JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

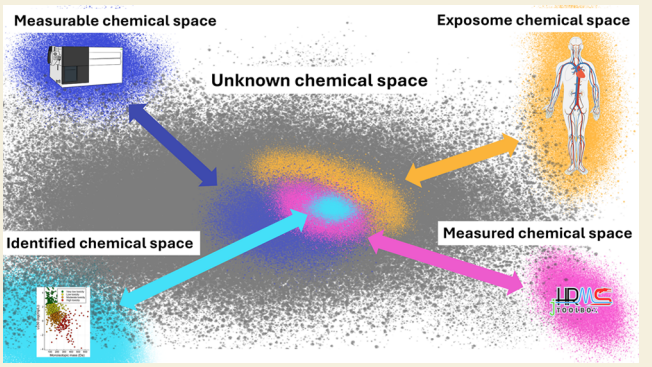
pubs.acs.org/jacsau
Perspective

Exploring the Chemical Space of the Exposome: How Far Have We Gone?

Saer Samanipour,* Leon Patrick Barron, Denice van Herwerden, Antonia Praetorius, Kevin V. Thomas, and Jake William O'Brien

Cite This: *JACS Au* 2024, 4, 2412−2425
Read Online

ABSTRACT: Around two-thirds of chronic human disease can not be explained by genetics alone. The Lancet Commission on Pollution and Health estimates that 16% of global premature deaths are linked to pollution. Additionally, it is now thought that humankind has surpassed the safe planetary operating space for introducing human-made chemicals into the Earth System. Direct and indirect exposure to a myriad of chemicals, known and unknown, poses a significant threat to biodiversity and human health, from vaccine efficacy to the rise of antimicrobial resistance as well as autoimmune diseases and mental health disorders. The exposome chemical space remains largely uncharted due to the sheer number of possible chemical structures, estimated at over 10^{60} unique forms. Conventional methods have cataloged only a fraction of the exposome, overlooking transformation products and often yielding uncertain results. In this Perspective, we have reviewed the latest efforts in mapping the exposome chemical space and its subspaces. We also provide our view on how the integration of data-driven approaches might be able to bridge the identified gaps.



KEYWORDS: *Chemical space, Exposome, Data-driven, Measurability, NTA, Retrospective analysis, Structural elucidation*

INTRODUCTION

The number of chemical structures known to us is expanding exponentially; for example, the number of entries to the Chemical Abstracts Service (CAS) registry has crossed the threshold of 100 million substances in 2015 and continues to grow.[1−7] A similar trend is observed for other chemical families and databases.[8−13] For example, PubChem currently includes more than 115 million unique structures, and this number is growing. Even though these numbers may seem large, compared to the true size of the chemical space, more than 10^{60} for organic structures smaller than 500 Da, these lists cover less than 0.001% of the possible chemical space.[13−16] Furthermore, even for known structures, <1% of them have been experimentally evaluated for their environmental and biological activity (e.g., toxicity), due to the cost and complexities associated with such measurements.[3−5] In fact, according to Persson et al., around 80% of the chemicals defined as in use according to REACH have yet to be assessed, even though the data may be available.[13] Several studies have shown the negative impact of exposure to chemicals with long-term adverse health outcomes.[17−22] For example, exposure to per- and polyfluoroalkyl substances (PFAS) has shown to correlate with the symptoms of autoimmune disease as well as mental health issues.[18,23]

Our current chemical management strategy is mainly based on manual chemical registration and/or experimental measurements of those chemicals in environmental and biological samples.[1,3,4,24] Both approaches are extremely challenging, costly, and inherently passive or, at best, reactive. Chemical registration, with regulatory focus, takes place only for chemicals with large production volumes at the national or international level (e.g., REACH Regulation).[3,8,9,25,26] With digitalization, these chemical registries and patents as well as scientific publications have been mined to gain an approximate idea about the current exposome chemical space (i.e., all the organic chemicals that humans are exposed to during their lifetime).[27−29] For example, databases such as PubChem or US-EPA CompTox dashboard are constantly updated with new chemicals coming from these mining exercises.[9,27] This process, even though sophisticated and powerful, is mainly centered toward human-made chemicals, thus having limited

Received: March 8, 2024
Revised: May 29, 2024
Accepted: May 31, 2024
Published: June 20, 2024

