Critical assessment of the chemical space covered by LC-HRMS non-targeted analysis

Tobias Hulleman,[†,‖] Viktoriia Turkina,[*,†,‖] Jake W. O'Brien,[‡,†] Aleksandra Chojnacka,[†] Kevin V. Thomas,[‡] and Saer Samanipour[*,†,¶,§]

†*Van 't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, 1090 GD, Amsterdam, the Netherlands*

‡*Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, 20 Cornwall Street, Woolloongabba, QLD, 4102, Australia*

¶*UvA Data Science Center, University of Amsterdam, Amsterdam*

§*Queensland Alliance for Environmental Health Sciences (QAEHS), 20 Cornwall Street, Woolloongabba, QLD, 4102, Australia*

‖*Contributed equally to this work*

E-mail: v.turkina@uva.nl; s.samanipour@uva.nl

Abstract

Non-targeted analysis (NTA) has emerged as a valuable approach for comprehensive monitoring of chemicals of emerging concern (CECs) in the exposome. The NTA approach, theoretically, is able to identify compounds with diverse physicochemical properties and sources. Non-targeted analysis methods, even though generic and wide scoping, have been shown to have limitations in terms of their coverage of the chemical space, as the number of the identified chemicals in each sample is very low (e.g. $\leq 5\%$). Investigating the chemical space covered by each NTA assay is crucial for understanding the limitations and challenges associated with the workflow from experimental methods

to the data acquisition and data processing. In this review, we examined recent NTA studies published between 2017 and 2023 that employed liquid chromatography-high resolution mass spectrometry. The parameters used in each study were documented and reported chemicals at the confidence level 1 and 2 were retrieved. The chosen experimental setups and the quality of reporting were critically evaluated and discussed. The findings revealed that only around 2% of the estimated chemical space was covered by the NTA studies investigated. Little to no trend was found between the experimental setup and the observed coverage, due to the generic and wide scope of NTA studies. The limited coverage of chemical space by the NTA studies highlights the necessity for a more comprehensive approach in experimental and data processing setups to enable the exploration of a broader range of chemical space, with the ultimate goal of protecting human and environmental health. Recommendations to further explore a wider range of the chemical space were given.

Synopsis

The coverage of chemical space via non-target analysis studies and the impact of the experimental conditions on that is critically assessed

Introduction

The exposome is the measure of all the exposures, both chemical and non-chemical, of an individual in a lifetime and how those exposures relate to health[1]. The chemical space of exposome refers to the chemical space relevant to human and environmental exposure[2–4]. On the other hand, the chemical space generally refers to all possible organic structures, that are plausible from organic chemistry point of view[2,5]. Theoretical estimates of such structures suggests there are around 10^{60} unique structures with molecular weights less than 500 Da[5,6]. This theoretical chemical space incorporates both known and unknown unknowns[2,7] and may

include structures that can cause adverse effects depending on their exposure levels. In fact, when looking at the known unknowns (i.e. structures recorded in the chemical databases, but not initially known to be present in the sample), several of them have been shown to have adverse effects on environmental and human health[8–11].

Chemical prioritization has been one of the main means for dealing with the diversity of chemical space in the human and environmental exposome[3,12,13]. This consists of exploration of the literature for measured chemicals and their properties/toxicities as well as national/international chemical registries[14,15]. A combination of predicted properties and toxicity is used to rank chemicals in the databases based on their potential impact on the environment and human health[16]. Chemicals with a high potential of such impact are considered as chemicals of emerging concern (CECs)[17,18]. To facilitate chemical prioritization, several databases consisting of chemical structures, the associated physicochemical properties (both measured and predicted), and their biological activities have been made publicly available (e.g. PubChem,[19] NORMAN Databases,[20] and CompTox[14]). However, most of these known unknowns remain unmeasured in environmental and biological matrices due to difficulties associated with the inclusion of such a large number of chemicals in routine monitoring programs[11,13].

Non targeted analysis (NTA) combined with liquid chromatography coupled with high resolution mass spectrometry (LC-HRMS) is considered as one of the most comprehensive methods for the detection and identification of known and unknown unknowns in complex environmental and biological samples[21,22]. This approach utilizes a generic and wide scope strategy for the sample preparation and analysis to maximize the coverage of the chemical space of the sample[2,13,21,23–31]. This typically results in very large and complex datasets (e.g. 5 GB per sample) that must be pre-processed prior to the identification workflow[31–33]. The NTA data processing workflows include several steps from data conversion to library search and the confidence assessment of the candidate spectra[2,23,26–29]. Due to the complexity of such datasets and sheer size of the chemical databases, the NTA workflows are not very

sensitive and do not result in a high percentage of identified chromatographic features[34,35]. A more sensitive but less comprehensive data processing alternative is suspect screening where the chemicals of interest are known prior to the data processing workflow. This approach is more sensitive in terms of limits of detection but is unable to detect unknown unknowns[20,29,36]. These two strategies are commonly employed together for the screening of complex environmental and biological samples[23].

The NTA strategy, even though powerful, has not been widely accepted within the regulatory framework due to reproduciblity issues[30,34,37]. Recent studies have indicated that small changes in both experimental (e.g. data dependent vs data independent acquisition) and data processing parameters may result in different outcomes and thus conclusions[34,35]. Additionally, a recent study has postulated the potential impact of different experimental parameters on the measured chemical space[38]. In fact, the aforementioned issues with NTA assays have sparked a debate in the scientific community and have given start to a new wave of data processing tools development[25,39,40]. Additionally, several efforts have been put into better defining the much needed quality control and assurance for such experiments to be successful in detection and identification of the known and unknown unknowns in complex environmental samples, thus better understanding the coverage of the analyzed chemical space[23,38,41–44].

Several recently published reviews discuss in detail the impact of different steps on the chemical space coverage through different experimental approaches[2,23,26,27]. They cover both data processing and experimental parameters including study scope, sampling and sample treatment, instrumental conditions, data processing and treatment, and reporting. However, none of these reviews attempted to assess (i.e. quantify) the coverage of the identified chemical space reached by the already conducted NTA environmental studies. Quantification of the coverage of chemical space by an analytical method is not a trivial task. Theoretically, it can be quantified as the number of identified compounds in the given sample divided by the number of all compounds present in the chemical subspace of the sample. But

practically, this calculation is impossible, due to the complex chemical nature of samples and the number of unknown constituents. Nevertheless, the investigation of experimentally explored chemical space is highly relevant for the researchers to be aware of the limited coverage of the associated chemical space.

In this review, we aim to quantify the coverage of the identified chemical space by recent environmental studies and investigate the relationship between the selected experimental parameters and the explored chemical space. To quantify the covered chemical space via NTA, we collected all recent studies that perform NTA (not suspect screening) and reported levels 1 and 2[45], in terms of identification, structures. Additionally, we limited the scope of this study to semi-polar and polar chemicals analyzable with liquid chromatography coupled with high resolution mass spectrometry , resulting in a total of 57 papers. As an approximation of the chemical space the NORMAN SusDat database containing around 60k unique chemicals with available PubChem CIDs (compound ID number) was used (`https://www.NORMAN-network.com/nds/susdat/susdatSearchShow.php`). We collected a list of experimental and instrumental parameters, including sample preparation (i.e. storage and extraction conditions), chromatographic separation (e.g. eluents, gradient type, and injection volume), high resolution mass spectrometry settings (e.g. mass analyzer, data acquisition mode, and polarity), and data processing workflows (e.g. mass and retention time tolerance, retention time domain alignment and databases used for the search). We also noted any unreported parameters to identify the most commonly omitted settings. Furthermore, we extracted information on the scope of the studies and samples analyzed.

Finally, we estimated the coverage of chemical space explored by recent NTA studies by comparing the structures identified in these studies with the chemical space represented by the compounds in the NORMAN SusDat database, as shown in Figure 1. This figure provides an insight of the range of chemicals that may be present in environmental samples. To our knowledge, this is the first study "quantifying" the coverage of chemical space via NTA assays.

