. Landrum G. Molecular sanitization in the RDKit. [https://www.rdkit.org/docs/RDKit\\_Book.html#molecular-sanitization](https://www.rdkit.org/docs/RDKit_Book.html#molecular-sanitization). Accessed 20 Jul 2022
59. Bolton E, Schymanski E, Kondic T, Thiessen P, Zhang J (Jeff) (2022) PubChemLite for Exposomics. <https://doi.org/10.5281/zenodo.6383860>
60. NORMAN Network. PubChem Classification Browser - NORMAN Suspect List Exchange Tree. <https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101>. Accessed 4 Apr 2022
61. NORMAN Network. NORMAN suspect list exchange. <https://www.norman-network.com/nds/SLE/>. Accessed 1 Nov 2022
62. PubChem Identifier Exchange Service. <https://pubchemdocs.ncbi.nlm.nih.gov/identifier-exchange-service>. Accessed 21 Sept 2020
63. SmilesGenerator (cdk 2.7.1 API). <https://cdk.github.io/cdk/2.7/docs/api/org.openscience.cdk.smiles/SmilesGenerator.html>. Accessed 17 Aug 2022
64. Lai A, Schaub J, Steinbeck C, Schymanski EL (2022) Supplementary information for "An algorithm to classify homologous series within compound datasets" (OngLai). <https://doi.org/10.5281/zenodo.7035020>
65. Schymanski E (2014) S7 | EAWAGSURF | Eawag surfactants suspect list. <https://doi.org/10.5281/zenodo.3549934>
66. Alygizakis N (2018) S23 | EIUBASURF | surfactant suspect list from EI and UBA. <https://doi.org/10.5281/zenodo.2648765>
67. Wang Z (2018) S25 | OECDPFAS | List of PFAS from the OECD. <https://doi.org/10.5281/zenodo.6349061>
68. Beckers M, Fechner N, Stiefl N (2022) 25 Years of small molecule optimization at novartis: a retrospective analysis of chemical series evolution. 12th Int. Conf. Chem. Struct. Plenary Sess. -1, Noordwijkerhout, The Netherlands
69. Remove flourinated natural products found by Adelene - Issue #89 - mSorok/NaturalProductsOnline. GitHub. <https://github.com/mSorok/NaturalProductsOnline/issues/89>. Accessed 1 Jul 2022
70. Wang Z, Buser AM, Cousins IT, Demattio S, Drost W, Johansson O, Ohno K, Patlewicz G, Richard AM, Walker GW, White GS, Leinala E (2021) A new OECD definition for per- and polyfluoroalkyl substances. *Environ Sci Technol* 55(23):15575–15578. <https://doi.org/10.1021/acs.est.1c06896>
71. Organization for Economic Co-operation and Development (2021) Reconciling terminology of the universe of per- and polyfluoroalkyl substances: recommendations and practical guidance; series on risk management; No. 61 ENV/CBC/MONO(2021)25; p 45. [https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/CBC/MONO\(2021\)25&docLanguage=en](https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/CBC/MONO(2021)25&docLanguage=en). Accessed 29 Aug 2022
72. How to delete the same substructure in one molecule separately · Discussion #4685 · rdkit/rdkit. GitHub. <https://github.com/rdkit/rdkit/discussions/4685>. Accessed 29 Jun 2022
73. Koutsoukas A, Paricharak S, Galloway WRJD, Spring DR, Ijzerman AP, Glen RC, Marcus D, Bender A (2014) How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. *J Chem Inf Model* 54(1):230–242. <https://doi.org/10.1021/ci400469u>
74. Helmus R, ter Laak TL, van Wezel AP, de Voegt P, Schymanski EL (2021) PatRoom: open source software platform for environmental mass spectrometry based non-target screening. *J Cheminform* 13(1):1. <https://doi.org/10.1186/s13321-020-00477-w>
75. Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H (2007) The scaffold tree—visualization of the scaffold universe by hierarchical scaffold classification. *J Chem Inf Model* 47(1):47–58. <https://doi.org/10.1021/ci600338x>
76. Faulon J-L, Visco DP, Pophale RS (2003) The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J Chem Inf Comput Sci* 43(3):707–720. <https://doi.org/10.1021/ci020345w>
77. Morgan HL (1965) The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *J Chem Doc* 5(2):107–113. <https://doi.org/10.1021/c160017a018>

78. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50(5):742–754. <https://doi.org/10.1021/ci100050t>
79. Bremser W (1978) Hose—a novel substructure code. *Anal Chim Acta* 103(4):355–365. [https://doi.org/10.1016/S0003-2670\(01\)83100-7](https://doi.org/10.1016/S0003-2670(01)83100-7)
80. Rajan K, Zielesny A, Steinbeck C (2021) DECIMER 1.0: deep learning for chemical image recognition using transformers. *J Cheminform* 13(1):61. <https://doi.org/10.1186/s13321-021-00538-8>
81. Wang Z, Adu-Kumi S, Diamond ML, Guardans R, Harner T, Harte A, Kajiwara N, Klánová J, Liu J, Moreira EG, Muir DCG, Suzuki N, Pinas V, Seppälä T, Weber R, Yuan B (2022) Enhancing scientific support for the stockholm convention's implementation: an analysis of policy needs for scientific evidence. *Environ Sci Technol* 56(5):2936–2949. <https://doi.org/10.1021/acs.est.1c06120>

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

OngLai, an algorithm to classify homologous series within compound datasets, was successfully developed and implemented as a Python package built using the RDKit. OngLai is freely available on GitHub (<https://github.com/adelenelai/onglai-classify-homologues>), as are the datasets used in this study, results, and analysis via Zenodo (<https://doi.org/10.5281/zenodo.7035020>). Running the OngLai algorithm entails executing one line of code in the command line interface that specifies inputs and various customisable user settings. Besides the input dataset of interest, the main input OngLai requires is a repeating unit of choice expressed in SMARTS notation, which reflects the original intended application of the algorithm to identify environmental unknowns, where series with a constant  $m/z$  difference are frequently detected in HRMS data.

OngLai was applied to three datasets, COCONUT, NORMAN-SLE, and PubChemLite for Exposomics, containing chemicals representing natural products, environmental chemicals, and those relevant to the human exposome respectively. Thousands of homologous series with  $\text{CH}_2$  repeating units representing alkyl chains were classified in each dataset, revealing the prevalence of homologous compounds in these domains. More importantly, these classified homologous series may support their identification in HRMS data, as analytical chemists would now be more able to match their observations of environmental unknowns in mass spectrometry data with entities in chemical databases. Additionally, OngLai proved to have good potential as an automated PFAS categorisation tool by comparing its  $\text{CF}_2$  series classification results with the cheminformatics tool *splitPFAS*, and the manual categorisation performed by PFAS experts for the OECD. Compared to computational identification workflows designed to analyse HRMS data, algorithms such as OngLai represent a new 'wave' of cheminformatics applications to environmental chemistry and the identification of unknowns that has been made possible thanks to the constant development of environmental chemical databases and lists such as those on the NORMAN-SLE and cheminformatics toolkits like the RDKit.

## Chapter 6

### Discussion

Despite the benefits chemicals have brought to society in the form of, for example, medicines, high-performance materials, and the ability to enhance properties such as safety (*e.g.*, flame-retardancy) or aesthetics (*e.g.*, paints or other coatings) *etc.*, the emission of chemicals into the environment has given rise to environmental chemical pollution. Environmental chemical pollution is a highly complex, multi-faceted problem as the design, production, regulation, use, and consumption of chemicals is driven by complex socioeconomic factors. In turn, many areas of research across myriad disciplines have focused on dealing with the issue of environmental chemicals from different perspectives, including but not limited to the natural sciences, history, and policy. Especially in the chemical sciences, and specifically environmental chemistry, scientific questions regarding these chemical pollutants abound, particularly concerning their unknown identities, environmental fate, exposure mechanisms, and potential toxicity to living organisms including humans.

Routine environmental monitoring is a fundamental aspect underpinning these critical questions about chemical pollutants. For example in European countries under the EU Water Framework Directive, national regulatory agencies are tasked with routine monitoring of surface water for the occurrence of roughly 45 target chemicals that mostly consist of pesticides and related compounds.<sup>99</sup> However, it is well documented that within the EU (and elsewhere), environmental surface water samples typically contain more chemical pollutants than are being monitored - in fact, HRMS data frequently contain hundreds, if not thousands of signals that remain unidentified. The prevalence of these chemical mixtures in the environment with the majority of components remaining unknown is problematic, and warrants sustained efforts to advance the knowledge of their identities.

Fundamental knowledge gaps in the identities of these non-target chemical pollutants continue to plague environmental chemists. Developments in HRMS instrumentation have allowed the high-resolution detection of these compounds in environmental samples, but analysing the measured data is a persistent bottleneck towards

identification of environmental unknowns, especially suspect and non-target compounds. Non-target compounds in particular are characterised by the lack of *a priori* information available regarding their identities beyond the signals measured by HRMS. As manual structure elucidation based on fragment spectra would be time-consuming, if not virtually impossible, applying cheminformatics and computational methods to support data analysis are widely considered *de rigueur*. With the advent of chemical data resources including databases, spectral libraries, software packages, and tools, efforts to combine and continuously develop these resources towards supporting the identification of environmental unknowns represent active areas of research.

The work in this dissertation specifically exploited the potential of up-and-coming open digital chemical resources to tackle the challenge of unknown identification within environmental chemical mixtures. Each of the four publications presented in the preceding chapters represents efforts to capitalise on different current developments in open chemistry data, cheminformatics toolkits, and the constant evolution of environmental chemistry research in tandem with the increasing availability of computational tools. Chapters 2 and 3 focused on developing workflows that exploited cutting-edge online chemical resources for analysing environmental chemistry data using water samples collected in Switzerland and Luxembourg that were part of regulatory sampling campaigns carried out by local authorities. Concerted efforts to communicate the findings of these studies in a transparent and constructive way that could inform future regulatory actions were prioritised in these works.

Chapter 4 then focused on the impending challenge and 'next frontier of environmental unknowns' known as UVCBs - registered chemical substances that are *de facto* mixtures. UVCBs are especially challenging to assess for regulators because of their unknown, variable, and complex compositions, and one of the principal obstacles to their sound management is the difficulty in representing them chemically. Accurate chemical representation of UVCBs underpins many aspects of its assessment, including but not limited to its analytical characterisation, evaluation of hazards, exposure and mixture toxicity modelling, and the ability to prioritise and possibly restrict these substances from a regulatory perspective. All these aforementioned aspects were critically reviewed in this chapter, and proposals for

improved and FAIR UVCB data management using the open Mixture InChI format were highlighted and discussed.

Lastly, a specific type of UVCB substance, namely mixtures of homologous compounds (also known as homologous series), was the focus of Chapter 5. Homologous series' signals are frequently detected in environmental samples because they are found in many High Production Volume substances, but are often intentionally ignored or deliberately excluded from identification exercises because of the inability to match them to corresponding database entries of homologous series. Thus, a cheminformatics algorithm was developed to classify homologous chemicals within compound datasets, which represents a step towards being able to match measured homologous signals to database compounds that would allow for their identification in environmental samples.

Overall, the work in this dissertation contributed to the advancement of data processing, classification, and analysis methods for the identification of suspect and non-target compounds *i.e.*, unknown chemical pollutants, in the environment. The Aims outlined in Section 1.2 were achieved in two ways: 1) by combining and incorporating state-of-the-art, open chemical resources into computational workflows to analyse HRMS data, and 2) developing a cheminformatics algorithm to classify homologous compounds in existing databases to enhance future identification efforts. Additionally, the critical review of UVCB substances, notably containing unprecedented proposals for its chemical data representation, highlighted a roadmap for dealing with these challenging substances. Therefore, the strength of this dissertation work is its demonstrative nature, as it focused on developing a breadth of computational and cheminformatics solutions along the entire environmental analytics pipeline while showcasing how various chemical data science resources could be leveraged as they were being released or developed.