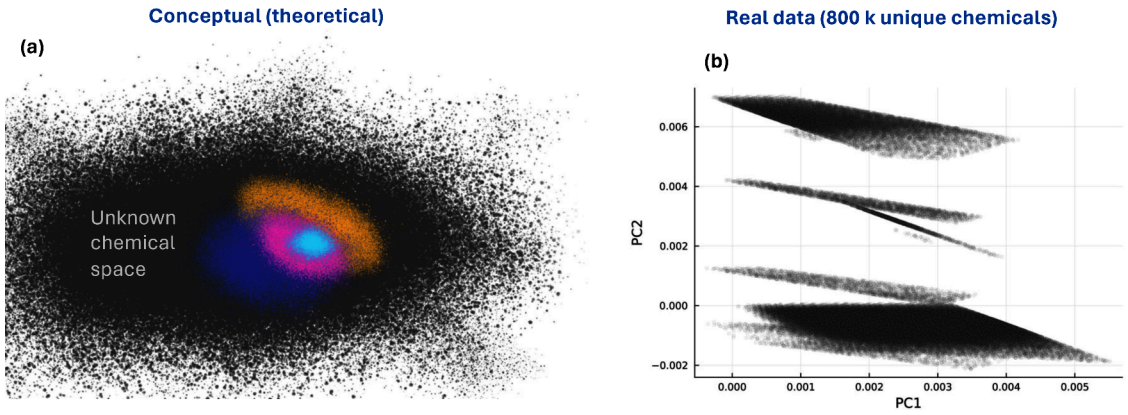


Figure 1. (a) A conceptual figure showing different chemical subspaces (i.e., "relevant chemical space" to exposome), including unknown chemical space (gray), exposome chemical space (orange), measurable chemical space (blue), measured chemical space (magenta), and identified/characterized chemical space (light blue) whereas (b) shows the chemicals in US-EPA CompTox with 800 k unique structures. The Principal Component (PC) plot was generated using six elemental mass defects and monoisotopic mass of the chemicals in US-EPA CompTox (details of these calculations and the scripts are provided elsewhere[37,38]). It should be noted that the size of subspaces in panel (a) are meant only for visualization purposes and are not representative of the true size of these spaces. The empty spots in the PC space (panel b) suggest that the exposome chemical space may not be a smooth and continuous space, mainly due to the organic chemistry rules.

coverage of structures produced via abiotic and biotic transformation and limited consideration for any future chemicals.

Chemical transformations can take place in the environment (e.g., photo- or biological degradation) or within human-made infrastructures, such as wastewater treatment plants.[30−33] Depending on the type of reactions and the structure of the parent compound, there may be more than 100 new chemicals formed at even the first level of the transformation tree.[31,34,35] Considering the costs and complexity associated with performing transformation experiments, the expansion of such methods to the exposome chemical space is impossible, Figure 1. An alternative to this experimentally driven approach has been the predictive models where a combination of machine learning and heuristic methods are used.[30,31,35] However, these methods are very uncertain, opaque, are limited to a few reaction pathways while stopping at shallow levels (e.g., first or second levels) in the transformation tree.[4,36] This implies that our current estimates of the coverage of the exposome chemical space are orders of magnitude smaller than its true size, given the number of possible reactions.

Thus far, measuring/monitoring chemicals in, across, and between different media (e.g., water, soil, air, or biological material) has been the main strategy to map the exposome chemical space. This strategy is reliant on three complementary approaches, namely, targeted, suspect, and nontargeted analysis (NTA).[39−41] Targeted analysis could be quantitative and focused on a limited number of preselected structures; for example, less than a few hundred chemicals are actively and routinely monitored in different matrices. Suspect screening/analysis, on the other hand, has been employed to identify chemicals based on user curated lists of preselected compounds and/or presence in databases/libraries using full-scan high-resolution mass spectrometry (HRMS) data.[40,41] Finally, NTA is considered the most agnostic approach for chemical measurement and identification in samples, where the collected signals are translated into candidate structures and confirmed via target analysis.[40,41] Both suspect and nontarget analysis are reliant on full scan data generated via HRMS coupled to a separation technique such as gas or liquid chromatography (GC/LC-HRMS). For a structurally unknown chemical to be identified via suspect or NTA, it must be measurable with our current analytical strategies (i.e., separation and detection).[16,40,42] What is measurable/analyzable with our current analytical technologies is unknown.[42] We are aware of only the identified fraction of the measurable chemical space (see the description below). Therefore, there may be chemicals highly relevant to the exposome that are not measurable with routine methods and may require the development of specific methods for their analysis (e.g., PFOS or glyphosate).[23,43,44]