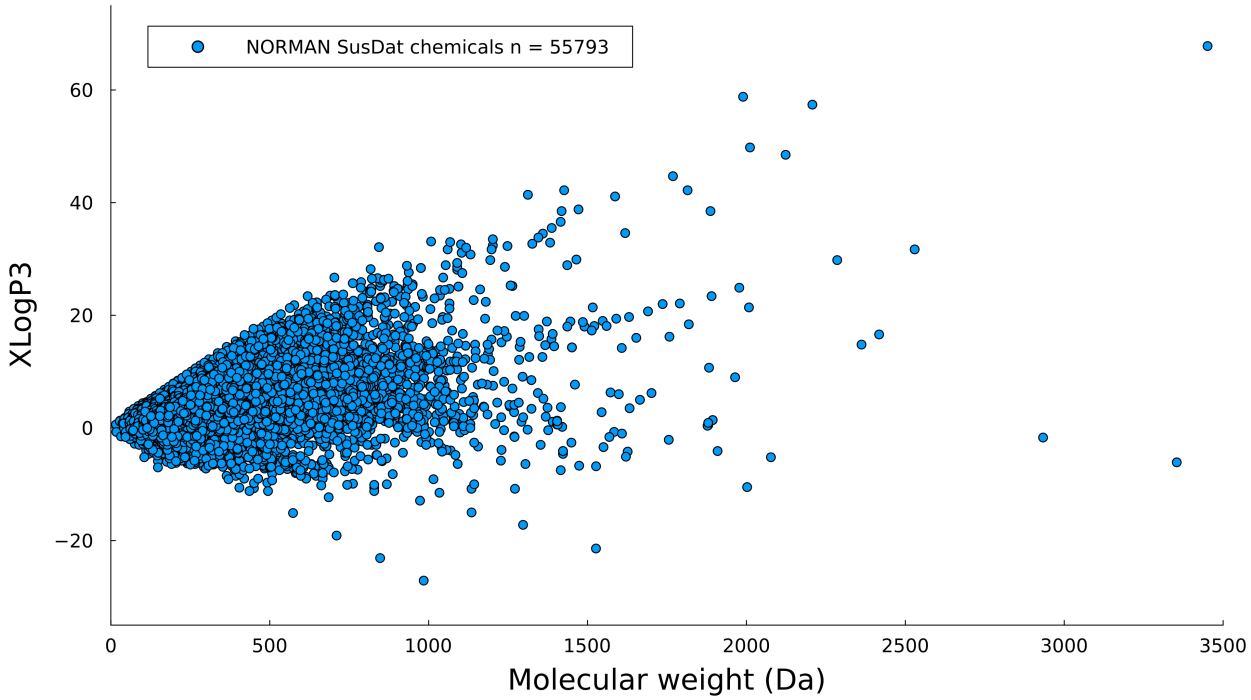

Figure 1: Distribution of all chemicals in the NORMAN SusDat database (n = 55793) based on their molecular weights (Da) and logP values.

Methods

Selection of NTA studies

This review is particularly focused on the development of the NTA approach in environmental studies, specifically after the discussions regarding reproducibility were initiated[39]. Thus, we used the citation database Scopus to search for relevant studies published from 2017 to 2023 in the field of non-target analysis (NTA) with a focus on environmental science. The search was limited to articles that contained the keywords "non targeted analysis", "non target analysis", "untargeted analysis", "untargeted screening", or "non-target screening" while excluding articles containing "metabolomics", "metabolic", or "gas chromatography". This initial search resulted in 377 publications adhering to the search parameters, which were then manually filtered to include only those that met a specific set of criteria.

The first criterion was that articles used non-target analysis to probe chemicals of emerging concern, preferably in environmental matrices. Secondly, the publications had to use a

non-target workflow. Some articles included the desired keywords in the title or abstract but were actually targeted studies with a very extensive list of target chemicals. For the same reason, the third criterion was that studies used LC-HRMS for sample analysis. Studies, which conducted GC-HRMS analysis were not included in the review, since such studies mainly employ suspect screening rather than non-targeted analysis. Furthermore, the recent development of NTA has been focused on LC rather than GC. Therefore, this review is focused on the coverage of the chemical space by NTA conducted via LC-HRMS. Additionally, direct infusion studies, studies that used rare setups, or heavily modified setups were excluded. Finally, review articles and studies which did not perform any identification were excluded as they did not contribute any additional methods or identified compounds. The search for relevant studies meeting these criteria was completed on March 1st, 2023, resulting in the inclusion of 61 studies in this review[46].

Collection of instrumental parameters

To capture the impact of each step of the NTA workflow on chemical space coverage, we extracted specific parameters used in the studies we reviewed. Sample preparation, chromatographic separation, data acquisition, and data processing were the four main steps where parameters were identified. Sample preparation parameters included the sample matrix, storage conditions, pre-storage modifications, extraction methods, and extraction conditions where applicable. Chromatographic separation parameters included the column used, eluent composition, gradient complexity, number of column volumes, column temperature, and injection volume. Gradients were classified as linear, semi-linear, or complex based on their complexity. The number of column volumes refers to the volume of solvent that passes through a chromatography column relative to the volume of the column itself. The calculation was performed using the equation 1.

$$Column\ volumes = \frac{F \times T\ run}{\pi \times (\frac{dc}{2})^2 \times L} \quad (1)$$

Where F is the flow rate (mL/min), $T\ run$ is the total run time of the method (min) - excluding equilibration time- dc is the internal diameter of the column (cm), and L is the length of the column in (cm). HRMS instrumental parameters included the mass analyzer, sampling rate (in the case of Q-TOF), resolution (in the case of Orbitrap), data acquisition mode, polarity, and mass range. Data processing parameters included mass tolerance, time domain alignment, mass calibration, retention time tolerance, databases used, and total database size (labeled as small if $\leq$1000 compounds or large if $>$1000 compounds). A summary of the collected parameters can be found in Figure 2. Furthermore, we made note of parameters that were not reported in order to identify which settings were commonly omitted. Lastly, we gathered information on the scope of the studies. The collected parameters along with the list of the publications are publicly available through this link[46].

Collection of reported structures

To assess the extent of chemical space coverage by recent NTA studies, we extracted the reported structures. To ensure the reliability and accuracy of our analysis, we only included structures identified with a high level of confidence (i.e levels one and two on the Schymanski scale), which is less susceptible to false positive identifications[45]. For each compound, SMILES, IUPAC name, and the regular names provided by the authors were extracted. Finally, we excluded articles from our chemical space coverage assessment if the authors did not specify the identification level, did not include the identified compounds in either the article or supplementary materials, or only reported compounds within their target list.

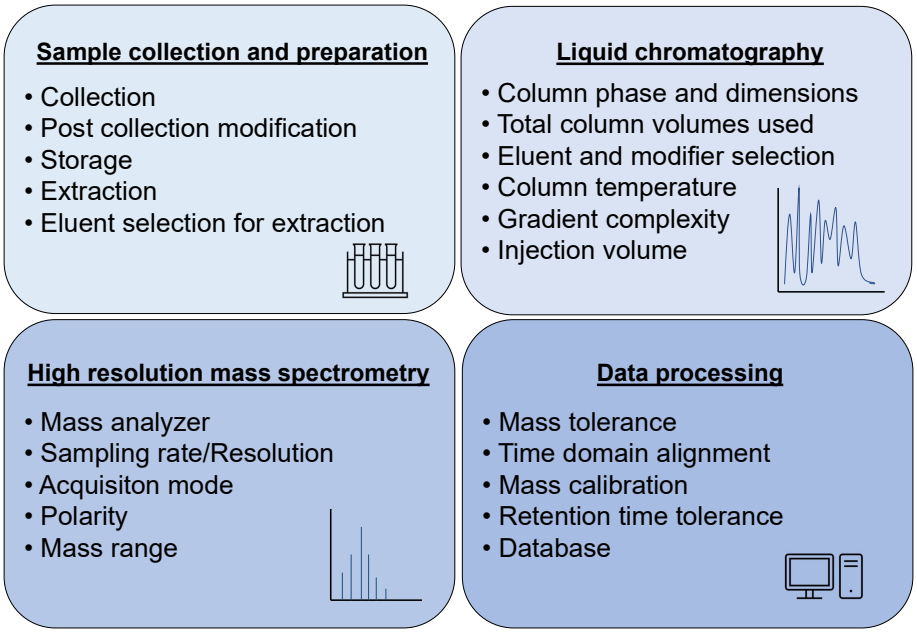

Figure 2: Summary of the main instrumental parameters collected from the reviewed NTA studies

Data processing

The list of the collected compounds was stored in CSV format, and Julia version 1.7 was used to import and process the data. A modified version of the PubChemCrawler.jl package was employed to retrieve chemical data such as XLogP3 and MW of the compounds from the PubChem database by using their available identifiers (SMILES, IUPAC, InChIKey, or regular name)[47]. logP values extracted from PubChem are generated using XlogP3 with an additive model starting from a reference compound[48]. Retrieved data along with the collected experimental parameters were combined into a dataset that included PubChem CIDs corresponding to the compounds, their logP values, molecular weights, and experimental parameters.

For the evaluation of the chemical space coverage, we additionally calculated elemental mass defects (EMD) of six elemental ratios (CO, CCl, CN, CS, CF, and CH) for each collected compound and the ones included in the NORMAN SusDat database[49]. EMD

values were used to cluster structurally similar compounds together and separate others, as they incorporate structural information and are used to compare compounds based on their elemental composition[50]. The combination of logP, MW, and EMDs was used for principal component analysis (PCA), which is an unsupervised algorithm for dimensional reduction combining variables into principal components[51]. This approach is able to identify trends and clusters in the data sets. Prior to the analysis the data was mean-centered and scaled to keep the initial weight of all variables comparable. PCA was performed using the ScikitLearn.jl julia package and in total three principal components were utilized.