On one hand, part of this dissertation features 'upstream' solutions involving the enhancement of database resources through possible annotation of related (homologous) compounds, but also on the other hand, 'downstream' solutions entailing development of identification workflows for specific environmental datasets. The alternative would have been to specialise or narrowly focus on fully identifying the chemical pollutants present in a specific environmental system or sample(s) to Level 1 certainty. However in practice, this would entail acquisition and measurement



of reference standards and further collaboration and feedback loops with analytical chemists. In other words, there would have been a tradeoff, at the expense of developing methods that are potentially more generalisable, applicable, and adaptable to other scientists' needs.

However, there exist methodological limitations in the current portfolio of scientific work that could be addressed in future efforts. These are thematically discussed in detail below.

## 6.1 Using Environmental Metadata in MetFrag for Unknown Identification

The *in silico* fragmentation tool MetFrag in its 'relaunched' version was used for identifying unknown masses in Chapters 2 and 3,<sup>85</sup> whereby the use of so-called 'environmental metadata' contributes to the identification procedure by increasing or decreasing the scores of candidate structures for a given unknown. Environmental metadata encompasses a broad range of information that is completely unrelated to the measured spectral information, and includes aspects like citation count, occurrence in patents, and presence or absence of the mass in user-defined suspect lists. In MetFrag, if a compound happens to be highly omnipresent in these areas, it will be up-prioritised in the list of candidates suggested in MetFrag's results. Essentially, this identification paradigm relies on the premise that the more documented a chemical compound is, either because it is widely produced, used, studied, or patented, the higher the likelihood it is a good candidate structure for a given unknown mass.

On one hand, this approach technically introduces bias in the identification workflow, as the information provided by environmental metadata is unrelated to the analytical measurement performed and the resulting mass spectra obtained. The fact that information unrelated to the analyte itself could be a contributing factor for unknown identification based on HRMS measurements does not necessarily invalidate any tentative identification made, but may introduce errors in the identification process by obscuring the relevance of candidates with matching analytical data, but that do not

have copious corresponding patent information, citations, presence in suspect lists *etc.*

However, on the other hand, considering the vastness of the possible chemical space of environmental pollutants that exacerbates the difficulty of non-target identification, environmental metadata may serve as a reasonable indication of the likelihood of a tentative identification being valid. Therefore, to mitigate the potential negative effects related to bias of introducing environmental metadata as a factor in unknown identification, appropriate weightings should be employed by MetFrag users to balance out the contribution of environmental metadata information with spectral data so as to obtain reasonable candidate structures. These weightings could be devised taking the quality of the spectral data into account, as discussed in Chapter 2.

## 6.2 Suspect Screening using Suspect Lists

The work in Chapter 3 involves suspect screening, which relied on a list of (pharmaceutical) products that had been registered for the Luxembourgish market published by the Luxembourgish *Caisse Nationale de Santé*, *i.e.*, a source completely distinct from sample measurement and spectral data generation. The use of such suspect lists in suspect screening, as performed in Chapter 3, is an approach that has become more widespread in the last 5-10 years, attributable to the intense development of chemical resources like the NORMAN Suspect List Exchange.<sup>82</sup> The NORMAN-SLE currently hosts 99 thematic lists of chemicals that were generated by NORMAN partners, including intense curation efforts that contributed to the findability, accessibility, interoperability, and reusability of this information, and continues to grow in its number of lists.

However, in light of these developments, the question of “screen big vs. screen smart” becomes ever more pertinent. “Screen big” refers to the use of a large suspect list for suspect screening, because presumably the larger the list, the larger the coverage of possible compound space, and the higher the possibility that unknown masses will match an element in the list. However, screening big in this way introduces more possibilities for false positives, as the likelihood of a matching mass becomes higher with an increasing list size, regardless of whether the compound it

represents is in fact a reasonable candidate. Conversely, “screen smart” implies a more focused approach, using a smaller but more meaningful list that may be tailored to the study based on domain knowledge. It remains challenging to strike a reasonable balance between screening big and screening smart, and analyses to identify potential false positives in suspect screening could be performed when using large lists.

## 6.3 Different Workflows for Different Studies? Harmonisation of NTA Towards Use in Regulatory Environmental Monitoring

Chapters 2 and 3 presented workflows for analysing non-target and suspect compounds that were deliberately developed for the respective studies. More specifically, different tools, datasets, and spectral libraries had to be integrated together into one workflow, as these discrete building blocks often tend to have general scopes of application that need to be refined for a given use, or are in fact originally from metabolomics or proteomics and must be repurposed for environmental analysis.

Such is the challenge in non-target analysis - that individual workflows are usually created for different datasets or are used by different research groups, depending on the types of analytical instruments and data formats, researchers' programming abilities, accessibility of software tools (licences are needed for proprietary software in some cases), and aims of the study. This heterogeneity is the basis of various collaborative trials and identification contests that have taken place over the last decade, where different researchers are invited to analyse the same dataset using a workflow of their choice, often producing different results.<sup>100–102</sup>

On one hand, the fact that researchers have flexibility in developing their own workflows is potentially beneficial, as each study could have different requirements that may not be suitably covered by a generic workflow. However, two distinct disadvantages pertain to this flexibility: 1) significant time and resources are dedicated to developing these highly-specialised workflows and 2) as a result of such

different workflows, their comparability and ability to assess their performance becomes limited.

Multiple calls for harmonisation and eventual standardisation of non-target analysis protocols, computational workflow methods, and reporting have been issued in the past decade, particularly if NTA is to be used in regulatory environmental monitoring as a means for improved chemicals management.<sup>103</sup> Software platforms such as patRoom represent possible solutions to this fragmented landscape of non-target analysis.<sup>104</sup> patRoom is an open software platform that combines multiple data analysis routines and algorithms for environmental non-target analysis. However, voluntary universal adoption of patRoom is unlikely, especially in the short- to medium-term, as researchers tend to prefer maintaining their respective *status quo* and to continue using workflows they have developed or used in the past, likely because of the significant effort required to learn to deploy new tools.

Meanwhile, multiple guidance documents and study reporting tools have been proposed towards the harmonisation of NTA approaches.<sup>105</sup> The prospect of full automation of these workflows, possibly by applications of machine learning as well as increased functionality and ability to interface chemical databases is anticipated. Furthermore, such automation and harmonisation would likely facilitate data processing and analysis, as custom solutions would not have to be developed for each different approach. Overall, the scientific community, together with regulatory stakeholders, likely need to reach some consensus if NTA is ever to be adopted as a routine water monitoring approach.

## 6.4 Open Science and FAIR Data Approaches of this Dissertation

All the work produced over the course of this dissertation adheres to Open Science and FAIR Data principles. In practice, this means that all tools, software and databases that were used and/or produced in this dissertation work are open source, as are all resulting research products *i.e.*, the peer-reviewed publications themselves,

in addition to the preprints of these manuscripts, an oral presentation slide deck,<sup>106</sup> as well as a scientific poster.<sup>107</sup>

Besides the open access nature of all publications produced in this work, all code that was produced is freely available online on repositories such as GitHub or GitLab, and archived using stable URLs or DOIs where appropriate. In particular, the work in Chapter 5 involving the development and implementation of the OngLai algorithm was published as a Python package. Packaging in this way not only ensures that OngLai can be easily downloaded, installed, and used, but that future development of OngLai by the open source community is facilitated because of the structured organisation of code and corresponding documentation.

Additionally, all data analysed were uploaded onto open repositories, namely MassIVE for mass spectral data, as well as archived versions of open data on Zenodo, all with corresponding Digital Object Identifiers (DOIs). Jupyter Notebooks are included to demonstrate how the code and analyses of these data were carried out, and can be executed to reproduce the same results as obtained in the respective publications.

Considerable time and effort was required to prepare data, code, analyses, and results that adhere to the FAIR Data and Open Science principles. Documentation in the form of metadata, code comments, or otherwise is an essential aspect of this approach. Furthermore, uploads to open repositories require 'clean' and organised outputs (*e.g.*, via software packaging) that are ideally easily understood by potential future users.

In spite of these added requirements, the importance and benefits of FAIR Data and Open Science are undeniable. On one hand, research such as the present dissertation that was funded by public funding sources is produced under the obligation of maintaining accessibility to all its outputs. On the other hand, ensuring all results are FAIR and Open not only increases the transparency of research, but may also enable other researchers working on similar problems to benefit from tools that have been developed, or the accessibility of data that can be shared and (re)used by others. For example, upon the (open) publication of the manuscript in Chapter 2, including the list of non-target masses of interest that were investigated in the study, regulators from a different organisation in Switzerland contacted our author

team, as they too had been working on trying to elucidate the structure of a non-target mass identical to one of the masses in our study. This connection then sparked further joint work, which represents critical steps forward towards elucidating this chemical unknown.

Such collaborative aspects are critical to carrying out environmental research and ultimately solving environmental problems, as joint efforts between researchers is essential considering that environmental pollution problems do not occur in silos and likely affect multiple stakeholders simultaneously. Thus on balance, the resources dedicated to ensuring FAIR Data and Open Science principles in research, especially in the environmental domain, are paramount, justified, and have been exemplified in this dissertation.

## 6.5 Bottleneck in Availability of Data on Environmental Pollutants

Despite the increasing availability of chemical data and open resources that has been discussed at length throughout this dissertation, fundamental data gaps persist concerning chemicals in the form of incomplete lists of marketed products; their toxicity, environmental fate, and safety properties; their production and usage in terms of tonnages and emission routes; reference standards and analytical data; and in the specific case of UVCBs, chemical structure information. Together, these gaps likely hinder the identification of these substances in the environment, which has consequences for trying to understand their potential toxic (mixture) effects.

In the particular case of UVCBs, which was the focus of Chapters 4 and 5 and likely represent the subject of much future work, the lacuna of information on their identities is attributable to the fact that regulatory agencies do not always require them, and even if this information is disclosed to regulators, is not made available in the public domain, including researchers. The latter could be a result of protective Confidential Business Information clauses, or the fact that testing is not required when a substance is declared a UVCB. Alternatively, this gap could also be a factor of the current technology used for disclosing and storing this information. In Europe, the International Uniform Chemical Information Database (IUCLID) system is the



predominant tool used in this context, but its relatively nonspecific data collection scheme, particularly for the variable composition of UVCB substances, has been found to complicate or in some cases hinder comparisons between, for example, lead and member country dossiers of the same substance.<sup>108</sup>

Moving forward, the availability of UVCB data will remain a challenge.<sup>67,109</sup> While there have been some efforts to elucidate the structures of UVCBs *in silico* based on substance name, the fragmented availability of information across the various online databases posed severe obstacles to these efforts.<sup>106</sup> Ultimately, analytical verification may be the only definitive method to determine the composition and/or structure of a UVCB. Thus, only if analytical verification is made a requirement of product registration, would this information become more readily available.

Overall, the work in this dissertation deliberately addressed the most challenging of environmental unknowns, while at the same time, offered holistic perspectives on the upcoming challenges within environmental chemistry concerning the specific case of UVCBs. As UVCBs are so varied in nature, focusing on a specific subtype through the OngLai algorithm to classify homologous series provided some measure of tractability that can hopefully benefit future researchers and alleviate the identification gap of chemical unknowns in non-target HRMS data.