Currently less than 10% of the signals acquired for suspect and NTA assays are successfully identified/annotated.[40,41,45,46] For example, even for a single surface water sample, the preprocessing reveals thousands of chemical signals to be identified. The existing identification workflows, at best, can confirm less than a few hundred chemicals in complex samples. Therefore, the signal of the unidentified chemicals in those samples remains unused.[38,40,47] These unidentified signals may be of high enough quality to be identified using updated preprocessing strategies or expanded spectral libraries (i.e., retrospective analysis).[48,49] In fact, the retrospective analysis of combined data from multiple human cohorts resulted in additional inferences on the connection of chemicals and health outcomes.[49]

In this Perspective, we critically assess the knowledge and technological gaps in comprehensive characterization/mapping of the exposome chemical space. We thereby aim at helping to provide means for future developments toward more proactive chemical management.

CHEMICAL SPACE

The concept of the chemical space was initially introduced within the field of drug discovery, where the central role was the exploration of drug-like structures.[50,51] Those efforts were based on using brute force and known organic chemistry rules to generate all possible structures within set boundaries, for example, the number and the type of elements.[15,50−52] This approach resulted in an extremely large number of possible structures ranging between 10^{20} and 10^{60} for molecules containing 30 atoms or less. Chemical databases such as GDB-20 or Zinc are generated using this strategy, and they

https://doi.org/10.1021/jacsau.4c00220
JACS Au 2024, 4, 2412−2425

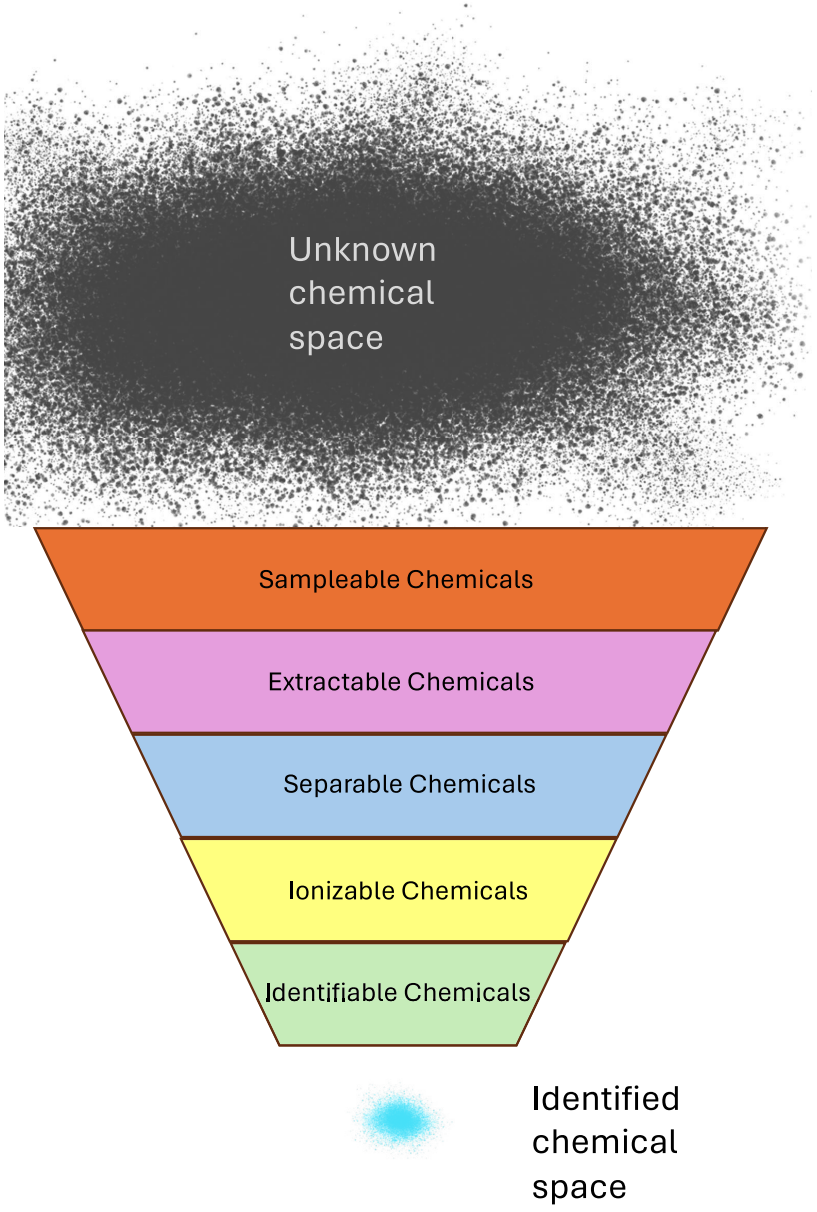


Figure 2. Depicts all the criteria for a chemical to be measured and identified. Each step results in the size of accessible chemical space.