The NORMAN SusDat database was used for the approximation of the chemical space of environmental samples. While the chemical space comprises both known and unknown compounds, it is practically impossible to include the latter in our approximations. The Norman SusDat database includes CECs that have either been detected in various environmental compartments or have been identified as potential CECs, providing a comprehensive set of chemicals with a wide coverage of physical and chemical properties, and structures[20].

Finally, the classes of the collected compounds were defined to illustrate the frequency of identification of specific classes. To obtain the class of each CEC, the corresponding InChIKey was used to generate information on superclasses, classes, and sub-classes of each compound via ClassyFire. ClassyFire divides a given chemical compound into classes based on its structural features (i.e. functional groups)[52].

Discussion

In this review, we estimated the coverage of the chemical space of environmental samples by investigating recent NTA studies. To evaluate the impact of selected workflow parameters on the coverage of chemical space, we collected information on these parameters (e.g. mass analyzer, data acquisition mode, ionization mode and size of the database used) from the studies[46]. The identified compounds were categorized into classes and their relative

frequency of occurrence was determined. XLogP3, MW, and EMDs were used to represent the vastness of the chemical space, approximated with the NORMAN SusDat Database. PCA was employed to illustrate the coverage of the space of chemicals detected in recent environmental studies.

Overview of the studies

In total, 61 studies were collected, with 55 of them published since 2020. Only studies using NTA were included, while those using screening or targeted approaches but claiming to be untargeted were excluded. This indicates that $\approx$ 90% of the reviewed studies were published in the last three years, yielding an average of more than 15 studies per year. In contrast, during the period from 2017 to 2020, only six of the selected studies were published. Therefore, the significant increase in the number of such studies in recent years reflects the successful applications of NTA workflows in exposome analysis. The scope of these studies varies, with 30 studies focusing on a wide range of chemicals and another 21 studies specifically targeting groups among which are per- and polyfluoroalkyl substances (PFAS), pesticides, pharmaceuticals, and illicit drugs. Such prior prioritization influences the choice of experimental setup. The remaining 10 studies focused on NTA workflow development, indicating a growing interest and the need for further advancements in this field.

Overview of selected parameters

Sample collection and preparation

The collection and preparation of samples in the non-targeted analysis (NTA) workflow can introduce potential sources of loss of chemical information. Issues such as ensuring sample representativeness (e.g. selecting appropriate grab or passive sampling techniques), addressing potential sample contamination, accounting for matrix effects, optimizing extraction methods for selectivity, and avoiding bias towards specific chemical groups are important

considerations in NTA[2,23,26]. These challenges may impact the accuracy and reliability of NTA results, potentially affecting the comprehensiveness and quality of the chemical information obtained from the analysis. Therefore, careful attention to sample collection and preparation steps are essential to minimize potential sources of bias and ensure robust and reliable NTA outcomes.

The majority of the collected studies (67%) analyzed water samples (n = 41). Other matrices that were investigated include biota (n = 5), dust (n = 3), urine (n = 3), atmospheric particulate matter (n = 2), paper (n = 2), serum (n = 2), blood (n = 1), human hair (n = 1), ovarian follicular fluid (n = 1), sewage sludge (n = 1), snow (n = 1), and surface soil (n = 1), turtle tissue (n = 1).

To prevent microbiological growth, the studies on water samples reported a conservation step, which involved either adding an acid or storing the sample at a temperature of -20°C or 4°C. Out of the 41 water studies, 7 studies either did not include a step to stop microbiological growth or did not report it. If this step was omitted, it could significantly alter the sample's final composition when it is eventually analyzed in the laboratory[53,54].

Around 54% of publications analyzing water included a sample filtering step prior to analysis. This step is a compromise to preserve the LC system and column but may lead to the loss of the chemicals adsorbed to the particle's surface. Approximately 67% of studies included solid phase extraction (SPE) in their sample preparation, out of which 73% used reversed-phase hydrophilic-lipophilic balance (HLB) SPE. However, only 29% of studies with SPE used acidic and/or basic modifiers in the extraction eluents. That implies that most studies using only HLB SPE are potentially leaving ionizable compounds on the sorbent and may exclude them from the analysis. The remaining studies employed alternative preptreatment techniques among which vacuum-assisted evaporation, centrifugation, liquid-liquid extraction (LLE), ultrasonic extraction as well as their combination. These choices are mostly dictated by the sample nature/matrix. There were three studies that performed no sample extraction and injected directly into the LC-MS with a higher injection volume[55–57].

While this protocol minimizes sample adulteration and keeps the sampling of chemical space more comprehensive, it can also pose a challenge to detection sensitivity due to the low analyte concentration[23].

Overall the sample collection and preparation section is well reported in the selected studies. However, many of the studies focused on analyzing a wide range of chemicals do not explore alternative extraction methods to ensure a more comprehensive coverage of the chemical space. This could result in a bias towards specific compounds, rather than capturing a more diverse set of chemicals.

Liquid chromatography

Chromatographic separation is employed to minimize sample complexity by spreading analytes across the time axis. This helps to reduce ion suppression (matrix effect) and provides additional information (retention time) for the identification of the analytes. The chemistry of the stationary phase along with the elution conditions affects the quality of separation and the type of analytes being retained. Thus, the selection of chromatographic conditions heavily influences the coverage of the chemical space of the sample[23].

The majority of NTA studies use conventional reverse-phase separation with a generic C18 column. Optimization of the separation includes proper selection of eluents and modifiers, including suitable elution power and gradient setup, to avoid co-elution and excessive or insufficient retention of chemicals[58]. A simple linear gradient of an aqueous phase and methanol or acetonitrile from low to high percentage is most widely accepted for the wide scope screening. This method proved its reproducibility across different scopes of the studies[26]. However, this strategy focuses on polar to semipolar compounds, potentially excluding very polar (i.e logP smaller than -2) and very hydrophobic substances (i.e. logP larger 6) from the comprehensive investigation of the chemical composition of samples[59]. To cover the polar part of the chemical space, orthogonal methods such as hydrophilic interaction chromatography (HILIC) become more popular while for hydrophobic volatile chemicals,

GC is a widely used technique[55,60,61]. Finally, to ensure the reproducibility and reliability of the studies parameters such as injection volume and column temperature should be properly reported[62].

More than 90% of the collected studies used a C18 column for the separation, among which almost all were endcapped with a column length of 50mm (20%), 100mm (49%), or 150mm (30%). Column diameters were either 2.1mm (78% of the studies), 3mm (16%), or 2mm (3%) with the particle diameter under 3.5 µm. Additionally, two different studies reported 4.6mm and 0.05mm column diameters. Although applying a simple gradient ensures higher reproducibility of the method, only half of the studies (replaced≈ 51%≈ 49%) used a linear gradient, while around 32% used a semi-linear gradient and the remaining (18 %) used a more complex type of gradient.

The median number of column volumes eluted in the studies is 15.9, with an interquartile range of 15.9. The use of a sufficient number of column volumes should ensure the complete elution of most hydrophobic compounds (high logP and MW) and the absence of carryover. The optimal number depends on the stationary phase, eluent power, and analytes themselves[63]. Nevertheless, the widely accepted hypothesis is that there is a linear relationship between logP and retention/number of column volumes used. The hypothesis is applied for the reverse phase mode with comparable C18 selectivity, similar gradients, and eluent composition[64,65]. However, our results do not indicate the presence of a linear relationship between the number of column volumes and logP of the chemicals, since no clear linear pattern could be identified between these parameters (Figure S1).

In addition, the column temperatures used were all slightly above room temperature which is favorable for repeatability and reproducibility[62]. 31% of publications used 40°C, 16% used 35°C, replaced1113% used 30°C, two studies held the column at 25°C, one at 20°C, one at 45°C and one at 50°C. About 29% of papers did not report the column temperature, which hinders the reproducibility of experiments.

Finally, 18% of the studies did not report the injection volume used. Injection volume

should not have a large effect on the final observed chemical space as they depend on the extraction method and efficiencies. Nevertheless, the success of the method's transfer depends on it. Most of the studies used either 5 (n = 17) or 10 µL (n = 13) injection volume, which is adequate when using SPE extraction. The remaining were spread across 1, 3, 4, 7, 20, 100, 140, and 660 µL.