## Chapter 7

### Conclusion & Perspectives

Pollution of the environment caused by chemicals is a persistent, multi-faceted problem. Until now, their production, use, and disposal has largely been dictated by historical, socio-economic, and political forces that tended to prioritise their benefits: the prospect of stronger, safer, better-performing materials and products. However, despite these advantages to society, concerns regarding their (environmental) fate, and consequently, the potential impacts these compounds might have on ecosystems and human health have been largely overshadowed. Six decades have elapsed since the publication of Rachel Carson's *Silent Spring*, the book that brought incidents of chemical pollution caused by pesticides and their consequences on wildlife and ecosystems to the mainstream public's attention. However, despite some regulatory initiatives to register, assess, monitor, and manage chemicals that were initially inspired by *Silent Spring*, in addition to significant grassroots efforts, numerous cases of scientific research have shown that chemicals are more ubiquitous than ever, occurring not just in the environment and within wildlife, but also increasingly within ourselves.

That chemicals are virtually everywhere is a fact - megatonnes are registered for production, sale, and use each year, as well as being routinely detected in the environment. What remains elusive, particularly in the public domain, is a complete understanding of their identities; more specifically, their chemical structures. This ignorance plagues two key stakeholders. Firstly, regulators often do not have full knowledge of the structures of the chemicals registered under their purview because of Confidential Business Information clauses, or simply because they were not characterised upon registration and do not need to be, as is widely the case for UVCB substances. Without knowledge of chemical structure, opportunities to screen or validate studies on their properties, let alone prescribe effective restriction or mitigation measures based on the results of environmental monitoring are hampered, which may compromise their sound management.

The second key stakeholder concerned is the scientific research community; the challenge of assigning structures to the unknown chemicals detected in environmental samples is ever persistent, difficult, and overwhelming in scale given there are so many of them. Not knowing the structures of these chemicals impedes our mechanistic understanding of their potential effects on human health and possible links to the onset of disease, not to mention their impacts on wildlife and the environment as a whole.

Importantly, it is not that this knowledge of chemical structure is completely unknown. In most cases, the producers of these chemicals likely know the structures of the compounds they manufacture, or at least have some partial information, because such knowledge is vital to their product development pipeline. For example, the producers likely had to screen compounds based on potency in the initial stages of product design, and then devise a synthesis route to achieve them. This situation likely describes the cases of agrochemicals and pharmaceuticals in particular. In the case of industrial chemicals, it is likely that at least the starting materials are known, since they must be acquired for production, and knowledge of how they will react together probably dictate how they can be manufactured at scale. It may be that during product development, certain functional groups or substructures are deliberately included or removed to optimise for desirable end-product properties. Finally, knowledge of chemical structure when developing the scope of the patent that would be eventually filed for the compound is likely important. Thus, it is not that knowledge of chemical structures is non-existent; rather the issue is that systemic legal and socioeconomic barriers to freely accessing this knowledge exist. Following the current paradigm, generating knowledge of these structures has become to a great extent the *de facto* occupation of contemporary environmental science research.

Therefore, this dissertation attempts to address this lack of knowledge concerning the chemical structures of substances in our environment. Its main contributions are threefold. First, cutting-edge digital chemical resources were exploited within computational workflows to enhance the identification of unknown compounds in the environment. These workflows consisted of open software tools, chemical databases, and environmental chemical lists that were integrated together within an analysis pipeline to be deployed on environmental samples. In two separate collaborations

with local regulators, surface water samples from Switzerland and Luxembourg that were measured using LC-HRMS were analysed using these workflows as a means of performing non-target and suspect screenings that eventually identified pharmaceutical and industrial compounds. Notably, transformation products of pharmaceuticals were incorporated into the suspect screening of the Luxembourgish surface water samples, several of which were identified. This ability to screen for transformation products based on information available in open chemical databases and the literature was a result of advances in data mining that were developed in this work.

Second, a significantly large class of substances that make up 20-40% of chemical registries but whose chemical structures are ambiguous or unknown was critically reviewed from the perspectives of risk assessment, cheminformatics, toxicology, analytical chemistry and current regulatory practice. A first of its kind in terms of breadth of scope, this review captured the interdisciplinary challenges that mark this 'next frontier of environmental unknowns', and proposed several main areas of further research as well as some constructive solutions for their improved management. With this, UVCBs are anticipated to attain a higher position on the collective research agenda of the environmental chemistry community.

Finally, a cheminformatics algorithm, OngLai, was developed towards bridging analytical detection with database identification of homologous series. OngLai was successfully designed and openly implemented as a Python package built using the open cheminformatics toolkit RDKit. This algorithm, which classifies homologous series within compound datasets, represents a step closer towards assigning chemical structures to the numerous homologous compound signals detected in environmental samples using LC-HRMS. Thus by uncovering their identities, OngLai may play a role in reducing the number of environmental unknowns so that future research can better prioritise relevant features for non-target analysis.

Continued efforts in compound identification, such as those presented in this dissertation, are imperative for multiple reasons. For example, if they are to be deliberately removed from the environment or converted into more benign transformation products, understanding of chemical structure is foundational for elucidating their reaction mechanisms, which is needed to inform wastewater

treatment processes,<sup>110</sup> or application of other remediation measures such as the use of microbial degradation agents.<sup>111</sup> Furthermore, as has been mentioned, the links of chemical exposure to the onset of disease can be better understood on a mechanistic level with available knowledge of chemical structure, which in turn could pave the way for the development of successful therapies.

Additionally, if chemicals in the environment are to be managed effectively, the ability to identify and quantify them during routine environmental monitoring is crucial, as such information would determine the necessary mitigation measures, and the effectiveness of their implementation over time. Especially considering that the chemical industry's pace of production far exceeds that at which they can be assessed, the concept of chemical grouping is a plausible next approach, but one that would require knowledge of chemical structure as a basis.<sup>112</sup> All in all, knowledge of chemical identities would boost various scientific and regulatory pursuits regarding the safe management of chemicals.

The advent of Big Data, continuous development of cheminformatics toolkits, and use of Machine Learning in further applications of AI-driven molecular informatics is cause for further optimism in pursuing knowledge of chemical structures in the context of environmental chemistry. The culture of data sharing is fundamentally shifting towards being more progressive, structured, and open thanks to both technical and non-technical factors: infrastructure improvements in chemical data deposition and management,<sup>113</sup> as well as enhanced open publishing policies mandated by initiatives like *Plan S*.<sup>114</sup> Multiple applications of machine learning to chemical screening,<sup>65,115–117</sup> as well as structure elucidation have been pursued;<sup>118</sup> in fact, as more data become available, it may even appear possible to circumvent structure elucidation altogether in favour of directly predicting ecotoxicological properties based on mass spectra.<sup>119</sup>

That said, it remains critical to 'close the loop' by connecting research outcomes to policy and regulatory needs with respect to environmental chemical pollution. Thus, a functional mechanism that can transfer empirical evidence and regulatory needs back and forth between scientists and policymakers will be essential. At present, it has been observed that the lack of a science-policy interface between these stakeholders has contributed to ineffective policy initiatives and scientific endeavours that do not

directly address immediate regulatory needs.<sup>120,121</sup> Initiatives to found a science-policy body in the form of an 'Intergovernmental Panel on Climate Change for chemicals' in the hope of more concrete and concerted collaborative action are in progress.

Nevertheless, even if there were a panacea for solving the chemical identification problem, the problem of chemical pollution would still persist, as society will likely continue its patterns of consumption and emission. Thus, a fundamental paradigm shift is likely needed to address the crisis of chemicals as a whole in the first place.<sup>3</sup> Calls have been made to simply reduce the amount of chemicals used in products to begin with within the context of defined essential use,<sup>122-124</sup> which would make their disposal and recycling more feasible within an ideal circular economy. Even more ambitious is the push to make chemicals Safe and Sustainable by Design (SSbD), a paradigm that considers the possible effects of a chemical as early as its conception. There have been preliminary steps in this area,<sup>125</sup> and the making of SSbD a research priority, for example within the current European Partnership for the Assessment of Risks from Chemicals<sup>126</sup> in coming years promises further developments towards this proactive approach.



## 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.* **2020**, *54* (5), 2575–2584. <https://doi.org/10.1021/acs.est.9b06379>.
- (2) European Commission. *White Paper - Strategy for a future Chemicals Policy - 52001DC0088*. <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A52001DC0088> (accessed 2022-10-22).
- (3) Carney Almroth, B.; Cornell, S. E.; Diamond, M. L.; de Wit, C. A.; Fantke, P.; Wang, Z. Understanding and Addressing the Planetary Crisis of Chemicals and Plastics. *One Earth* **2022**, *5* (10), 1070–1074. <https://doi.org/10.1016/j.oneear.2022.09.012>.
- (4) Rockström, J.; Steffen, W.; Noone, K.; Persson, Å.; Chapin, F. S.; Lambin, E. F.; Lenton, T. M.; Scheffer, M.; Folke, C.; Schellnhuber, H. J.; Nykvist, B.; de Wit, C. A.; Hughes, T.; van der Leeuw, S.; Rodhe, H.; Sörlin, S.; Snyder, P. K.; Costanza, R.; Svedin, U.; Falkenmark, M.; Karlberg, L.; Corell, R. W.; Fabry, V. J.; Hansen, J.; Walker, B.; Liverman, D.; Richardson, K.; Crutzen, P.; Foley, J. A. A Safe Operating Space for Humanity. *Nature* **2009**, *461* (7263), 472–475. <https://doi.org/10.1038/461472a>.
- (5) Diamond, M. L.; de Wit, C. A.; Molander, S.; Scheringer, M.; Backhaus, T.; Lohmann, R.; Arvidsson, R.; Bergman, Å.; Hauschild, M.; Holoubek, I.; Persson, L.; Suzuki, N.; Vighi, M.; Zetzsch, C. Exploring the Planetary Boundary for Chemical Pollution. *Environ. Int.* **2015**, *78*, 8–15. <https://doi.org/10.1016/j.envint.2015.02.001>.
- (6) Persson, L.; Carney Almroth, B. M.; Collins, C. D.; Cornell, S.; de Wit, C. A.; Diamond, M. L.; Fantke, P.; Hassellöv, M.; MacLeod, M.; Ryberg, M. W.; Søgaard Jørgensen, P.; Villarrubia-Gómez, P.; Wang, Z.; Hauschild, M. Z. Outside the Safe Operating Space of the Planetary Boundary for Novel Entities. *Environ. Sci. Technol.* **2022**, *56* (3), 1510–1521. <https://doi.org/10.1021/acs.est.1c04158>.
- (7) Bernhardt, E. S.; Rosi, E. J.; Gessner, M. O. Synthetic Chemicals as Agents of Global Change. *Front. Ecol. Environ.* **2017**, *15* (2), 84–90. <https://doi.org/10.1002/fee.1450>.
- (8) Lenoir, A.; Boulay, R.; Dejean, A.; Touchard, A.; Cuvillier-Hot, V. Phthalate Pollution in an Amazonian Rainforest. *Environ. Sci. Pollut. Res.* **2016**, *23* (16), 16865–16872. <https://doi.org/10.1007/s11356-016-7141-z>.
- (9) Xu, Z.; Mi, W.; Mi, N.; Fan, X.; Zhou, Y.; Tian, Y. Comprehensive Evaluation of Soil Quality in a Desert Steppe Influenced by Industrial Activities in Northern China. *Sci. Rep.* **2021**, *11* (1), 17493. <https://doi.org/10.1038/s41598-021-96948-7>.
- (10) Gallen, C.; Heffernan, A. L.; Kaserzon, S.; Dogruer, G.; Samanipour, S.; Gomez-Ramos, M. J.; Mueller, J. F. Integrated Chemical Exposure Assessment of Coastal Green Turtle Foraging Grounds on the Great Barrier Reef. *Sci. Total Environ.* **2019**, *657*, 401–409. <https://doi.org/10.1016/j.scitotenv.2018.11.322>.
- (11) Wong, F.; Hung, H.; Dryfhout-Clark, H.; Aas, W.; Bohlin-Nizzetto, P.; Breivik, K.; Mastromonaco, M. N.; Lundén, E. B.; Ólafsdóttir, K.; Sigurðsson, Á.; Vorkamp, K.; Bossi, R.; Skov, H.; Hakola, H.; Barresi, E.; Sverko, E.; Fellin, P.; Li, H.; Vlasenko, A.; Zapevalov, M.; Samsonov, D.; Wilson, S. Time Trends of Persistent Organic Pollutants (POPs) and Chemicals of Emerging Arctic