also contain a few estimated molecular descriptors such as logP (i.e., partitioning coefficient between water and an organic phase).[14,53,54] The main objective of such databases is to be able to query them for structural similarity and/or a specific functional group.[54] Given the approach used for generating such databases, they consist of structures that go beyond drug-like chemicals.[55]

The chemical space contains a myriad of structures that may or may not be relevant to a specific application case.[16,56] For example, ozonation degradation products of a natural product, although highly relevant to exposomics, may not be relevant to drug discovery. These selected subspaces of the chemical space are defined as "relevant chemical space" (Figure 1), which are field/question dependent, in the case of exposome it is "exposome chemical space".[51,57] Another chemical subspace is the "measurable chemical space" (Figure 1). This subspace represents the chemicals within the chemical space that can be measured using current analytical techniques. The measurable chemical space focuses only on whether a chemical can be separated and ionized via one of the existing mass spectrometry ionization technologies. The measured chemical space is the chemical subspace where all the structures have been previously measured (Figure 2). Being part of the measured chemical space does not imply that these chemicals have been identified (i.e., structurally confirmed). As an example, features in chromatograms that are not identified during NTA assays are part of the measured chemical space. The most well-known chemical subspace is the structurally confirmed/identified chemical space. This is composed of chemicals that are well-known and studied, for example, pharmaceuticals and pesticides. Except for the identified chemical space, other subspaces are mostly unexplored and thus unknown. The relevant and measurable subspaces may overlap depending on the field. For example, in the field of exposomics, the relevant chemical space may be larger than the

https://doi.org/10.1021/jacsau.4c00220
JACS Au 2024, 4, 2412−2425

measurable one, while in drug discovery this may not hold true.

Exploration of Chemical Space

To explore such a vast chemical space, several cheminformatics tools have been built.[51,52,54,58,59] These techniques range from simple nearest neighbor search to generative models based on large language models.[52,58,60−63] These tools are mainly built either to focus on a very specific subspace of the total chemical space (e.g., drugs) or to explore the chemical space as a whole, as a visualization strategy. Recent developments in graph-based methods such as molecular networks have provided the means of a more detailed exploration of the chemical space.[15,62−64] However, the application of these tools has been limited to mostly visualizing the chemical space and to the drug discovery area, due to the sheer size and the diversity of the chemical space.

Within the exposomics community, the concept of exploration of the chemical space has been focused on building large chemical databases (e.g., PubChem).[9,10,16,25] Additionally, recent works have used text mining approaches to further enrich these lists and thus expand the known chemical space, including chemical classification based on either repeating units or the presence of specific functional groups. These tools have also enabled the classification of chemicals based on their functional groups or repeating units.[9,29] The ultimate goal of these efforts has been detailed characterization of the exposome chemical space. However, they are inherently limited to the registered or identified chemicals.

EXPOSOME CHEMICAL SPACE

The exposome chemical space is the chemical subspace which humans are exposed to from conception to death.[4,5] The exposome chemical space is mostly unknown and may include human-made and natural chemicals, as well as their transformation products. The efforts to explore/map the exposome chemical space have been divided into computational and experimental.[27,40] The computational approaches focus on building chemical databases of mainly human-made chemicals and then ranking (i.e., chemical prioritization) those structures based on the available metadata (e.g., volume of production).[3,9,10] As for the experimental strategies, the focus has been on actively measuring the chemicals in different environmental compartments, including the transformation products.[40,41]