To conclude, despite the rising discussion about reporting quality[23,66–68] chromatographic separation parameters in the collected studies were not always properly reported. Proper harmonized reporting ensures successful method transfer, whereas inconsistent reporting raises questions related to the reproducibility of the study, reliability of the results, and the possibility of retrospective studies. While the majority of the studies seek to comprehensively investigate the chemical composition of the samples, only approximately 10% employ an alternative to the conventional approach to analyze the samples. Lastly, the hypothetical linear trend between logP and retention was not confirmed, indicating the need for more sophisticated strategies for method development and optimization.

High resolution mass spectrometry

The Orbitrap and the quadrupole time of flight (QTOF) equipped with electrospray ionization (ESI) are the two most commonly used HRMS instruments in liquid chromatography-based (LC) NTA experiments. For complimentary analysis, it is recommended to perform separate experiments in both positive and negative modes[69]. The mass resolution of Orbitrap mass analyzers is generally higher than that of QTOF, but both can provide high-resolution mass spectra (Resolution $\geq 30,000$)[70].

In QTOF, resolution is determined by the architecture of the mass analyzer[71], while for Orbitrap, the resolving power depends on a user specified resolution. In the case of Orbitrap, the speed of scans is directly related to the spectral resolution. However, the increase in mass resolution is limited by the time required for scanning operations. For QTOF, a crucial parameter for data quality is the sampling speed, which is reported as

spectra per second in Hz. If the scan rate is too high, fewer ions are sampled, which can lead to a sensitivity issue. Conversely, if the scan rate is too low, fewer data points on the time axis are recorded, potentially causing missed detection of analytes eluting in a narrow time range[72].

MS/MS spectra for structure elucidation are recorded using either data-dependent acquisition (DDA) or data-independent acquisition (DIA). DDA mode records fragments of pre-selected precursor ions (which are chosen based on their abundance or via an inclusion list) while DIA mode fragments all precursor ions within a certain mass range. The latter is preferable for comprehensive investigations of complex samples. However, DDA mode is currently the preferred choice in environmental studies, partly due to the limited availability of processing tools for DIA files and also because the DIA experimental setup is not commonly employed with Orbitrap mass analyzers[26]. QTOF analyzers are more commonly used for DIA due to higher data acquisition rates.

Roughly, half of the collected studies (n = 31) utilized an Orbitrap mass analyzer, while the other half employed a QTOF mass analyzer. However, a significant proportion (approximately 74%) of the studies reported using DDA, which inherently limits their results to predefined ions. The scan rate for QTOF analyzers was mostly set at 4 Hz, although some studies operated at lower rates of 3, 2, or 1 Hz. Many studies using Orbitrap analyzers operated at a resolution of 70,000, while some studies used lower resolutions with a minimum of 35,000 and higher with a maximum of 240,000. Approximately 22% of the studies did not report either resolution or scan rate.

Less than half of the studies (around 42%) conducted separate experiments in positive and negative modes, utilizing multiple injections, different modifiers, and sometimes different columns, which is considered a more suitable scenario for achieving comprehensive coverage of chemical space. In approximately 30% of the studies, MS was operated only in positive mode. There were eleven publications where the analysis was reported in both modes, but the details were insufficient to determine if the experiment was performed simultaneously or separately

in both modes. In three other studies, an exclusively negative mode was used to prioritize a specific group of compounds of interest, such as PFAS[73–75], deliberately narrowing down the investigated chemical space. Finally, two of the reviewed studies employed simultaneous positive and negative ionization modes with formic acid as a modifier. This approach is not preferable for NTA given that acidic additives are not always the optimal for a negative ionization mode. Additionally, the acquired data becomes extremely complex and lacks quality for reliable and robust processing.

The selected mass range in the collected studies is between 50-1200 m/z, which is based on approximated chemical space covering the largest part. However, some studies set their maximum m/z at 1000 or lower, which leads to the exclusion of the part of chemical space with higher MW.

To conclude, despite recent advancements in DIA technology, DDA remains the predominant choice in the reviewed studies. However, the recommended approach for improved reproducibility and reliability of NTA studies, and to enhance coverage of chemical space in environmental and metabolomics research, is to acquire data in DIA mode for initial screening and then continue with DDA for individual feature identification[23,76–78]. Finally, in terms of reproducibility the lack of comprehensively reported information hinders method transfer and therefore it warrants actions towards a harmonized reporting strategy[66–68].

Data processing

Data processing is considered a major bottleneck in NTA workflows. It refers to a series of procedures that starts with the data conversion and ends with the feature identification[23]. One of the steps for reliable processing is the mass calibration, either external or internal. During this step the measured m/z values of known structures are compared against theoretical m/z values. These shifts/correction factors are applied to all mass channels, depending on the instrumental setup. This step ensures the quality of the spectra in terms of accurate mass measurement[79]. An inadequate mass calibration may result in false positive and

negative detections during the identification[80].

One of the last steps of CEC identification is the use of a database to relate the MS output to a known chemical structure. To proceed with the identification, experimental data undergoes pre-processing steps: data compression, to remove noise and blank peaks, feature detection, to find features in 3-dimensional data, componentization, to group fragments and isotopologues belonging to the same compound, and feature prioritization to reduce the number of irrelevant features[81]. Since most of the collected studies used vendor software for the latter four steps, which makes it almost impossible to retrieve the information of algorithms utilized, these parameters cannot be adequately discussed for their influence on the coverage of chemical space. For the identification of known unknowns, pre-processed data is compared with chemical databases and matched against references from available spectral libraries, utilizing a combination of features, retention time, accurate mass, and fragmentation pattern[45]. The mass tolerance is the initial parameter used for the candidates' list compilation. This parameter, along with the database used, heavily affects the results of the candidate search. The number of chemicals included in databases used in the evaluated studies differs from a few hundred structures in in-house libraries[82] to tens and hundreds of thousands in publicly available libraries[27] such as NORMAN,[20] MassBank[83] or PubChem[19]. These search algorithms result in a set of candidate structures that ultimately must be confirmed via either reference standard and/or an orthogonal method[45]. The retention tolerance is applied for level 1 confirmation employing either predicted or measured retention times.[23]

For the transparency and reproducibility of the method, proper reporting of applied setups for each data processing step is essential. Nevertheless, a significant part of the studies did not provide sufficient information to reproduce the results. Specifically, approximately 39% did not mention anything about mass calibration, while 25% reported that they performed calibration but did not describe the procedure. Only about 36% included a report on the mass calibration procedure. A large number of the papers (43%) also did not report

whether a retention alignment was performed. 34% did report the fact that a retention alignment was done but did not specify the algorithm that was used or provided the details on the parameters used. The remaining 23% of publications did report both the fact that one was performed and which algorithm was used.

In contrast, mass tolerance applied for the search was reported in almost all studies, around 95%. Among which around 76% used a mass tolerance for the database query of 5ppm, which is highly common in the NTA database search workflows. There were also studies that used a relatively high mass tolerance of 20 ppm, 17 ppm, or 10 ppm and some studies that used mass tolerances lower than 5 at 3ppm, 2ppm, and even 1ppm. Generally, the studies that were using the lower mass tolerances for the database search reported a higher resolution of the mass analyzer. However, there was no clear indication of whether the mass tolerance applied to formula assignment or structural identification. On the other hand, retention tolerances had much lower reporting rates as 45% of the studies did not include this information. The remaining studies used tolerances in a range between 0.1 min and 0.5 min. However, there are a few publications that used a wider tolerance, up to 1.8 min, which may result in a high false positive rate. Finally, approximately 9% did not report the databases used or referred to the software but not the databases that the software was using. The majority, 82%, used a total database size containing more than 5000 compounds, while only 5 studies used databases with less than one thousand compounds.

The data processing step is one of the main bottlenecks for the NTA approach and thus requires greater attention within the community. Nevertheless, the reporting quality needs improvement. Furthermore, it was found that around 70% of the identified chemicals are available in MassBank EU. That means that roughly 30% of the HRMS spectra acquired for the identified compounds have not been deposited in public databases such as MassBank. For NTA to reach its full potential, the expansion of publicly available spectral databases is vital for the improvement of the coverage of chemical space at the identification step.

Explored chemical space

The studies yielded a total of 2657 compounds reported in the identification level 1 up to 2b. The contribution to the total number from each study varies between 1 and 370, with a mean and median of 50 and 30, respectively (Figure S2). Among these, 1606 compounds were identified as unique structures, accounting for $\approx 60\%$ of the total number of retrieved compounds. This finding implies that around half of the overall variety of compounds were detected more than once in various environmental compartments. However, in 7 studies, there was no report of either identification level, or any identifiers, which hinders the retrieval of the compounds from these studies. The class of each collected CEC was obtained and displayed in Figure 3. The most commonly found compounds were benzoids, followed by organocyclic compounds and then organic acids and derivatives. The latter category, along with organohalogen compounds, constitutes PFAS, which have been of particular interest in recent years. The median molecular weights of compounds from SusDat were 239 Da and 261 Da for those collected from the studies, with a median XLogP3 of 3.2 for SusDat and 2.2 for collected compounds. Based on histograms in Figure 4, compounds with the most frequently occurring properties are being identified in recent NTA studies, which can be partially explained by the generalized experimental workflows with reverse phase C18 columns.