- Concern (CEAC) in Arctic Air from 25 Years of Monitoring. *Sci. Total Environ.* **2021**, 775, 145109. <https://doi.org/10.1016/j.scitotenv.2021.145109>.
- (12) Abrahamsson, D. P.; Warner, N. A.; Jantunen, L.; Jahnke, A.; Wong, F.; MacLeod, M. Investigating the Presence and Persistence of Volatile Methylsiloxanes in Arctic Sediments. *Environ. Sci. Process. Impacts* **2020**, 22 (4), 908–917. <https://doi.org/10.1039/C9EM00455F>.
- (13) Washington, J. W.; Rosal, C. G.; McCord, J. P.; Strynar, M. J.; Lindstrom, A. B.; Bergman, E. L.; Goodrow, S. M.; Tadesse, H. K.; Pilant, A. N.; Washington, B. J.; Davis, M. J.; Stuart, B. G.; Jenkins, T. M. Nontargeted Mass-Spectral Detection of Chloroperfluoropolyether Carboxylates in New Jersey Soils. *Science* **2020**, 368 (6495), 1103–1107. <https://doi.org/10.1126/science.aba7127>.
- (14) Chiaia-Hernández, A. C.; Scheringer, M.; Müller, A.; Stieger, G.; Wächter, D.; Keller, A.; Pintado-Herrera, M. G.; Lara-Martin, P. A.; Bucheli, T. D.; Hollender, J. Target and Suspect Screening Analysis Reveals Persistent Emerging Organic Contaminants in Soils and Sediments. *Sci. Total Environ.* **2020**, 740, 140181. <https://doi.org/10.1016/j.scitotenv.2020.140181>.
- (15) Nishimuta, K.; Ueno, D.; Takahashi, S.; Kuwae, M.; Kadokami, K.; Miyawaki, T.; Matsukami, H.; Kuramochi, H.; Higuchi, T.; Koga, Y.; Matsumoto, H.; Ryuda, N.; Miyamoto, H.; Haraguchi, T.; Sakai, S.-I. Use of Comprehensive Target Analysis for Determination of Contaminants of Emerging Concern in a Sediment Core Collected from Beppu Bay, Japan. *Environ. Pollut.* **2021**, 272, 115587. <https://doi.org/10.1016/j.envpol.2020.115587>.
- (16) Fabregat-Safont, D.; Ibáñez, M.; Bijlsma, L.; Hernández, F.; Waichman, A. V.; de Oliveira, R.; Rico, A. Wide-Scope Screening of Pharmaceuticals, Illicit Drugs and Their Metabolites in the Amazon River. *Water Res.* **2021**, 200, 117251. <https://doi.org/10.1016/j.watres.2021.117251>.
- (17) Wilkinson, J. L.; Boxall, A. B. A.; Kolpin, D. W.; Leung, K. M. Y.; Lai, R. W. S.; Galbán-Malagón, C.; Adell, A. D.; Mondon, J.; Metian, M.; Marchant, R. A.; Bouzas-Monroy, A.; Cuni-Sanchez, A.; Coors, A.; Carriquiriborde, P.; Rojo, M.; Gordon, C.; Cara, M.; Moermond, M.; Luarte, T.; Petrosyan, V.; Perikhanyan, Y.; Mahon, C. S.; McGurk, C. J.; Hofmann, T.; Kormoker, T.; Iniguez, V.; Guzman-Otazo, J.; Tavares, J. L.; Gildasio De Figueiredo, F.; Razzolini, M. T. P.; Dougnon, V.; Gbaguidi, G.; Traoré, O.; Blais, J. M.; Kimpe, L. E.; Wong, M.; Wong, D.; Ntchantcho, R.; Pizarro, J.; Ying, G.-G.; Chen, C.-E.; Páez, M.; Martínez-Lara, J.; Otamonga, J.-P.; Poté, J.; Ifo, S. A.; Wilson, P.; Echeverría-Sáenz, S.; Udikovic-Kolic, N.; Milakovic, M.; Fatta-Kassinou, D.; Ioannou-Ttofa, L.; Belušová, V.; Vymazal, J.; Cárdenas-Bustamante, M.; Kassa, B. A.; Garric, J.; Chaumot, A.; Gibba, P.; Kunchulia, I.; Seidensticker, S.; Lyberatos, G.; Halldórsson, H. P.; Melling, M.; Shashidhar, T.; Lamba, M.; Nastiti, A.; Supriatin, A.; Pourang, N.; Abedini, A.; Abdullah, O.; Gharbia, S. S.; Pilla, F.; Chefetz, B.; Topaz, T.; Yao, K. M.; Aubakirova, B.; Beisenova, R.; Olaka, L.; Mulu, J. K.; Chatanga, P.; Ntuli, V.; Blama, N. T.; Sherif, S.; Aris, A. Z.; Looi, L. J.; Niang, M.; Traore, S. T.; Oldenkamp, R.; Ogunbanwo, O.; Ashfaq, M.; Iqbal, M.; Abdeen, Z.; O'Dea, A.; Morales-Saldaña, J. M.; Custodio, M.; de la Cruz, H.; Navarrete, I.; Carvalho, F.; Gogra, A. B.; Koroma, B. M.; Cerkvenik-Flajs, V.; Gombač, M.; Thwala, M.; Choi, K.; Kang, H.; Ladu, J. L. C.; Rico, A.; Amerasinghe, P.; Sobek, A.; Horlitz, G.; Zenker, A. K.; King, A. C.; Jiang, J.-J.; Kariuki, R.; Tumbo, M.; Tezel, U.; Onay, T. T.; Lejju, J. B.; Vystavna, Y.; Vergeles, Y.; Heinzen, H.; Pérez-Parada, A.; Sims, D. B.; Figy, M.; Good, D.; Teta, C. Pharmaceutical Pollution of the World's Rivers. *Proc. Natl. Acad. Sci.* **2022**, 119 (8), e2113947119. <https://doi.org/10.1073/pnas.2113947119>.
- (18) Kandie, F. J.; Krauss, M.; Beckers, L.-M.; Massei, R.; Fillinger, U.; Becker, J.;

- Liess, M.; Torto, B.; Brack, W. Occurrence and Risk Assessment of Organic Micropollutants in Freshwater Systems within the Lake Victoria South Basin, Kenya. *Sci. Total Environ.* **2020**, *714*, 136748. <https://doi.org/10.1016/j.scitotenv.2020.136748>.
- (19) Fang, W.; Peng, Y.; Muir, D.; Lin, J.; Zhang, X. A Critical Review of Synthetic Chemicals in Surface Waters of the US, the EU and China. *Environ. Int.* **2019**, *131*, 104994. <https://doi.org/10.1016/j.envint.2019.104994>.
- (20) Liu, L.; Aljathelah, N. M.; Hassan, H.; Giraldez, B. W.; Leitão, A.; Bayen, S. Targeted and Suspect Screening of Contaminants in Coastal Water and Sediment Samples in Qatar. *Sci. Total Environ.* **2021**, *774*, 145043. <https://doi.org/10.1016/j.scitotenv.2021.145043>.
- (21) Feng, X.; Sun, H.; Liu, X.; Zhu, B.; Liang, W.; Ruan, T.; Jiang, G. Occurrence and Ecological Impact of Chemical Mixtures in a Semiclosed Sea by Suspect Screening Analysis. *Environ. Sci. Technol.* **2022**, *56* (15), 10681–10690. <https://doi.org/10.1021/acs.est.2c00966>.
- (22) Weigel, S.; Kuhlmann, J.; Hühnerfuss, H. Drugs and Personal Care Products as Ubiquitous Pollutants: Occurrence and Distribution of Clofibric Acid, Caffeine and DEET in the North Sea. *Sci. Total Environ.* **2002**, *295* (1), 131–141. [https://doi.org/10.1016/S0048-9697\(02\)00064-5](https://doi.org/10.1016/S0048-9697(02)00064-5).
- (23) Brumovský, M.; Bečanová, J.; Kohoutek, J.; Thomas, H.; Petersen, W.; Sørensen, K.; Sáňka, O.; Nizzetto, L. Exploring the Occurrence and Distribution of Contaminants of Emerging Concern through Unmanned Sampling from Ships of Opportunity in the North Sea. *J. Mar. Syst.* **2016**, *162*, 47–56. <https://doi.org/10.1016/j.jmarsys.2016.03.004>.
- (24) Muir, D.; Miaz, L. T. Spatial and Temporal Trends of Perfluoroalkyl Substances in Global Ocean and Coastal Waters. *Environ. Sci. Technol.* **2021**, *55* (14), 9527–9537. <https://doi.org/10.1021/acs.est.0c08035>.
- (25) Pike, K. A.; Edmiston, P. L.; Morrison, J. J.; Faust, J. A. Correlation Analysis of Perfluoroalkyl Substances in Regional U.S. Precipitation Events. *Water Res.* **2021**, *190*, 116685. <https://doi.org/10.1016/j.watres.2020.116685>.
- (26) Chen, M.; Wang, C.; Gao, K.; Wang, X.; Fu, J.; Gong, P.; Wang, Y. Perfluoroalkyl Substances in Precipitation from the Tibetan Plateau during Monsoon Season: Concentrations, Source Regions and Mass Fluxes. *Chemosphere* **2021**, *282*, 131105. <https://doi.org/10.1016/j.chemosphere.2021.131105>.
- (27) Bangma, J. T.; Ragland, J. M.; Rainwater, T. R.; Bowden, J. A.; Gibbons, J. W.; Reiner, J. L. Perfluoroalkyl Substances in Diamondback Terrapins (*Malaclemys Terrapin*) in Coastal South Carolina. *Chemosphere* **2019**, *215*, 305–312. <https://doi.org/10.1016/j.chemosphere.2018.10.023>.
- (28) Carrizo, J. C.; Vo Duy, S.; Munoz, G.; Marconi, G.; Amé, M. V.; Sauvé, S. Suspect Screening of Pharmaceuticals, Illicit Drugs, Pesticides, and Other Emerging Contaminants in Argentinean *Piaractus Mesopotamicus*, a Fish Species Used for Local Consumption and Export. *Chemosphere* **2022**, *309*, 136769. <https://doi.org/10.1016/j.chemosphere.2022.136769>.
- (29) Hornek-Gausterer, R.; Oberacher, H.; Reinstadler, V.; Hartmann, C.; Liebmann, B.; Lomako, I.; Scharf, S.; Posautz, A.; Kübber-Heiss, A. A Preliminary Study on the Detection of Potential Contaminants in the European Brown Hare (*Lepus Europaeus*) by Suspect and Microplastics Screening. *Environ. Adv.* **2021**, *4*, 100045. <https://doi.org/10.1016/j.envadv.2021.100045>.
- (30) Zhao, J.-H.; Hu, L.-X.; He, L.-X.; Wang, Y.-Q.; Liu, J.; Zhao, J.-L.; Liu, Y.-S.; Ying, G.-G. Rapid Target and Non-Target Screening Method for Determination of Emerging Organic Chemicals in Fish. *J. Chromatogr. A* **2022**, *1676*, 463185.