Computational Approaches for Database Building

The main strategy to add human-made chemicals to the list of chemicals associated with the exposome chemical space has been the mining of national and international chemical registries as well as the mining of patents and scientific publications. One of the earliest efforts to keep track of human-made chemicals has been Chemical Abstract Services (CAS). Formed in 1904, CAS collects structural information for chemicals synthesized as early as 1800 (source: Chemical Abstract Services). The main source of CAS is the scientific literature, where newly reported chemicals are registered and given an identification number (i.e., CAS number). Other similarly formed databases such as PubChem,[9] ChemSpider,[65] NORMAN SusDat, and US-EPA CompTox act as hubs where the data from different chemical registries is gathered, curated, and made available for public use. While these larger databases are more generic and tend to have different chemical families, there are also more specialized chemical databases such as drug bank,[66] human metabolome database,[67] and FORIDENT.[68]

In addition to the literature-based chemical databases, there are also national and international chemical registries with mainly regulatory focus. For example, the European Chemical Agency (ECHA) formed in 2007 is the registry of chemicals used, imported, and/or exported into the European Union as well as the Organization for Economic Co-operation and Development (OECD) and/or eChemportal. As these chemical databases are meant for regulatory purposes, they also include information regarding the volume of production/use as well as biological activity (e.g., toxicity) and physiochemical properties. However, these databases may have different volumes of production/use registration thresholds.[8,69] Furthermore, these databases rarely include the transformation products of human-made chemicals unless they are being actively produced and used for other purposes. However, these databases are limited to human-made chemicals, and their size is increasing by around 1500 new structures a year.[3]

Not all human-made chemicals, registered or not, are part of the exposome chemical space due to their total volume of production, physiochemical properties, and use type. For example, the potential of exposure to a chemical with a very small volume of production may be very low, as this chemical once released into the environment is infinitely diluted. Between 2010 and 2012, Howard and Muir published three very influential manuscripts in which they reported lists of high priority chemicals to be further studied.[8,69,70] In those studies, the authors investigated all the existing North American chemical databases and selected the chemicals with a volume of production larger than 1 ton a year. This threshold was set to ensure the environmental detection of these chemicals. Additional filtering (i.e., chemical prioritization) based on physiochemical properties and expert knowledge were employed to narrow down these chemicals to those pertinent to the environmental and human exposome. Similar efforts have been carried out globally for mapping the exposome relevant chemical space (e.g., SusDat).[71,72] It should be noted that these chemical prioritization approaches are designed to direct the monitoring programs given the costs and difficulties associated with them. Consequently, these databases cover only a small portion of the exposome chemical space.

Experimental Approaches for Exposome Assessment

Detection, identification, and quantification of chemicals in exposure media and biological samples are additional approaches for mapping the exposome chemical space (Figure 2).[40,41] Typically, a combination of target, suspect, and NTA using HRMS is employed for the structural elucidation of the chemicals in the exposome chemical space.[39,56,73] Each of these approaches has its advantages and limitations, and they are usually combined together to maximize their coverage of the exposome chemical space.

Targeted analysis is a top-down approach where all of the necessary information for the unequivocal identification and potential quantification of a chemical in a sample is available to the analyst prior to the analysis. Targeted analysis is the main strategy for routine monitoring of chemicals in environmental and biological samples.[39,40,47,74,75] On the other hand, suspect and NTA are less certain and also tend to be only qualitative,[40,47,76] even though there have been several new developments in semiquantification of known and unknown

https://doi.org/10.1021/jacsau.4c00220
JACS Au 2024, 4, 2412−2425

chemicals.[77] For suspect analysis/screening, a list of suspect analytes with as much information as possible (e.g., predicted retention behavior, fragmentation spectra) is compiled, implying that suspect analysis, similar to the target analysis, is a top-down approach. The generated suspect list is used in a later stage for the detection and tentative identification of the chemicals in the analyzed samples. NTA is the most comprehensive but uncertain approach for the identification of chemicals in environmental and biological samples. NTA is a bottom up approach where minimum or no prior knowledge about the structure of the chemicals in samples is used during the identification process. The ultimate goal of most NTA workflows within the exposomics area is the unequivocal identification of all chemicals present in a sample. However, this process is extremely difficult, time-consuming, and uncertain,[40,41] especially when applied across multiple environmental compartments (e.g., air, water, soil, biological fluids, etc.) and spatiotemporally. Consequently, when looking at the number of new structures discovered in environmental samples using NTA strategies in the past five years, those studies resulted in less than 2% of a database such as Norman SusDat.[38]