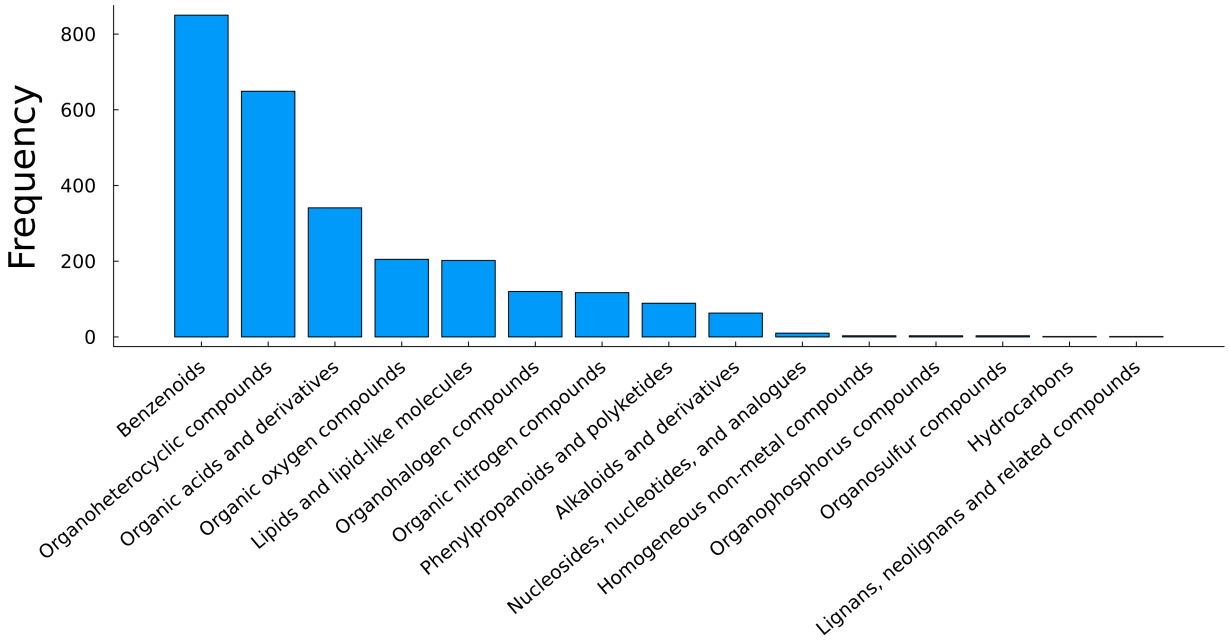

Figure 3: Histogram of all of the classes obtained from the Classyfire search for the detected CECs in reviewed studies

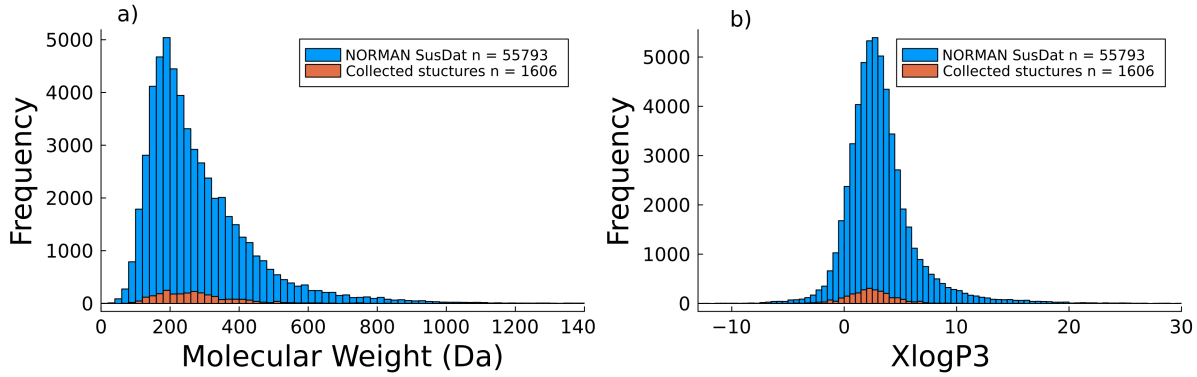

Figure 4: Molecular weights (a) and logP (b) distributions for the collected compounds (orange) and ones included in NORMAN SusDat database (blue).

Most of the compounds detected in the studies clustered closely together, with only a few compounds found further away from this main cluster (Figure 5). The collected compounds were analyzed in relation to their properties and plotted on a chemical space approximation represented by the NORMAN SusDat database. Figure 5 shows the plot in dimensions of

molecular weight (MW) and XLogP3, which emphasizes the limited space that is currently explored using current non-target analysis workflows. To examine the effect of some of the mass spectrometry (MS) parameters used on the explored chemical space, all compounds were plotted and clustered based on factors such as the mass analyzer used, acquisition mode, ionization mode, and the total database size used (Figure S3-S6). However, neither of these parameters showed an unambiguous influence on the coverage of the chemical space. It should be noted that the representation in MW and logP dimensions does not provide information about the elemental composition of compounds or their classes, which may result in an over-representation of the covered chemical space. Therefore, it is important to consider other parameters beyond MW and logP when evaluating the coverage of the chemical space by the collected structures.

The PCA scores plot in Figure 6 reveals that many regions of the chemical space are unexplored. The PCA was applied to the dataset combining the collected compounds with the ones from the NORMAN SusDat, with MW, XLogP3, and the EMDs as input variables. The first two principal components in the analysis were found to be primarily influenced by the elemental mass differences (EMDs) associated with compounds containing chlorine (Cl), fluorine (F), cyanide (CN), and sulfur (S). These EMDs represent the high variability in the elemental composition of the compounds and were identified as the most important variables in the PCA. This indicates that fewer compounds in the dataset contain halogens, nitrogen (N), and sulfur, while hydrogen (H), which is present in nearly every compound, does not contribute significantly to the variability in the data. The third principal component is primarily influenced by MW and XlogP3 (Figure 76). In total, the first three principal components explain 74% of the variance (Figure S8). In Figure S9-S11, the coverage of chemical space by different compound classes is displayed. Figure S11 specifically highlights the coverage by organic acids and derivatives as well as organohalogen compounds. The majority of PFAS, not exclusively, fall into these classes. The figures reveal that the distribution of compound classes across the chemical space is not homogeneous, suggesting

an over-representation of certain classes of compounds. This observation can be attributed to the prior prioritization of specific classes, which may bias the identification towards those classes of compounds.

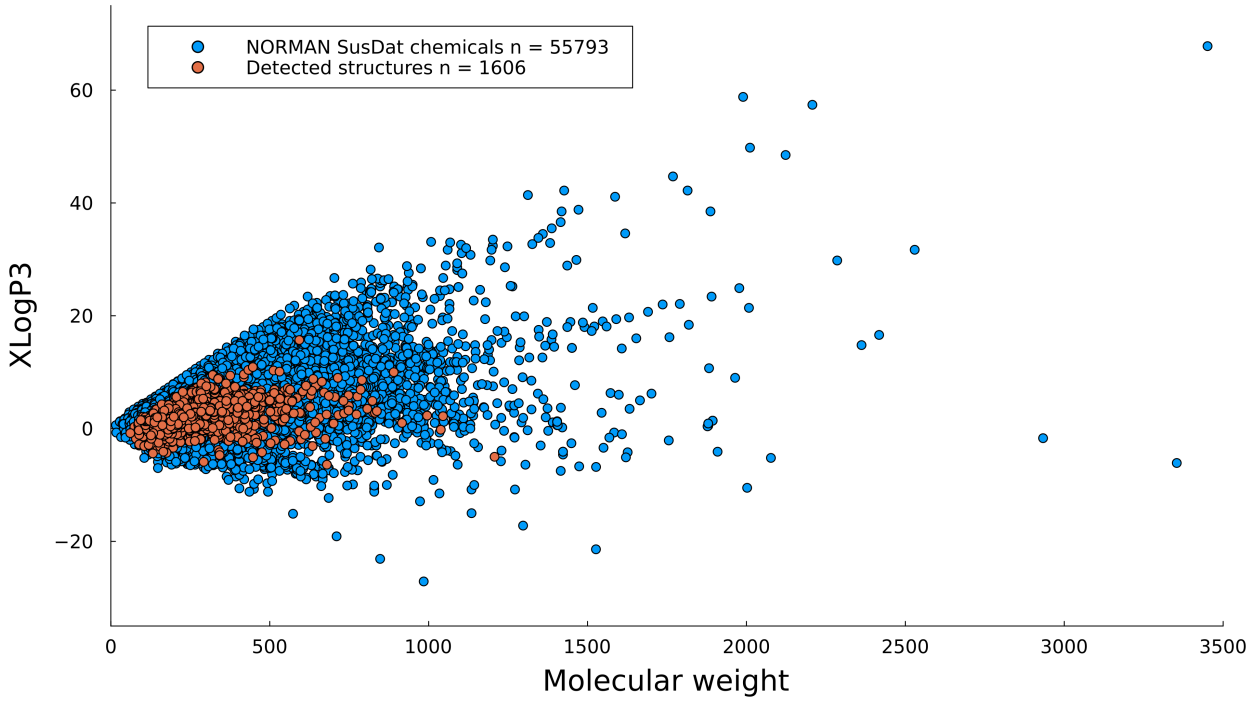

Figure 5: Distribution of all chemicals found in the reviewed articles at level 1 to 2b (orange) overlayed on NORMAN SusDat database chemicals (blue) based on their molecular weights and XlogP3 value