- <https://doi.org/10.1016/j.chroma.2022.463185>.
- (31) Roos, A. M.; Gamberg, M.; Muir, D.; Kärrman, A.; Carlsson, P.; Cuyler, C.; Lind, Y.; Bossi, R.; Rigét, F. Perfluoroalkyl Substances in Circum-Arctic Rangifer: Caribou and Reindeer. *Environ. Sci. Pollut. Res.* **2022**, *29* (16), 23721–23735. <https://doi.org/10.1007/s11356-021-16729-7>.
- (32) Grenier, P.; Elliott, J. E.; Drouillard, K. G.; Guigueno, M. F.; Muir, D.; Shaw, D. P.; Wayland, M.; Elliott, K. H. Long-Range Transport of Legacy Organic Pollutants Affects Alpine Fish Eaten by Ospreys in Western Canada. *Sci. Total Environ.* **2020**, *712*, 135889. <https://doi.org/10.1016/j.scitotenv.2019.135889>.
- (33) Jiang, T.; Wang, M.; Wang, A.; Abrahamsson, D.; Kuang, W.; Morello-Frosch, R.; Park, J.-S.; Woodruff, T. J. Large-Scale Implementation and Flaw Investigation of Human Serum Suspect Screening Analysis for Industrial Chemicals. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (9), 2425–2435. <https://doi.org/10.1021/jasms.1c00135>.
- (34) Woodruff, T. J.; Zota, A. R.; Schwartz, J. M. Environmental Chemicals in Pregnant Women in the United States: NHANES 2003–2004. *Environ. Health Perspect.* **2011**, *119* (6), 878–885. <https://doi.org/10.1289/ehp.1002727>.
- (35) Liu, Y.; Li, A.; Buchanan, S.; Liu, W. Exposure Characteristics for Congeners, Isomers, and Enantiomers of Perfluoroalkyl Substances in Mothers and Infants. *Environ. Int.* **2020**, *144*, 106012. <https://doi.org/10.1016/j.envint.2020.106012>.
- (36) Peng, F.-J.; Emond, C.; Hardy, E. M.; Sauvageot, N.; Alkerwi, A.; Lair, M.-L.; Appenzeller, B. M. R. Population-Based Biomonitoring of Exposure to Persistent and Non-Persistent Organic Pollutants in the Grand Duchy of Luxembourg: Results from Hair Analysis. *Environ. Int.* **2021**, *153*, 106526. <https://doi.org/10.1016/j.envint.2021.106526>.
- (37) Peng, F.-J.; Hardy, E. M.; Mezzache, S.; Bourokba, N.; Palazzi, P.; Stojiljkovic, N.; Bastien, P.; Li, J.; Soeur, J.; Appenzeller, B. M. R. Exposure to Multiclass Pesticides among Female Adult Population in Two Chinese Cities Revealed by Hair Analysis. *Environ. Int.* **2020**, *138*, 105633. <https://doi.org/10.1016/j.envint.2020.105633>.
- (38) Hardy, E. M.; Dereumeaux, C.; Guldner, L.; Briand, O.; Vandentorren, S.; Oleko, A.; Zaros, C.; Appenzeller, B. M. R. Hair versus Urine for the Biomonitoring of Pesticide Exposure: Results from a Pilot Cohort Study on Pregnant Women. *Environ. Int.* **2021**, *152*, 106481. <https://doi.org/10.1016/j.envint.2021.106481>.
- (39) Yusa, V.; Millet, M.; Coscolla, C.; Pardo, O.; Roca, M. Occurrence of Biomarkers of Pesticide Exposure in Non-Invasive Human Specimens. *Chemosphere* **2015**, *139*, 91–108. <https://doi.org/10.1016/j.chemosphere.2015.05.082>.
- (40) Bonmatin, J.-M.; Mitchell, E. A. D.; Glauser, G.; Lumawig-Heitzman, E.; Claveria, F.; Bijleveld van Lexmond, M.; Taira, K.; Sánchez-Bayo, F. Residues of Neonicotinoids in Soil, Water and People's Hair: A Case Study from Three Agricultural Regions of the Philippines. *Sci. Total Environ.* **2021**, *757*, 143822. <https://doi.org/10.1016/j.scitotenv.2020.143822>.
- (41) Polledri, E.; Mercadante, R.; Nijssen, R.; Consonni, D.; Mol, H.; Fustinoni, S. Hair as a Matrix to Evaluate Cumulative and Aggregate Exposure to Pesticides in Winegrowers. *Sci. Total Environ.* **2019**, *687*, 808–816. <https://doi.org/10.1016/j.scitotenv.2019.06.061>.
- (42) Lemke, N.; Murawski, A.; Schmied-Tobies, M. I. H.; Rucic, E.; Hoppe, H.-W.; Conrad, A.; Kolossa-Gehring, M. Glyphosate and Aminomethylphosphonic Acid (AMPA) in Urine of Children and Adolescents in Germany – Human Biomonitoring Results of the German Environmental Survey 2014–2017 (GerES V). *Environ. Int.* **2021**, *156*, 106769. <https://doi.org/10.1016/j.envint.2021.106769>.

- (43) Martins, C.; Vidal, A.; De Boevre, M.; De Saeger, S.; Nunes, C.; Torres, D.; Goios, A.; Lopes, C.; Assunção, R.; Alvito, P. Exposure Assessment of Portuguese Population to Multiple Mycotoxins: The Human Biomonitoring Approach. *Int. J. Hyg. Environ. Health* **2019**, *222* (6), 913–925. <https://doi.org/10.1016/j.ijheh.2019.06.010>.
- (44) Wongta, A.; Sawarng, N.; Tongchai, P.; Sutan, K.; Kerdnoi, T.; Prapamontol, T.; Hongsibsong, S. The Pesticide Exposure of People Living in Agricultural Community, Northern Thailand. *J. Toxicol.* **2018**, *2018*, e4168034. <https://doi.org/10.1155/2018/4168034>.
- (45) Savvaides, T.; Koelmel, J. P.; Zhou, Y.; Lin, E. Z.; Stelben, P.; Aristizabal-Henao, J. J.; Bowden, J. A.; Godri Pollitt, K. J. Prevalence and Implications of Per- and Polyfluoroalkyl Substances (PFAS) in Settled Dust. *Curr. Environ. Health Rep.* **2021**, *8* (4), 323–335. <https://doi.org/10.1007/s40572-021-00326-4>.
- (46) Besis, A.; Botsaropoulou, E.; Balla, D.; Voutsas, D.; Samara, C. Toxic Organic Pollutants in Greek House Dust: Implications for Human Exposure and Health Risk. *Chemosphere* **2021**, *284*, 131318. <https://doi.org/10.1016/j.chemosphere.2021.131318>.
- (47) Yang, J.; Ching, Y. C.; Kadokami, K. Occurrence and Exposure Risk Assessment of Organic Micropollutants in Indoor Dust from Malaysia. *Chemosphere* **2022**, *287*, 132340. <https://doi.org/10.1016/j.chemosphere.2021.132340>.
- (48) Sen, P.; Qadri, S.; Luukkonen, P. K.; Ragnarsdottir, O.; McGlinchey, A.; Jäntti, S.; Juuti, A.; Arola, J.; Schlezinger, J. J.; Webster, T. F.; Orešič, M.; Yki-Järvinen, H.; Hyötyläinen, T. Exposure to Environmental Contaminants Is Associated with Altered Hepatic Lipid Metabolism in Non-Alcoholic Fatty Liver Disease. *J. Hepatol.* **2022**, *76* (2), 283–293. <https://doi.org/10.1016/j.jhep.2021.09.039>.
- (49) Silva de Carvalho, T. G.; Tavares, N. H. C.; Bastos, M. L. A.; Rodrigues de Oliveira, B. B.; Araújo, L. F.; Ferreira, M. J. M. Exposure to Chemical and Biological Agents at Work and Cardiovascular Disease in Brazil: A Population-Based Study. *J. Occup. Environ. Med.* **2021**, *63* (6), e341. <https://doi.org/10.1097/JOM.0000000000002210>.
- (50) Caballero, M.; Amiri, S.; Denney, J. T.; Monsivais, P.; Hystad, P.; Amram, O. Estimated Residential Exposure to Agricultural Chemicals and Premature Mortality by Parkinson's Disease in Washington State. *Int. J. Environ. Res. Public Health* **2018**, *15* (12), 2885. <https://doi.org/10.3390/ijerph15122885>.
- (51) Polemi, K. M.; Nguyen, V. K.; Heidt, J.; Kahana, A.; Jolliet, O.; Colacino, J. A. Identifying the Link between Chemical Exposures and Breast Cancer in African American Women via Integrated in Vitro and Exposure Biomarker Data. *Toxicology* **2021**, *463*, 152964. <https://doi.org/10.1016/j.tox.2021.152964>.
- (52) Videnros, C.; Selander, J.; Wiebert, P.; Albin, M.; Plato, N.; Borgquist, S.; Manjer, J.; Gustavsson, P. Postmenopausal Breast Cancer and Occupational Exposure to Chemicals. *Scand. J. Work. Environ. Health* **2019**, *45* (6), 642–650.
- (53) Ball, N.; Teo, W.-P.; Chandra, S.; Chapman, J. Parkinson's Disease and the Environment. *Front. Neurol.* **2019**, *10*.
- (54) Paul, K. C.; Sinsheimer, J. S.; Rhodes, S. L.; Cockburn, M.; Bronstein, J.; Ritz, B. Organophosphate Pesticide Exposures, Nitric Oxide Synthase Gene Variants, and Gene–Pesticide Interactions in a Case–Control Study of Parkinson's Disease, California (USA). *Environ. Health Perspect.* **2016**, *124* (5), 570–577. <https://doi.org/10.1289/ehp.1408976>.
- (55) Ahmed, F.; Tschärke, B.; O'Brien, J.; Thompson, J.; Samanipour, S.; Choi, P.; Li, J.; Mueller, J. F.; Thomas, K. Wastewater-Based Estimation of the Prevalence of