Transformation Products

Transformation products, natural or based on human-made processes, theoretically constitute a large portion of the exposome chemical space. Each human-made chemical could potentially have a large number of different transformation products, depending on the reaction pathways and the environmental conditions (e.g., biotic or abiotic).[30,31,35,78,79] Some of these transformation products may be more persistent than their parent compounds and hence be even more relevant for the exposome chemical space. However, most of these structures remain unknown, even though their importance to environmental and human health has been previously demonstrated (e.g., DDT and its metabolites DDE and DDD or disinfection byproducts).[80−82]

A combination of experimental and *in silico* approaches is typically employed for the structural elucidation of the transformation products of chemicals.[30,31,34−36] This task is carried out for one chemical and one reaction type at a time due to the complexity of such systems (e.g., photodegradation of pharmaceuticals). To elucidate the generated transformation products, a combination of NTA/suspect analysis and *in silico* prediction tools is used.[83−86] The currently available *in silico* tools are able to estimate the structure of a potential transformation product based on the parent structure and the reaction type.[85] These transformation product structures are used either for the generation of suspect lists or as potential candidate structures during the NTA workflows. Additionally, the generated transformation product structures may not have their chemical standard available or may not have been measured before, increasing the complexity and uncertainty of this task. Additionally, due to the uncertainties associated with the *in silico* transformation product structure estimation tools and the NTA workflows, the addition of the transformation products to the list of chemicals in the exposome chemical space has been an extremely slow process. In fact, most of the chemicals present in the databases such as PubChem or Norman SusDat consist of the parent structures rather than transformation products,[36] indicating the need for their expansion with transformation products.

MEASURABLE EXPOSOME CHEMICAL SPACE

The measurable exposome chemical space is the subspace of chemicals that can be measured via existing analytical strategies, in particular, GC and/or LC-HRMS. A recent review by Manz et al. highlighted that the majority of human exposome-related studies that employed HRMS-based NTA used LC-MS only (51%), followed then by GC-MS only (32%), ≈16% used both techniques together, and 1% used direct injection-HRMS without any separation.[87] Of 76 HRMS-based studies reviewed in total, there was no consistency in application of different analytical platforms across chemical classes or the environmental compartment studied. The majority of applications lay in the food and consumer products space (n = 19 studies), followed by air (n = 15), soil/sediment (n = 13), dust and human samples (each, n = 10), and then water (n = 9). Fundamentally therefore, it seems that researchers have assumed that, at exposome relevant concentrations, a large component of chemicals can be separated via a chromatographic approach and be ionized/fragmented via HRMS technology. It should be noted that slight deviation from the optimal experimental conditions may have extreme impact on the measurable subspace explored by the used method.[16,40,41,87,88] There are several such examples where more generic methods fail to cover highly relevant chemicals in the exposome chemical space (e.g., PFOS or glyphosate).

The separation space is dominated by reversed-phased liquid chromatography (RPLC) and gas chromatography. For RPLC in particular, there is often a linearity assumption between the hydrophobicity and the size of a chemical and its retention within the set experimental setup.[89,90] There are several reversed-phase liquid chromatography studies where the retention times of the internal standards are correlated to their octanol/water partitioning coefficient.[89,91,92] These linear models are then extrapolated to infer which portion of the chemical space is covered. This linearity assumption has been challenged by different studies focusing on retention time modeling.[93−95] A recent study has shown that chemicals with similar retention behavior in RPLC may have up to 6 orders of magnitude variance in their predicted partition coefficients. In the same study, a data-driven approach showed that 20,000 chemicals present in Norman SusDat (around 100 k unique structures) are not analyzable with RPLC.[95]

In the detection space, the ionization efficiency (IE) is the main determining factor for the measurability of a chemical, which fundamentally and potentially significantly limits the measurable space covered.[77,96] The IE is a structure-dependent parameter indicating the magnitude of the generated signal for a specific chemical. There have been several recent studies that have successfully predicted the IE of known and unknown structures. Therefore, the IE could be used as a parameter for categorizing chemicals as detectable or not detectable based on their IE. It should be noted that there has been a study classifying the chemicals to be analyzable via GC-MS vs LC-MS.[42] However, that study did not emphasize the measurability bottleneck and has been trained based on well-known structures (e.g., simple hydrocarbons and pharmaceuticals).