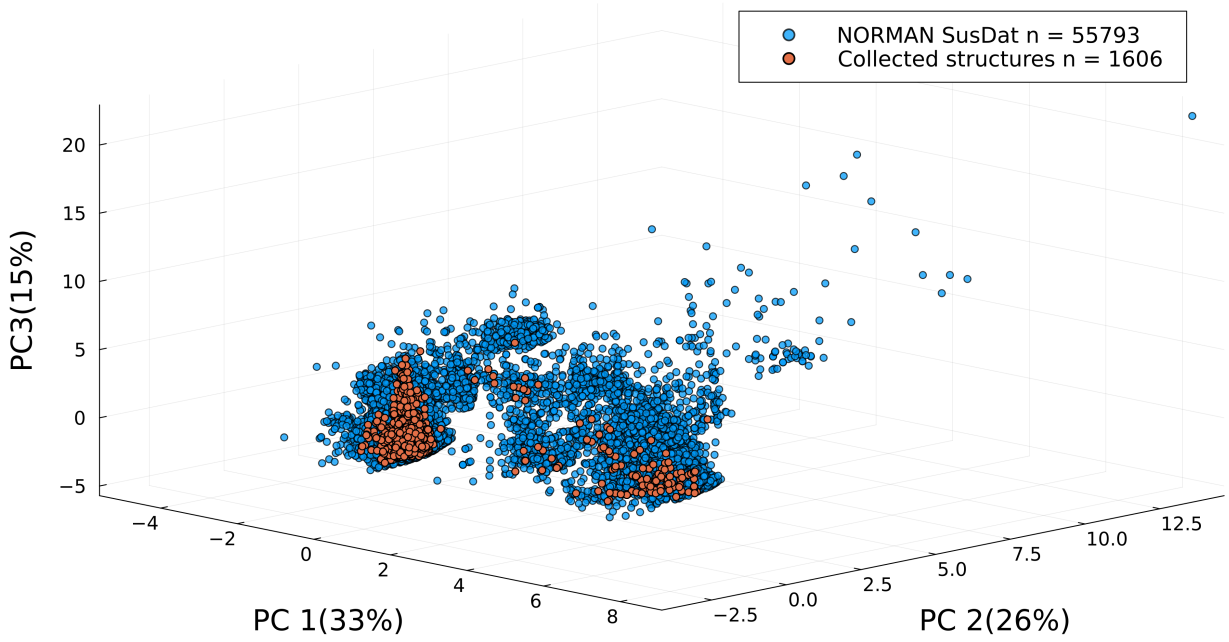

Figure 6: Scores plot of three principal components of the NORMAN SusDat database (blue) and the collected structures (orange).

Overall only around 2% of the estimated chemical space was covered by NTA studies investigated in this review. The coverage was defined as the number of unique structures retrieved from the reviewed studies versus the number of structures in NORMAN SusDat database as an approximation of chemical space. We used NORMAN SusDat as an approximation of chemical space as it contains a set of highly relevant and curated structures for environmental and human exposome. It should be noted that this is a small subspace of the total chemical space and serves as means of approximation for the true chemical space of exposome. No clear relationship between experimental conditions and coverage of the chemical space was discovered, which may indicate that the used experimental approaches are generic enough for the NTA assays. On the other hand, this may be caused by the lack of detailed and standardized reporting of the experimental conditions. Therefore, a more rigorous investigation of the parameters and standardization of reporting criteria has to be designed and performed. Although the most widely accepted properties of compounds such as logP and MW are widely used while discussing chemical space[23], in this study we showed

that they may not be the most relevant markers for assessing the coverage of chemical space. Finally, such a low coverage emphasizes the need for more comprehensive approaches to experimental and data processing workflows in order to explore a broader range of the chemical space and ultimately protect human and environmental health.

Recommendations and Outlook

Despite the ability of NTA to provide holistic information about the chemical composition of the samples, their true coverage of the chemical space has not been investigated. Furthermore, the NTA studies have suffered from issues related to their reproducibility, due to the complexity of both experimental and computational approaches employed in NTA assays. One of the main bottlenecks for a more reproducible NTA assay is the lack of standardization of the reporting criteria (including the experimental conditions). Our detailed investigation of the previously published NTA studies further suggests the need for such criteria. Minimum accepted experimental criteria and data processing parameters should be reported to ensure the transparency and reliability of the results. The utilization of harmonized reporting tools such as BP4NTA SRT or NORMAN suspect screening reporting tools can help the reproduciblity and transparency of future NTA studies[68,77,78]. This will potentially lead to the acceptance of the NTA approach by the regulatory bodies.

The potential coverage of the chemical space should be assessed during the design of the experimental setups. Most of the recent studies focused their experimental setups based on the conventional workflow including HLB SPE for sample preparation, reverse phase separation with C18 columns, and DDA acquisition mode, without considering alternative approaches. The best practice would be an application of alternative extraction methods, implementation of orthogonal techniques (e.g. RPLC and HILIC), DIA acquisition mode as the first screening approach, and the application of reliable/robust data processing tools, preferably open source/access. For the identification part of the workflow, the sharing of

experimental mass spectra of identified compounds along with their acquisition conditions is vital to the progress of the community. Additionally, archiving the raw data in public repositories for both the retrospective analysis as well as data processing tool development is highly essential.

To our knowledge, no other study has evaluated the coverage of the chemical space via NTA studies in such detail. However, due to the lack of standardized reporting criteria, the direct impact of different experimental choices on the covered chemical space could not be established. Also, our study is limited to the works published after 2017 and we only included studies with clear level 1 and 2 identification reporting. Furthermore, we excluded the suspect screening studies, which may result in an underestimation of the coverage of NTA studies. However, our study, even though limited, clearly shows the shortcomings of the current NTA practices and the need for further development in different areas - including experimental setup.

Acknowledgement

The authors are thankful to the members of Environmental Modeling & Computational Mass Spectrometry (www.emcms.info). S.S. and V.T. are thankful to the UvA Data Science Center and ChemistryNL TKI for the financial support (projects Edified and SCOPE). The authors thank Denice van Herwerden for her help in setting up the calculations of elemental mass defects. J.W.O. is the recipient of a National Health and Medical Research Council (NHMRC) Investigator Grant (EL12009209) funded by the Australian Government. K.V.T, and S.S. also acknowledge funding support from the Australian National Health and Medical Research Council (NHMRC; APP1185347). The Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland (UQ), gratefully acknowledges the financial support of Queensland Health. This research was funded in whole or in part by NHMRC Emerging Leadership Fellowship EL1 2009209. For the purposes of open access,

the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

Notes

Information retrieved in this study can be found at `https://doi.org/10.5281/zenodo.7774345`. References to the reviewed studies and collected experimental parameters are at All experimental parameters.xlsx. The script to perform the calculations is available at `https://github.com//tobihul//Code-for-Critical-assessment-of-covered-chemical-space-with-LC-HRMS-non-targeted-analysis`. PubChemCrawler package is available at`https://github.com/JuliaHealth/PubChemCrawler.jl`.

Supporting Information Available

The Supporting Information with figures (S1 - S10) showing the relationship between experimental parameters and the covered chemical space is available at `XXX`.

References

(1) Wild, C. P. Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. *Cancer Epidemiology, Biomarkers & Prevention* **2005**, *14*, 1847–1850.

(2) Milman, B. L.; Zhurkovich, I. K. The chemical space for non-target analysis. *Trends Analyt. Chem.* **2017**, *97*, 179–187.

(3) Wambaugh, J. F. et al. New approach methodologies for exposure science. *Curr. Opin. Toxicol.* **2019**, *15*, 76–92.