- Gout in Australia. *Sci. Total Environ.* **2020**, *715*, 136925.  
<https://doi.org/10.1016/j.scitotenv.2020.136925>.
- (56) Richmond, E. K.; Rosi, E. J.; Walters, D. M.; Fick, J.; Hamilton, S. K.; Brodin, T.; Sundelin, A.; Grace, M. R. A Diverse Suite of Pharmaceuticals Contaminates Stream and Riparian Food Webs. *Nat. Commun.* **2018**, *9* (1), 4491.  
<https://doi.org/10.1038/s41467-018-06822-w>.
- (57) Tian, Z.; Zhao, H.; Peter, K. T.; Gonzalez, M.; Wetzel, J.; Wu, C.; Hu, X.; Prat, J.; Mudrock, E.; Hettinger, R.; Cortina, A. E.; Biswas, R. G.; Kock, F. V. C.; Soong, R.; Jenne, A.; Du, B.; Hou, F.; He, H.; Lundeen, R.; Gilbreath, A.; Sutton, R.; Scholz, N. L.; Davis, J. W.; Dodd, M. C.; Simpson, A.; McIntyre, J. K.; Kolodziej, E. P. A Ubiquitous Tire Rubber–Derived Chemical Induces Acute Mortality in Coho Salmon. *Science* **2020**.  
<https://doi.org/10.1126/science.abd6951>.
- (58) McCallum, E. S.; Krutzmann, E.; Brodin, T.; Fick, J.; Sundelin, A.; Balshine, S. Exposure to Wastewater Effluent Affects Fish Behaviour and Tissue-Specific Uptake of Pharmaceuticals. *Sci. Total Environ.* **2017**, *605–606*, 578–588.  
<https://doi.org/10.1016/j.scitotenv.2017.06.073>.
- (59) Gessner, M. O.; Tlili, A. Fostering Integration of Freshwater Ecology with Ecotoxicology. *Freshw. Biol.* **2016**, *61* (12), 1991–2001.  
<https://doi.org/10.1111/fwb.12852>.
- (60) Escher, B. I.; Stapleton, H. M.; Schymanski, E. L. Tracking Complex Mixtures of Chemicals in Our Changing Environment. *Science* **2020**, *367* (6476), 388–392.  
<https://doi.org/10.1126/science.aay6636>.
- (61) *Guidance on Information Requirements and Chemical Safety Assessment - ECHA*.  
<https://echa.europa.eu/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment> (accessed 2022-10-22).
- (62) US EPA. *EPI Suite™-Estimation Program Interface*.  
<https://www.epa.gov/tsca-screening-tools/epi-suite-estimation-program-interface> (accessed 2022-10-22).
- (63) Moon, J.; Lee, B.; Ra, J.-S.; Kim, K.-T. Predicting PBT and CMR Properties of Substances of Very High Concern (SVHCs) Using QSAR Models, and Application for K-REACH. *Toxicol. Rep.* **2020**, *7*, 995–1000.  
<https://doi.org/10.1016/j.toxrep.2020.08.014>.
- (64) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA Models for Predicting Physicochemical Properties and Environmental Fate Endpoints. *J. Cheminformatics* **2018**, *10* (1), 10. <https://doi.org/10.1186/s13321-018-0263-1>.
- (65) Stempel, S.; Scherlinger, M.; Ng, C. A.; Hungerbühler, K. Screening for PBT Chemicals among the “Existing” and “New” Chemicals of the EU. *Environ. Sci. Technol.* **2012**, *46* (11), 5680–5687. <https://doi.org/10.1021/es3002713>.
- (66) European Chemicals Agency. *Assessment of regulatory needs (Bisphenols)*.  
[https://echa.europa.eu/documents/10162/3448017/GMT\\_109\\_Bisphenols\\_Report\\_public\\_23502\\_en.pdf/1bd5525c-432c-495d-9dab-d7806bf34312?t=1647590013566](https://echa.europa.eu/documents/10162/3448017/GMT_109_Bisphenols_Report_public_23502_en.pdf/1bd5525c-432c-495d-9dab-d7806bf34312?t=1647590013566) (accessed 2022-10-22).
- (67) Wang, Z.; Wiesinger, H.; Groh, K. Time to Reveal Chemical Identities of Polymers and UVCBs. *Environ. Sci. Technol.* **2021**.  
<https://doi.org/10.1021/acs.est.1c05620>.
- (68) *Registered substances information - ECHA*.  
<https://echa.europa.eu/information-on-chemicals/registered-substances/information> (accessed 2022-10-22).
- (69) *Substance Identification and Inventory Information. CAS*.  
<https://www.cas.org/support/documentation/regulated-chemicals/substlist>

- (accessed 2022-10-30).
- (70) *CosIng - Cosmetics - GROWTH - European Commission*. <https://ec.europa.eu/growth/tools-databases/cosing/index.cfm> (accessed 2022-10-30).
- (71) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. <https://doi.org/10.1021/ci00057a005>.
- (72) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminformatics* **2015**, *7* (1), 23. <https://doi.org/10.1186/s13321-015-0068-4>.
- (73) Krauss, M.; Singer, H.; Hollender, J. LC–High Resolution MS in Environmental Analysis: From Target Screening to the Identification of Unknowns. *Anal. Bioanal. Chem.* **2010**, *397* (3), 943–951. <https://doi.org/10.1007/s00216-010-3608-9>.
- (74) Little, J. L.; Cleven, C. D.; Brown, S. D. Identification of “Known Unknowns” Utilizing Accurate Mass Data and Chemical Abstracts Service Databases. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (2), 348–359. <https://doi.org/10.1007/s13361-010-0034-3>.
- (75) Purschke, K.; Zoell, C.; Leonhardt, J.; Weber, M.; Schmidt, T. C. Identification of Unknowns in Industrial Wastewater Using Offline 2D Chromatography and Non-Target Screening. *Sci. Total Environ.* **2020**, *706*, 135835. <https://doi.org/10.1016/j.scitotenv.2019.135835>.
- (76) Chiaia-Hernandez, A. C.; Schymanski, E. L.; Kumar, P.; Singer, H. P.; Hollender, J. Suspect and Nontarget Screening Approaches to Identify Organic Contaminant Records in Lake Sediments. *Anal. Bioanal. Chem.* **2014**, *406* (28), 7323–7335. <https://doi.org/10.1007/s00216-014-8166-0>.
- (77) Pouchet, M.; Debrauwer, L.; Klanova, J.; Price, E. J.; Covaci, A.; Caballero-Casero, N.; Oberacher, H.; Lamoree, M.; Damont, A.; Fenaille, F.; Vlaanderen, J.; Meijer, J.; Krauss, M.; Sarigiannis, D.; Barouki, R.; Le Bizec, B.; Antignac, J.-P. Suspect and Non-Targeted Screening of Chemicals of Emerging Concern for Human Biomonitoring, Environmental Health Studies and Support to Risk Assessment: From Promises to Challenges and Harmonisation Issues. *Environ. Int.* **2020**, *139*, 105545. <https://doi.org/10.1016/j.envint.2020.105545>.
- (78) 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.* **2018**, *52* (12), 6881–6894. <https://doi.org/10.1021/acs.est.7b06598>.
- (79) MassBank Consortium; NORMAN Association. *MassBank | MassBank Europe Mass Spectral DataBase*. <https://massbank.eu/MassBank/> (accessed 2022-11-01).
- (80) *MassBank of North America*. <https://mona.fiehnlab.ucdavis.edu/> (accessed 2022-11-01).
- (81) Mass Spectrometry Data Center. *NIST Libraries and Software*. <https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:start> (accessed 2022-10-30).
- (82) Mohammed Taha, H.; Aalizadeh, R.; Alygizakis, N.; Antignac, J.-P.; Arp, H. P. H.; Bade, R.; Baker, N.; Belova, L.; Bijlsma, L.; Bolton, E. E.; Brack, W.; Celma, A.; Chen, W.-L.; Cheng, T.; Chirsir, P.; Ćirka, L.; D’Agostino, L. A.; Djoumbou Feunang, Y.; Dulio, V.; Fischer, S.; Gago-Ferrero, P.; Galani, A.; Geueke, B.; Glowacka, N.; Glüge, J.; Groh, K.; Grosse, S.; Haglund, P.; Hakkinen, P. J.; Hale, S. E.; Hernandez, F.; Janssen, E. M.-L.; Jonkers, T.; Kiefer, K.; Kirchner, M.; Koschorreck, J.; Krauss, M.; Krier, J.; Lamoree, M. H.; Letzel, M.; Letzel, T.;



- Li, Q.; Little, J.; Liu, Y.; Lunderberg, D. M.; Martin, J. W.; McEachran, A. D.; McLean, J. A.; Meier, C.; Meijer, J.; Menger, F.; Merino, C.; Muncke, J.; Muschket, M.; Neumann, M.; Neveu, V.; Ng, K.; Oberacher, H.; O'Brien, J.; Oswald, P.; Oswaldova, M.; Picache, J. A.; Postigo, C.; Ramirez, N.; Reemtsma, T.; Renaud, J.; Rostkowski, P.; Rüdell, H.; Salek, R. M.; Samanipour, S.; Scheringer, M.; Schliebner, I.; Schulz, W.; Schulze, T.; Sengl, M.; Shoemaker, B. A.; Sims, K.; Singer, H.; Singh, R. R.; Sumarah, M.; Thiessen, P. A.; Thomas, K. V.; Torres, S.; Trier, X.; van Wezel, A. P.; Vermeulen, R. C. H.; Vlaanderen, J. J.; von der Ohe, P. C.; Wang, Z.; Williams, A. J.; Willighagen, E. L.; Wishart, D. S.; Zhang, J.; Thomaidis, N. S.; Hollender, J.; Slobodnik, J.; Schymanski, E. L. The NORMAN Suspect List Exchange (NORMAN-SLE): Facilitating European and Worldwide Collaboration on Suspect Screening in High Resolution Mass Spectrometry. *Environ. Sci. Eur.* **2022**, *34* (1), 104. <https://doi.org/10.1186/s12302-022-00680-6>.
- (83) NORMAN Network. *NORMAN Suspect List Exchange*. <https://www.norman-network.com/nds/SLE/> (accessed 2022-11-01).
- (84) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. In Silico Fragmentation for Computer Assisted Identification of Metabolite Mass Spectra. *BMC Bioinformatics* **2010**, *11* (1), 148. <https://doi.org/10.1186/1471-2105-11-148>.
- (85) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag Relunched: Incorporating Strategies beyond in Silico Fragmentation. *J. Cheminformatics* **2016**, *8*, 3. <https://doi.org/10.1186/s13321-016-0115-9>.
- (86) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information. *Nat. Methods* **2019**, *16* (4), 299. <https://doi.org/10.1038/s41592-019-0344-8>.
- (87) Wang, F.; Liigand, J.; Tian, S.; Arndt, D.; Greiner, R.; Wishart, D. S. CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Anal. Chem.* **2021**, *93* (34), 11692–11700. <https://doi.org/10.1021/acs.analchem.1c01465>.
- (88) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **2014**, *48* (4), 2097–2098. <https://doi.org/10.1021/es5002105>.
- (89) Vermeulen, R.; Schymanski, E. L.; Barabási, A.-L.; Miller, G. W. The Exposome and Health: Where Chemistry Meets Biology. *Science* **2020**, *367* (6476), 392–396. <https://doi.org/10.1126/science.aay3164>.
- (90) Escher, B. I.; Hackermüller, J.; Polte, T.; Scholz, S.; Aigner, A.; Altenburger, R.; Böhme, A.; Bopp, S. K.; Brack, W.; Busch, W.; Chadeau-Hyam, M.; Covaci, A.; Eisenträger, A.; Galligan, J. J.; Garcia-Reyero, N.; Hartung, T.; Hein, M.; Herberth, G.; Jahnke, A.; Kleinjans, J.; Klüver, N.; Krauss, M.; Lamoree, M.; Lehmann, I.; Luckenbach, T.; Miller, G. W.; Müller, A.; Phillips, D. H.; Reemtsma, T.; Rolle-Kampczyk, U.; Schüürmann, G.; Schwikowski, B.; Tan, Y.-M.; Trump, S.; Walter-Rohde, S.; Wambaugh, J. F. From the Exposome to Mechanistic Understanding of Chemical-Induced Adverse Effects. *Environ. Int.* **2017**, *99*, 97–106. <https://doi.org/10.1016/j.envint.2016.11.029>.
- (91) Oltermann, P. *Oder river: mystery of mass die-off of fish lingers as no toxic substances found*. The Guardian. <https://www.theguardian.com/environment/2022/aug/15/oder-river-mystery-of-mass-die-off-of-fish-lingers-as-toxic-substances-ruled-out> (accessed 2022-10-30).
- (92) Singh, R. R. S76 | LUXPHARMA | Pharmaceuticals Marketed in Luxembourg,