Instrumental Perspective

In a simplistic sense, our current measurements for the chemical space rely on getting separated compounds into a mass spectrometer under conditions that are sufficient for them to be measured. This has heavily relied on the ability to

https://doi.org/10.1021/jacsau.4c00220
JACS Au 2024, 4, 2412−2425

easily introduce the chemicals into liquid (mainly RPLC) and gas chromatography and separation based on interactions between the chemicals, a solid phase, and a mobile phase.[38,87,97] This implies that the compounds are already in a liquid or gaseous solution and hence already exclude particles, low solubility chemicals, and nonvolatiles. Mass spectrometry requires chemicals to be in a gaseous ionized state, and ionization has largely focused on soft ionization for LC (mainly electrospray (ESI) and atmospheric pressure chemical ionization (APCI)) to produce pseudo/molecular ions (precursor ions) and hard ionization (electron impact (EI))) for GC which typically fragments the molecules. As such, chemicals with poor ionization efficiency will likely not be introduced to the mass spectrometer, and those that have very high ionization efficiency may fragment too much to provide sufficient identification of the parent compound. Even those that do ionize may be unstable or rearrange. Beyond ionization, precursors are typically fragmented using collision-induced dissociation, also known as collisionally activated dissociation, and then measured via tandem mass spectrometry. Only recently have we seen the introduction of electron-activated dissociation (EAD) into commercial mass spectrometers,[98] and as such, new libraries need to be developed to identify chemicals based off the spectrum produced through EAD.

Rarely in the NTA space have we seen the application of other separation (e.g., electrophoresis) and ionization techniques such as inductively coupled plasma mass spectrometry (ICPMS) and matrix-assisted laser desorption/ionization (MALDI). For identification of chemicals that may impact humans through surface contact or inhalation of particles, other ambient ionization techniques have emerged including direct-analysis in real-time (DART) and desorption electrospray ionization (DESI).[99,100] These platforms have been used heavily in forensic science for the identification of bulk drug or explosive materials or diagnostic markers in contact evidence, such as illicit drug metabolites in fingermarks.[40,87] Other more recent developments include rapid evaporative ionization (REIMS) or laser desorption ionization (LDI) which have both been integrated within surgical blades to rapidly classify biological materials in real-time through identification of discriminating biomolecules.[101] While other ionization sources exist, these are rarely used in parallel or tandem configurations with the conventional sources. Even the application of positive/negative switching ionizations remains limited. Besides instrumental ionization techniques, chemical ionization such as derivatization can be conducted offline or even online and in some cases may even lead to better chromatographic separation.[102]

MEASURED CHEMICAL SPACE

Within the measured subspace of the exposome chemical space, a large portion of the detected chemicals remain unidentified. This implies that a large portion of the collected analytical signals, even though of high quality, remain unidentified. An example of these cases is the C6 to C16 PFAS chemicals in lipidomics studies, as both studies use a very similar experimental setup. Recent studies have indicated that the retrospective analysis can be employed to further annotate/identify the unknowns in the archive data sets.[48,103−105] The retrospective analysis of the archived data, even though it has shown great potential, has not been widely applied for the exploration of the measured chemical space. This shortcoming mainly has been due to inadequate data processing tools, limited chemical and spectral databases, the hypothesis-driven approaches used in NTA experiments, and limited computational power available to different research groups.

The data processing strategies used for the retrospective analysis of the archived data mainly consist of typical NTA workflows.[40,48] There have been several extensive reviews on such workflows and all the included steps.[40,45,106−108] Once these data are processed and further annotated, the identified signals are aligned over multiple data sets for trend analysis and/or inference.[109−111] However, the confidence levels associated with that identification may not be the same across different samples.[48,104,106] Depending on the quality of the generated data (e.g., signal-to-noise ratio), those identifications may be less reliable than others.[46] In addition to that, the unidentified signals cannot be aligned unless generated under the same experimental conditions.[46] These challenges have resulted in several attempts in assessing the data quality as well as the chromatogram alignment.[40,89,92]

The quality of the collected data for NTA assays defines whether the generated signals can be confidently identified or not.[112,113] The quality of the acquired data may be compromised due to heavy matrix effects, instrumental issues, and/or the nature of the chemical itself.[40,41] For some chemicals to generate an HRMS signal of acceptable quality, they must be analyzed using specific conditions.[88] These issues with the data quality can be detected only once the data go through a complete data processing workflow. A common approach to approximately assess the quality of the collected data is to look for the added internal standards in the samples.[45,46,104,106] This approach, even though effective, is computationally expensive and requires detailed metadata about the experimental setup (e.g., the separation selectivity, ionization efficiency, data acquisition conditions). Such information may or may not be available for a specific study depending on the objectives. Moreover, depending on the number and spread of internal standards, they may not be enough for assessing the quality of the collected data. Some efforts have been put toward standardizing the NTA data generation procedures.[40,114−116] However, the proposed procedures are specific to different communities and do not translate across. As an example, the metabolomics and exposomics communities use different retention index scales (alkylamide[117] vs University of Athens scale[118,119]), making comparison of such data sets extremely difficult. Furthermore, these measures tend to be overly conservative, resulting in low levels of implementation within each community.