(4) Escher, B. I.; Stapleton, H. M.; Schymanski, E. L. Tracking complex mixtures of chemicals in our changing environment. *Science* **2020**, *367*, 388–392.

(5) Reymond, J. L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*, 722–730.

(6) Van Deursen, R.; Reymond, J. L. Chemical space travel. *ChemMedChem* **2007**, *2*, 636–640.

(7) Reymond, J. L.; Awale, M. Exploring chemical space for drug discovery using the chemical universe database. *ACS Chem. Neurosci.* **2012**, *3*, 649–657.

(8) McEachran, A. D.; Sobus, J. R.; Williams, A. J. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. *Anal. Bioanal. Chem.* **2017**, *409*, 1729–1735.

(9) Vermeulen, R.; Schymanski, E. L.; Barabási, A.-L.; Miller, G. W. The exposome and health: Where chemistry meets biology. *Science* **2020**, *367*, 392–396.

(10) Landrigan, P. J. et al. The Lancet Commission on pollution and health. *Lancet* **2018**, *391*, 462–512.

(11) Petrie, B.; Barden, R.; Kasprzyk-Hordern, B. A review on emerging contaminants in wastewaters and the environment: Current knowledge, understudied areas and recommendations for future monitoring. *Water Research* **2015**, *72*, 3–27.

(12) Samanipour, S.; O'Brien, J. W.; Reid, M. J.; Thomas, K. V.; Praetorius, A. From Molecular Descriptors to Intrinsic Fish Toxicity of Chemicals: An Alternative Approach to Chemical Prioritization. *Environ. Sci. Technol.* **2022**,

(13) Dulio, V.; van Bavel, B.; Brorström-Lundén, E.; Harmsen, J.; Hollender, J.; Schlabach, M.; Slobodnik, J.; Thomas, K.; Koschorreck, J. Emerging pollutants in the EU: 10 years of NORMAN in support of environmental policies and regulations. *Environ Sci Eur* **2018**, *30*.

(14) 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*.

(15) Richard, A. M. et al. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **2016**, *29*, 1225–1251.

(16) Tian, Z. et al. A ubiquitous tire rubber–derived chemical induces acute mortality in coho salmon. *Science* **2021**, *371*, 185–189.

(17) Sauvé, S.; Desrosiers, M. A review of what is an emerging contaminant. *Chem. Cent. J.* **2014**, *8*, 15.

(18) Maddela, N. R.; Ramakrishnan, B.; Kakarla, D.; Venkateswarlu, K.; Megharaj, M. Major contaminants of emerging concern in soils: a perspective on potential health risks. *RSC Advances* **2022**, *12*, 12396–12415.

(19) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, 623–633.

(20) Mohammed Taha, H. et al. 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*.

(21) Aceña, J.; Stampachiacchiere, S.; Pérez, S.; Barceló, D. Advances in liquid chromatography - High-resolution mass spectrometry for quantitative and qualitative environmental analysis. *Anal. Bioanal. Chem.* **2015**, *407*, 6289–6299.

(22) Picó, Y.; Barceló, D. Transformation products of emerging contaminants in the envi-

ronment and high-resolution mass spectrometry: A new horizon. *Anal Bioanal Chem* **2015**, *407*, 6257–6273.

(23) Schulze, B.; Jeon, Y.; Kaserzon, S.; Heffernan, A. L.; Dewapriya, P.; O'Brien, J.; Gomez Ramos, M. J.; Ghorbani Gorji, S.; Mueller, J. F.; Thomas, K. V.; Samanipour, S. An assessment of quality assurance/quality control efforts in high resolution mass spectrometry non-target workflows for analysis of environmental samples. *TrAC Trends in Analytical Chemistry* **2020**, *133*, 116063.

(24) Muir, D. C.; Howard, P. H. Are there other persistent organic pollutants? A challenge for environmental chemists. *Environ. Sci. Technol.* **2006**, *40*, 7157–7166.

(25) Samanipour, S.; Martin, J. W.; Lamoree, M. H.; Reid, M. J.; Thomas, K. V. Letter to the Editor: Optimism for Nontarget Analysis in Environmental Chemistry. *Environ. Sci. Technol.* **2019**, *53*, 5529–5530.

(26) Menger, F.; Gago-Ferrero, P.; Wiberg, K.; Ahrens, L. Wide-scope screening of polar contaminants of concern in water: A critical review of liquid chromatography-high resolution mass spectrometry-based strategies. *Trends Environ. Anal. Chem.* **2020**, *28*, e00102.

(27) 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*, 11505–11512.

(28) Zedda, M.; Zwiener, C. Is nontarget screening of emerging contaminants by LC-HRMS successful? A plea for compound libraries and computer tools. *Anal. Bioanal. Chem.* **2012**, *403*, 2493–2502.

(29) 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*, 943–951.

(30) McCord, J. P.; Groff, L. C.; Sobus, J. R. Quantitative non-targeted analysis: Bridging the gap between contaminant discovery and risk characterization. *Environ. Int.* **2022**, *158*, 107011.

(31) Schmidt, T. C. Recent trends in water analysis triggering future monitoring of organic micropollutants. *Anal. Bioanal. Chem.* **2018**, *410*, 3933–3941.

(32) Samanipour, S.; Reid, M. J.; Thomas, K. V. Statistical Variable Selection: An Alternative Prioritization Strategy during the Nontarget Analysis of LC-HR-MS Data. *Anal. Chem.* **2017**, *89*, 5585–5591.

(33) van Herwerden, D.; O'Brien, J. W.; Lege, S.; Pirok, B. W. J.; Thomas, K. V.; Samanipour, S. Cumulative Neutral Loss Model for Fragment Deconvolution in Electrospray Ionization High-Resolution Mass Spectrometry Data. *Anal. Chem.* **2023**,

(34) Hohrenk, L. L.; Itzel, F.; Baetz, N.; Tuerk, J.; Vosough, M.; Schmidt, T. C. Comparison of Software Tools for Liquid Chromatography-High-Resolution Mass Spectrometry Data Processing in Nontarget Screening of Environmental Samples. *Anal. Chem.* **2020**, *92*, 1898–1907.

(35) Samanipour, S.; Baz-Lomba, J. A.; Alygizakis, N. A.; Reid, M. J.; Thomaidis, N. S.; Thomas, K. V. Two stage algorithm vs commonly used approaches for the suspect screening of complex environmental samples analyzed via liquid chromatography high resolution time of flight mass spectroscopy: A test study. *J. Chromatograph A* **2017**, *1501*, 68–78.

(36) Hernández, F. et al. The role of analytical chemistry in exposure science: Focus on the aquatic environment. *Chemosphere* **2019**, *222*, 564–583.

(37) Brack, W.; Hollender, J.; de Alda, M. L.; Müller, C.; Schulze, T.; Schymanski, E.; Slobodnik, J.; Krauss, M. High-resolution mass spectrometry to complement monitor-

ing and track emerging chemicals and pollution trends in European water resources. *Environ Sci Eur* **2019**, *31*, 62.

(38) Black, G. et al. Exploring chemical space in non-targeted analysis: a proposed ChemSpace tool. *Anal. Bioanal. Chem.* **2023**, *415*, 35–44.

(39) Hites, R. A.; Jobst, K. J. Is Nontargeted Screening Reproducible? *Environ. Sci. Technol.* **2018**, *52*, 11975–11976.

(40) 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*, 853–866.

(41) Fisher, C. M.; Peter, K. T.; Newton, S. R.; Schaub, A. J.; Sobus, J. R. Approaches for assessing performance of high-resolution mass spectrometry–based non-targeted analysis methods. *Anal. Bioanal. Chem.* **2022**, *414*, 6455–6471.

(42) Knolhoff, A. M.; Premo, J. H.; Fisher, C. M. A proposed quality control standard mixture and its uses for evaluating nontargeted and suspect screening LC/HR-MS method performance. *Anal. Chem.* **2021**, *93*, 1596–1603.

(43) 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.

(44) Newton, S. R.; Sobus, J. R.; Ulrich, E. M.; Singh, R. R.; Chao, A.; McCord, J.; Laughlin-Toth, S.; Strynar, M. Examining NTA performance and potential using fortified and reference house dust as part of EPA's Non-Targeted Analysis Collaborative Trial (ENTACT). *Anal. Bioanal. Chem.* **2020**, *412*, 4221–4233.

(45) 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*, 2097–2098.

(46) Hulleman, T.; Turkina, V.; O'Brien, J.; Chojnacka, A.; Thomas, K. V.; Saminopour, S. Data files for: Critical assessment of covered chemical space with LC-HRMS non-targeted analysis. 2022; `https://doi.org/10.5281/zenodo.7774345`.