2021. <https://doi.org/10.5281/zenodo.4587356>.
- (93) Krier, J.; Singh, R. R.; Kondić, T.; Lai, A.; Diderich, P.; Zhang, J.; Thiessen, P. A.; Bolton, E. E.; Schymanski, E. L. Discovering Pesticides and Their TPs in Luxembourg Waters Using Open Cheminformatics Approaches. *Environ. Int.* **2022**, *158*, 106885. <https://doi.org/10.1016/j.envint.2021.106885>.
- (94) Schymanski, E. L.; Kondić, T.; Neumann, S.; Thiessen, P. A.; Zhang, J.; Bolton, E. E. Empowering Large Chemical Knowledge Bases for Exposomics: PubChemLite Meets MetFrag. *J. Cheminformatics* **2021**, *13* (1), 19. <https://doi.org/10.1186/s13321-021-00489-0>.
- (95) Anliker, S.; Loos, M.; Comte, R.; Ruff, M.; Fenner, K.; Singer, H. Assessing Emissions from Pharmaceutical Manufacturing Based on Temporal High-Resolution Mass Spectrometry Data. *Environ. Sci. Technol.* **2020**, *54* (7), 4110–4120. <https://doi.org/10.1021/acs.est.9b07085>.
- (96) Schymanski, E. L.; Singer, H. P.; Longrée, P.; Loos, M.; Ruff, M.; Stravs, M. A.; Ripollés Vidal, C.; Hollender, J. Strategies to Characterize Polar Organic Contamination in Wastewater: Exploring the Capability of High Resolution Mass Spectrometry. *Environ. Sci. Technol.* **2014**, *48* (3), 1811–1818. <https://doi.org/10.1021/es4044374>.
- (97) Mairinger, T.; Loos, M.; Hollender, J. Characterization of Water-Soluble Synthetic Polymeric Substances in Wastewater Using LC-HRMS/MS. *Water Res.* **2021**, *190*, 116745. <https://doi.org/10.1016/j.watres.2020.116745>.
- (98) Schinkel, L.; Lara-Martín, P. A.; Giger, W.; Hollender, J.; Berg, M. Synthetic Surfactants in Swiss Sewage Sludges: Analytical Challenges, Concentrations and per Capita Loads. *Sci. Total Environ.* **2022**, *808*, 151361. <https://doi.org/10.1016/j.scitotenv.2021.151361>.
- (99) *Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy.* <http://data.europa.eu/eli/dir/2000/60/2014-11-20/eng> (accessed 2022-11-01).
- (100) Rostkowski, P.; Haglund, P.; Aalizadeh, R.; Alygizakis, N.; Thomaidis, N.; Arandes, J. B.; Nizzetto, P. B.; Booi, P.; Budzinski, H.; Brunswick, P.; Covaci, A.; Gallampois, C.; Grosse, S.; Hindle, R.; Ipolyi, I.; Jobst, K.; Kaserzon, S. L.; Leonards, P.; Lestremau, F.; Letzel, T.; Magnér, J.; Matsukami, H.; Moschet, C.; Oswald, P.; Plassmann, M.; Slobodnik, J.; Yang, C. The Strength in Numbers: Comprehensive Characterization of House Dust Using Complementary Mass Spectrometric Techniques. *Anal. Bioanal. Chem.* **2019**, *411* (10), 1957–1977. <https://doi.org/10.1007/s00216-019-01615-6>.
- (101) Ulrich, E. M.; Sobus, J. R.; Grulke, C. M.; Richard, A. M.; Newton, S. R.; Strynar, M. J.; Mansouri, K.; Williams, A. J. EPA's Non-Targeted Analysis Collaborative Trial (ENTACT): Genesis, Design, and Initial Findings. *Anal. Bioanal. Chem.* **2019**, *411* (4), 853–866. <https://doi.org/10.1007/s00216-018-1435-6>.
- (102) Schymanski, E. L.; Ruttkies, C.; Krauss, M.; Brouard, C.; Kind, T.; Dührkop, K.; Allen, F.; Vaniya, A.; Verdegem, D.; Böcker, S.; Rousu, J.; Shen, H.; Tsugawa, H.; Sajed, T.; Fiehn, O.; Ghesquière, B.; Neumann, S. Critical Assessment of Small Molecule Identification 2016: Automated Methods. *J. Cheminformatics* **2017**, *9* (1), 22. <https://doi.org/10.1186/s13321-017-0207-1>.
- (103) 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.* **2017**, *51* (20), 11505–11512. <https://doi.org/10.1021/acs.est.7b02184>.
- (104) Helmus, R.; ter Laak, T. L.; van Wezel, A. P.; de Voogt, P.; Schymanski, E. L.

- PatRoom: Open Source Software Platform for Environmental Mass Spectrometry Based Non-Target Screening. *J. Cheminformatics* **2021**, *13* (1), 1. <https://doi.org/10.1186/s13321-020-00477-w>.
- (105) Peter, K. T.; Phillips, A. L.; Knolhoff, A. M.; Gardinali, P. R.; Manzano, C. A.; Miller, K. E.; Pristner, M.; Sabourin, L.; Sumarah, M. W.; Warth, B.; Sobus, J. R. Nontargeted Analysis Study Reporting Tool: A Framework to Improve Research Transparency and Reproducibility. *Anal. Chem.* **2021**, *93* (41), 13870–13879. <https://doi.org/10.1021/acs.analchem.1c02621>.
- (106) Lai, A.; Sachdev, N.; Clark, A.; McEwen, L.; Fernandez, M.; Okonski, A.; Sullivan, K.; Schymanski, E. In Silico Structure Elucidation & FAIR Information Management for Improved UVCB Assessment, 2022. <https://doi.org/10.5281/zenodo.6469349>.
- (107) Lai; Schymanski; Steinbeck. Algorithm for Automated Classification of Homologous Chemical Series, 2022. <https://doi.org/10.5281/zenodo.6491204>.
- (108) Oertel, A.; Maul, K.; Menz, J.; Kronsbein, A. L.; Sittner, D.; Springer, A.; Müller, A.-K.; Herbst, U.; Schlegel, K.; Schulte, A. *REACH Compliance: Data availability in REACH registrations Part 2: Evaluation of data waiving and adaptations for chemicals  $\geq 1000$  tpa (Series 64/2018)*. <https://www.umweltbundesamt.de/en/publikationen/reach-compliance-data-availability-in-reach> (accessed 2021-08-14).
- (109) Lai, A.; Clark, A. M.; Escher, B. I.; Fernandez, M.; McEwen, L. R.; Tian, Z.; Wang, Z.; Schymanski, E. L. The Next Frontier of Environmental Unknowns: Substances of Unknown or Variable Composition, Complex Reaction Products, or Biological Materials (UVCBs). *Environ. Sci. Technol.* **2022**, *56* (12), 7448–7466. <https://doi.org/10.1021/acs.est.2c00321>.
- (110) S. McLachlan, M.; Li, Z.; Jonsson, L.; Kaserzon, S.; W. O'Brien, J.; F. Mueller, J. Removal of 293 Organic Compounds in 15 WWTPs Studied with Non-Targeted Suspect Screening. *Environ. Sci. Water Res. Technol.* **2022**, *8* (7), 1423–1433. <https://doi.org/10.1039/D2EW00088A>.
- (111) Zumstein, M.; Battagliarin, G.; Kuenkel, A.; Sander, M. Environmental Biodegradation of Water-Soluble Polymers: Key Considerations and Ways Forward. *Acc. Chem. Res.* **2022**, *55* (16), 2163–2167. <https://doi.org/10.1021/acs.accounts.2c00232>.
- (112) Fenner, K.; Scheringer, M. The Need for Chemical Simplification As a Logical Consequence of Ever-Increasing Chemical Pollution. *Environ. Sci. Technol.* **2021**. <https://doi.org/10.1021/acs.est.1c04903>.
- (113) Schymanski, E. L.; Bolton, E. E. FAIR Chemical Structures in the Journal of Cheminformatics. *J. Cheminformatics* **2021**, *13* (1), 50. <https://doi.org/10.1186/s13321-021-00520-4>.
- (114) Brinkhaus, H. O.; Rajan, K.; Schaub, J.; Zielesny, A.; Steinbeck, C. Open Data and Algorithms for Open Science in AI-Driven Molecular Informatics. **2022**. <https://doi.org/10.26434/chemrxiv-2022-dgcm6>.
- (115) Sun, X.; Zhang, X.; Muir, D. C. G.; Zeng, E. Y. Identification of Potential PBT/POP-Like Chemicals by a Deep Learning Approach Based on 2D Structural Features. *Environ. Sci. Technol.* **2020**, *54* (13), 8221–8231. <https://doi.org/10.1021/acs.est.0c01437>.
- (116) Li, X.; Chevez, T.; De Silva, A. O.; Muir, D. C. G.; Kleywegt, S.; Simpson, A.; Simpson, M. J.; Jobst, K. J. Which of the (Mixed) Halogenated n-Alkanes Are Likely To Be Persistent Organic Pollutants? *Environ. Sci. Technol.* **2021**, *55* (23), 15912–15920. <https://doi.org/10.1021/acs.est.1c05465>.
- (117) Muir, D.; Zhang, X.; de Wit, C. A.; Vorkamp, K.; Wilson, S. Identifying Further Chemicals of Emerging Arctic Concern Based on 'in Silico' Screening of

- Chemical Inventories. *Emerg. Contam.* **2019**, *5*, 201–210.  
<https://doi.org/10.1016/j.emcon.2019.05.005>.
- (118) Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. *MSNovelist: De Novo Structure Generation from Mass Spectra*; preprint; Bioinformatics, 2021.  
<https://doi.org/10.1101/2021.07.06.450875>.
- (119) Peets, P.; Wang, W.-C.; MacLeod, M.; Breitholtz, M.; Martin, J. W.; Krueve, A. MS2Tox Machine Learning Tool for Predicting the Ecotoxicity of Unidentified Chemicals in Water by Nontarget LC-HRMS. *Environ. Sci. Technol.* **2022**.  
<https://doi.org/10.1021/acs.est.2c02536>.
- (120) Wang, Z.; Altenburger, R.; Backhaus, T.; Covaci, A.; Diamond, M. L.; Grimalt, J. O.; Lohmann, R.; Schäffer, A.; Scheringer, M.; Selin, H.; Soehl, A.; Suzuki, N. We Need a Global Science-Policy Body on Chemicals and Waste. *Science* **2021**, *371* (6531), 774–776. <https://doi.org/10.1126/science.abe9090>.
- (121) Wang, Z.; Adu-Kumi, S.; Diamond, M. L.; Guardans, R.; Harner, T.; Harte, A.; Kajiwara, N.; Klánová, J.; Liu, J.; Moreira, E. G.; Muir, D. C. G.; Suzuki, N.; Pinas, V.; Seppälä, T.; Weber, R.; Yuan, B. Enhancing Scientific Support for the Stockholm Convention's Implementation: An Analysis of Policy Needs for Scientific Evidence. *Environ. Sci. Technol.* **2022**, *56* (5), 2936–2949.  
<https://doi.org/10.1021/acs.est.1c06120>.
- (122) Kümmerer, K.; Dionysiou, D. D.; Olsson, O.; Fatta-Kassinos, D. Reducing Aquatic Micropollutants – Increasing the Focus on Input Prevention and Integrated Emission Management. *Sci. Total Environ.* **2019**, *652*, 836–850.  
<https://doi.org/10.1016/j.scitotenv.2018.10.219>.
- (123) Kümmerer, K.; Clark, J. H.; Zuin, V. G. Rethinking Chemistry for a Circular Economy. *Science* **2020**, *367* (6476), 369–370.  
<https://doi.org/10.1126/science.aba4979>.
- (124) Cousins, I. T.; Goldenman, G.; Herzke, D.; Lohmann, R.; Miller, M.; Ng, C. A.; Patton, S.; Scheringer, M.; Trier, X.; Vierke, L.; Wang, Z.; DeWitt, J. C. The Concept of Essential Use for Determining When Uses of PFASs Can Be Phased Out. *Environ. Sci. Process. Impacts* **2019**, *21* (11), 1803–1815.  
<https://doi.org/10.1039/C9EM00163H>.
- (125) van Dijk, J.; Flerlage, H.; Beijer, S.; Slootweg, J. C.; van Wezel, A. P. Safe and Sustainable by Design: A Computer-Based Approach to Redesign Chemicals for Reduced Environmental Hazards. *Chemosphere* **2022**, *296*, 134050. <https://doi.org/10.1016/j.chemosphere.2022.134050>.
- (126) Dulio, V.; Koschorreck, J.; van Bavel, B.; van den Brink, P.; Hollender, J.; Munthe, J.; Schlabach, M.; Aalizadeh, R.; Agerstrand, M.; Ahrens, L.; Allan, I.; Alygizakis, N.; Barcelo, D.; Bohlin-Nizzetto, P.; Boutroup, S.; Brack, W.; Bressy, A.; Christensen, J. H.; Cirka, L.; Covaci, A.; Derksen, A.; Deviller, G.; Dingemans, M. M. L.; Engwall, M.; Fatta-Kassinos, D.; Gago-Ferrero, P.; Hernández, F.; Herzke, D.; Hilscherová, K.; Hollert, H.; Junghans, M.; Kasprzyk-Hordern, B.; Keiter, S.; Kools, S. A. E.; Krueve, A.; Lambropoulou, D.; Lamoree, M.; Leonards, P.; Lopez, B.; López de Alda, M.; Lundy, L.; Makovinská, J.; Marigómez, I.; Martin, J. W.; McHugh, B.; Miège, C.; O'Toole, S.; Perkola, N.; Polesello, S.; Posthuma, L.; Rodriguez-Mozaz, S.; Roessink, I.; Rostkowski, P.; Ruedel, H.; Samanipour, S.; Schulze, T.; Schymanski, E. L.; Sengl, M.; Tarábek, P.; Ten Hulscher, D.; Thomaidis, N.; Togola, A.; Valsecchi, S.; van Leeuwen, S.; von der Ohe, P.; Vorkamp, K.; Vrana, B.; Slobodnik, J. The NORMAN Association and the European Partnership for Chemicals Risk Assessment (PARC): Let's Cooperate! *Environ. Sci. Eur.* **2020**, *32* (1), 100.  
<https://doi.org/10.1186/s12302-020-00375-w>.