(47) Hulleman, T. Code for: Critical assessment of covered chemical space with LC-HRMS non-targeted analysis. 2023; `https://github.com/tobihul/Code-for-Critical-assessment-of-covered-chemical-space-with-LC-HRMS-non-targeted-analysis`.

(48) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *J. Chem. Inf. Model.* **2007**, *47*, 2140–2148.

(49) van Herwerden, D.; O'Brien, J. W.; Choi, P. M.; Thomas, K. V.; Schoenmakers, P. J.; Samanipour, S. Naive Bayes classification model for isotopologue detection in LC-HRMS data. *Chemom. Intell. Lab. Syst.* **2022**, *223*, 104515.

(50) Sleno, L. The use of mass defect in modern mass spectrometry. *J. Mass. Spectrom.* **2012**, *47*, 226–236.

(51) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*; Springer, 2009.

(52) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **2016**, *8*, 61.

(53) Kruve, A. Semi-quantitative non-target analysis of water with liquid chromatography/high-resolution mass spectrometry: How far are we? *Rapid Commun. Mass Spectrom.* **2019**, *33*, 54–63.

(54) Lyytikäinen, M.; Kukkonen, J. V.; Lydy, M. J. Analysis of pesticides in water and sediment under different storage conditions using gas chromatography. *Arch. Environ. Contam. Toxicol.* **2003**, *44*, 437–444.

(55) Been, F.; Kruve, A.; Vughs, D.; Meekel, N.; Reus, A.; Zwartsen, A.; Wessel, A.; Fischer, A.; ter Laak, T.; Brunner, A. M. Risk-based prioritization of suspects detected in riverine water using complementary chromatographic techniques. *Water Res.* **2021**, *204*, 117612.

(56) Hu, L. X.; Olaitan, O. J.; Li, Z.; Yang, Y. Y.; Chimezie, A.; Adepoju-Bello, A. A.; Ying, G. G.; Chen, C. E. What is in Nigerian waters? Target and non-target screening analysis for organic chemicals. *Chemosphere* **2021**, *284*, 131546.

(57) Köppe, T.; Jewell, K. S.; Dietrich, C.; Wick, A.; Ternes, T. A. Application of a non-target workflow for the identification of specific contaminants using the example of the Nidda river basin. *Water Research* **2020**, *178*, 115703.

(58) Kunzelmann, M.; Winter, M.; Åberg, M.; Hellenäs, K. E.; Rosén, J. Non-targeted analysis of unexpected food contaminants using LC-HRMS. *Anal. Bioanal. Chem.* **2018**, *410*, 5593–5602.

(59) Reemtsma, T.; Berger, U.; Arp, H. P. H.; Gallard, H.; Knepper, T. P.; Neumann, M.; Quintana, J. B.; Voogt, P. D. Mind the Gap: Persistent and Mobile Organic Compounds - Water Contaminants That Slip Through. *Environ. Sci. Technol.* **2016**, *50*, 10308–10315.

(60) Brüggen, S.; Schmitz, O. J. A New Concept for Regulatory Water Monitoring Via High-

Performance Liquid Chromatography Coupled to High-Resolution Mass Spectrometry. *J Anal Test* **2018**, *2*, 342–351.

(61) Badea, S. L.; Geana, E. I.; Niculescu, V. C.; Ionete, R. E. Recent progresses in analytical GC and LC mass spectrometric based-methods for the detection of emerging chlorinated and brominated contaminants and their transformation products in aquatic environment. *Sci. Total Environ.* **2020**, *722*, 137914.

(62) Greibrokk, T.; Andersen, T. High-temperature liquid chromatography. *J. Chromatogr. A* **2003**, *1000*, 743–755.

(63) Snyder, L. R.; Kirkland, J. J.; Dolan, J. W. *Introduction to Modern Liquid Chromatography*; John Wiley and Sons, 2010.

(64) Bade, R.; Bijlsma, L.; Sancho, J. V.; Hernández, F. Critical evaluation of a simple retention time predictor based on LogKow as a complementary tool in the identification of emerging contaminants in water. *Talanta* **2015**, *139*, 143–149.

(65) Kaliszan, R.; Haber, P.; Tomasz, B.; Siluk, D.; Valko, K. Lipophilicity and pKa estimates from gradient high-performance liquid chromatography. *J. Chromatogr. A* **2002**, *965*, 117–127.

(66) Phillips, A. L.; Peter, K. T.; Sobus, J. R.; Fisher, C. M.; Manzano, C. A.; McEachran, A. D.; Williams, A. J.; Knolhoff, A. M.; Ulrich, E. M. Standardizing non-targeted analysis reporting to advance exposure science and environmental epidemiology. *J. Expo. Sci. Environ. Epidemiol.* **2023**,

(67) 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*, 13870–13879.

(68) Place, B. J. et al. An Introduction to the Benchmarking and Publications for Non-Targeted Analysis Working Group. *Anal. Chem.* **2021**, *93*, 16289–16296.

(69) Malm, L.; Palm, E.; Souihi, A.; Plassmann, M.; Liigand, J.; Kruve, A. Guide to Semi-Quantitative Non-Targeted Screening Using LC/ESI/HRMS. *Molecules* **2021**, *26*, 3524.

(70) Zubarev, R. A.; Makarov, A. Orbitrap mass spectrometry. *Anal. Chem.* **2013**, *85*, 5288–5296.

(71) Boesl, U. Time-of-flight mass spectrometry: Introduction to the basics. *Mass Spectrom. Rev.* **2017**, *36*, 86–109.

(72) Gosetti, F.; Mazzucco, E.; Gennaro, M. C.; Marengo, E. Contaminants in water: non-target UHPLC/MS analysis. *Environ. Chem. Lett.* **2016**, *14*, 51–65.

(73) Jeong, Y.; Da Silva, K. M.; Iturrospe, E.; Fuiji, Y.; Boogaerts, T.; van Nuijs, A. L.; Koelmel, J.; Covaci, A. Occurrence and contamination profile of legacy and emerging per- and polyfluoroalkyl substances (PFAS) in Belgian wastewater using target, suspect and non-target screening approaches. *J. Hazard. Mater.* **2022**, *437*, 129378.

(74) Xia, X.; Zheng, Y.; Tang, X.; Zhao, N.; Wang, B.; Lin, H.; Lin, Y. Nontarget Identification of Novel Per- and Polyfluoroalkyl Substances in Cord Blood Samples. *Environ. Sci. Technol.* **2022**, *56*, 17061–17069.

(75) Yu, N.; Wen, H.; Wang, X.; Yamazaki, E.; Taniyasu, S.; Yamashita, N.; Yu, H.; Wei, S. Nontarget Discovery of Per- And Polyfluoroalkyl Substances in Atmospheric Particulate Matter and Gaseous Phase Using Cryogenic Air Sampler. *Environ. Sci. Technol.* **2020**, *54*, 3103–3113.

(76) Alygizakis, N. A.; Oswald, P.; Thomaidis, N. S.; Schymanski, E. L.; Aalizadeh, R.; Schulze, T.; Oswaldova, M.; Slobodnik, J. NORMAN digital sample freezing platform: A European virtual platform to exchange liquid chromatography high resolution-mass

spectrometry data and screen suspects in "digitally frozen" environmental samples. *TrAC Trends in Analytical Chemistry* **2019**, *115*, 129–137.

(77) Rostkowski, P. et al. The strength in numbers: comprehensive characterization of house dust using complementary mass spectrometric techniques. *Anal. Bioanal. Chem.* **2019**, *411*, 1957–1977.

(78) Alygizakis, N. A.; Samanipour, S.; Hollender, J.; Ibáñez, 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*, 5135–5144.

(79) Podwojski, K.; Fritsch, A.; Chamrad, D. C.; Paul, W.; Sitek, B.; Stühler, K.; Mutzel, P.; Stephan, C.; Meyer, H. E.; Urfer, W.; Ickstadt, K.; Rahnenführer, J. Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics* **2009**, *25*, 758–764.

(80) Clauser, K. R.; Baker, P.; Burlingame, A. L. Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **1999**, *71*, 2871–2882.

(81) Minkus, S.; Bieber, S.; Letzel, T. Spotlight on mass spectrometric non-target screening analysis: Advanced data processing methods recently communicated for extracting, prioritizing and quantifying features. *Anal. Sci. Adv.* **2022**, *3*, 103–112.

(82) Kleis, J. N.; Hess, C.; Germerott, T.; Roehrich, J. Sensitive Screening of New Psychoactive Substances in Serum Using Liquid Chromatography-Quadrupole Time-of-Flight Mass Spectrometry. *J. Anal. Toxicol.* **2022**, *46*, 592–599.

(83) Horai, H. et al. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass. Spectrom.* **2010**, *45*, 703–714.

TOC Graphic