# Erklärungen

## **Selbstständigkeitserklärung**

Ich erkläre, dass ich die vorliegende Arbeit selbstständig und unter Verwendung der angegebenen Hilfsmittel, persönlichen Mitteilungen und Quellen angefertigt habe.

Luxembourg, \_\_\_\_\_

\_\_\_\_\_  
Adelene Lai Shuen Lyn

# Curriculum Vitae

## Adelene Lai Shuen Lyn

Birth Date & Place: 30.06.1993, Kuala Lumpur (Malaysia)

ORCID: <https://orcid.org/0000-0002-2985-6473>

GitHub: <https://github.com/adelenelai>

Google Scholar:  
<https://scholar.google.com/citations?user=qofOnu8AAAAJ&hl=en>

## Education

03.2019-12.2022	<b>Doctoral Researcher (Cotutelle PhD)</b> University of Luxembourg, Luxembourg & Friedrich Schiller University Jena, Germany  Environmental Cheminformatics & Cheminformatics and Computational Metabolomics  Supervised by Assoc. Prof. Dr. Emma Schymanski & Prof. Dr. Christoph Steinbeck
09.2015-09.2018	<b>MSc. Environmental Sciences</b> ETH Zurich, Switzerland Major: Biogeochemistry and Pollutant Dynamics ETH Master's Scholarship
09.2011-05.2015	<b>B.A. <i>magna cum laude</i></b> Wellesley College, United States of America Major: Chemistry Gulick Fund Scholarship

## Publications

Lai, A., Schaub, J., Steinbeck, C. & Schymanski, E. L. (2022). A Cheminformatics Algorithm to Classify Homologous Series. *Journal of Cheminformatics*, 14, 85. DOI: 10.1186/s13321-022-00663-y

Sigmund, G., Ågerstrand, M., Brodin, T., Diamond, M.L., Erdelen, W.R., Evers, D.C., Lai, A., Rillig, M.C., Schäffer, A., Soehl, A., Torres, J.P.M., Wang, Z. & Groh, K. J.

(2022). Broaden chemicals scope in biodiversity targets. *Science*, 376, 6599, pp.1280-1280. DOI:10.1126/science.add3070

Lai, A., Clark, A. M., Escher, B. I., Fernandez, M., McEwen, L. R., Tian, Z., Wang, Z. & Schymanski, E. L. (2022). The Next Frontier of Environmental Unknowns: Substances of Unknown or Variable Composition, Complex Reaction Products, or Biological Materials (UVCBs). *Environmental Science & Technology*, 56, 12, 7448-7466. DOI:10.1021/acs.est.2c00321

Krier, J., Singh, R. R., Kondić, T., Lai, A., Diderich, P., Zhang, J., Thiessen, P. A., Bolton, E. B. & Schymanski, E. L. (2021). Discovering Pesticides and their Transformation Products in Luxembourg Waters using Open Cheminformatics Approaches. *Environmental International*, 158, 106885. DOI:10.1016/j.envint.2021.106885

Singh, R. R., Lai, A., Krier, J., Kondić, T., Diderich, P., & Schymanski, E. L. (2021). Occurrence and Distribution of Pharmaceuticals and their Transformation Products in Luxembourgish Surface Waters. *ACS Environmental Au*, 0, 0, pp. DOI:10.1021/acsenvironau.1c00008

Lai, A., Singh, R. R., Kovalova, L., Jaeggi, O., Kondić, T., & Schymanski, E. L. (2021). Retrospective non-target analysis to support regulatory water monitoring: from masses of interest to recommendations via in silico workflows. *Environmental Sciences Europe*, 33, 1, 1-21. DOI:10.1186/s12302-021-00475-1

Wang, Y., Lai, A., Latino, D., Fenner, K., Helbling, D. (2018). Evaluating the environmental parameters that determine aerobic biodegradation half-lives of pesticides in soil with a multivariable approach. *Chemosphere*, 209: 430-438. DOI:10.1016/j.chemosphere.2018.06.077

Meylan, G., Lai, A., Hensley, J., Stauffacher, M., Krütli, P. (2018) Solid waste management of small island developing states - The Case of the Seychelles: A systemic and collaborative study of Swiss and Seychellois students to support policy. *Environmental Science Pollution Research*, 25, 36: 1-14. DOI:10.1007/s11356-018-2139-3

Persson, L., Karlsson-Vinkhuyzen, S., Lai, A., Persson, A., Fick, S. (2017). The Globally Harmonized System of Classification and Labelling of Chemicals - explaining the implementation gap. *Sustainability*, 9, 12: 2176. DOI:10.3390/su9122176

Zhumaev, U., Lai A.S., Pobelov, I.V., Kuzume, A., Rudnev, A.V., Wandlowski, Th. (2014). Quantifying perchlorate adsorption on Au(111). *Electrochimica Acta*, 146: 112-118. DOI:10.1016/j.electacta.2014.09.013

## Oral Presentations

Lai, A., Sachdev, N., Clark, A. M., Fernandez, M., McEwen, L. R., Okonski, A., Sullivan, K., Schymanski, E. L. In silico Structure Elucidation & FAIR Information Management for Improved UVCB Assessment. (05/2022). Platform Oral Presentation, Society of Environmental Toxicology and Chemistry Europe Annual Meeting, Copenhagen, DK. DOI: 10.5281/zenodo.6469348



Lai, A., Schymanski, E. L., Steinbeck, C. Algorithm for Automated Classification of Homologous Chemical Series. (05/2022) Poster Presentation, 17th German Cheminformatics Conference, Garmisch- Partenkirchen, DE. DOI: 10.5281/zenodo.6491203

Lai, A., Latino, D. A. R. S., Fenner, K. Using EFSA Regulatory Data to Explore Pesticide Biodegradation Half-life Variability. (05/2017). Poster Presentation, Society of Environmental Toxicology and Chemistry Europe Annual Meeting, Brussels, BE.

## Teaching & Supervision

- Assisted in the MetFrag in Practice Lab, Exposome Bootcamp, Columbia University Mailman School of Public Health, September 2019.
- Supervised three Bachelor's students for Wellesley College Hive Internships, January 2021.

Luxembourg, \_\_\_\_\_

\_\_\_\_\_  
Adelene Lai Shuen Lyn

# Acknowledgements

My first thanks go to my supervisors, whose Open Science ethos inspired me throughout this PhD. Thank you Emma Schymanski for the opportunities, critical and timely feedback, and overall guidance. I am honoured to have the special privilege of being the first PhD of the Environmental Cheminformatics Group. To Christoph Steinbeck, thank you for the wisdom, support, and examples. I am fortunate to have been included in Caffeine, and to have enjoyed group life and other fun parts of science. To you both, I am grateful for the Cotutelle setup and opportunities to travel and present my work. I am also thankful to my thesis committee members, Steffen Neumann and Reinhard Schneider, in addition to Egon Willighagen and Michael Stelter for their interest in my work and participation in my defence. My special gratitude goes to Rudi Balling for his support and leadership.

To my external collaborators, thank you for being part of my journey. I acknowledge Lubomira Kovalova, Oliver Jäggi, Philippe Diderich, Jessy Krier, Alex Clark, Beate Escher, Leah McEwen, Marc Fernandez, Zhenyu Tian, and Zhanyun Wang for all the joint efforts. I extend my gratitude to Véronique Briche, Céline Lecarpentier, Magali Guillaume, Sarah Tippner, Alexander Schwarzkopf, and Franziska Feldkamp for their help navigating administrative hurdles throughout this Cotutelle PhD. Thanks also to the Bachelor's students whom I supervised as part of the Wellesley Hive Internship Programme 2021: Amy Liu, Nina Sachdev, and Wen Li Yau.

I remain greatly indebted to Randolph Singh, for his infinite patience, guidance, and mentorship from Day 1. To Maria Sorokina, for her thought-provoking questions and ideas, and for being a ceaseless fount of (quotable) wisdom, *merci!* A special thanks to Zhanyun for all the advice and opportunities to work together that have shaped me as a scientist.

For their support and friendly atmosphere, I thank the entire ECI group: Corey, Todor, Lorenzo, Bego, Anjana, Emma P., Dagny, Gianfranco, Hiba, Parviel, and Simone. Special mention to the MFN and BCM groups for the nice times in the office and after-work events.

To the Caffeine-addicts, thank you for the epic travels, hikes, coffee breaks, dinners, movie nights, karaoke/dance sessions, and group retreats. Kohulan, Mahnoor, Jonas, Otto, Michael, Isa, Aziz, Chandu, Nisha, Noura, Luiz, and Christian - you made this experience special and you know what you mean to me.

Thank you to my friends in Luxembourg, Switzerland, the US, and beyond: Chrys, Semra, Kartik, Nirvana, Andrew, Gargee, Soracha, Sarena, Amanda, Shahida, Michelle, Nelleke, Tomo, Eileen, Flory, Oscar, Choong, and many more.

Lastly, thank you to my family in Malaysia and Leo for their love and support.

This Cotutelle PhD was carried out at the Environmental Cheminformatics Group, Luxembourg Centre for Systems Biomedicine and the Cheminformatics and Computational Metabolomics Group, Friedrich Schiller University, with the support of the Luxembourg National Research Fund (A18/BM/12341006). Portions of this thesis were proofread by Jonas Schaub, Zhanyun Wang, and Randolph Singh, whom I gratefully acknowledge.