Another very important challenge to be tackled during the processing is the signal alignment.[40] The aligned signals are needed for trend analysis and signal (feature) prioritization. Existing tools are limited to either a single batch or fully identified structures.[89−92] Retention indices using a set of calibrant chemicals or retention mapping have been widely utilized for such alignments.[89,92,117−119] However, both of these approaches need the presence of a set of chemicals (i.e., calibrants) in all the samples. Such solutions may be very effective for small scale chromatogram alignment but are not able to address the challenges associated with the alignment of archived data, given that these data rarely include the retention index calibrants or the same set of internal standards. Additionally, the few added internal standards are meant to represent all of the chemicals present in those samples. These

https://doi.org/10.1021/jacsau.4c00220
JACS Au 2024, 4, 2412−2425

limitations imply that the alignment of data sets acquired on different instruments using different experimental conditions is still an open question and requires further development.

The last step of NTA assays consists of the identification workflows, including spectral matching against the experimental or *in silico* predicted spectra.[40,41] Independently from the source of the reference spectra, i.e., experimental or predicted, the quality of the measured spectrum to be identified is essential.[112,113] In the field of exposomics, the efforts in assessing the quality of the recorded spectra is limited to a recent study with limited applicability.[112] In addition to the spectral quality, spectral matching algorithms also play an important role during the identification process. These algorithms range from simple dot product to spectral entropy and more data-driven approaches (e.g., deep learning).[106,120−124] For the spectral library, matching the size of the spectral library and its coverage of the relevant chemical space is an essential factor. For example, when merging all open and commercial libraries together, only 40% the metabolic networks is covered.[125] As for *in silico* fragmentation tools, access to a large and relevant chemical subspace is important.[25,126−128] These approaches employ the structure of the different candidates retrieved from the chemical databases for fragmentation or fingerprint prediction. However, these approaches, even though powerful, have limitations due to the limited coverage of the exposome chemical space by current chemical databases.

In addition to the previously mentioned approaches for the structural elucidation, recently, there has been a surge of data-driven tools for facilitation of *de novo* identification and/or annotation of the chromatographic signal. Many of these tools use a combination of machine learning, previously annotated spectra, and spectral similarity to provide additional inference into the structure of an unknown signal (e.g., molecular networking and Spec2LDA).[129−131] However, their applications have been limited to mainly metabolomics studies, and therefore, they have not been adequately tested for the exploration of exposome.

Identified Chemical Space

The identified chemical space is the subspace of measured chemical space, where those structures are fully characterized (e.g., measured via GC/LC-HRMS and available as analytical standard). This subspace is extremely small compared to the size of the exposome chemical space. A recent meta analysis showed that all NTA studies in the past 5 years have resulted in around 1600 (i.e., confidence levels one and two) unique new structures while every year around 700 new structures are introduced into the US market alone.[3,38] It should be noted that the true number of new structures introduced into the global market is extremely difficult to estimate. Considering the number of potential transformation products of these chemicals, the speed of NTA studies is far too low to be able to catch up with the rate of expansion of the exposome chemical space.

HOW TO MOVE FORWARD

The main categories of chemicals that are absent from the current exposome related chemical databases are the transformation products of anthropogenic chemicals. The structure-based molecular networks (SBMNs), from drug discovery, combined with synthetic accessibility (computational synthetic chemistry) can be used to build the transformation tree of a chemical.[62,132,133] The already well-known transformation products would provide the distance metrics for the SBMNs, and the synthetic accessibility calculation would enable pruning of the trees from the structures that are impossible.[133−135] Other data-driven approaches, such as generative models, can also provide the means of building such transformation trees. Ultimately, the structures in the pruned trees can be added to existing chemical databases.

In terms of the measurable exposome chemical space, the combination of the modeled separation (e.g., retention time) and mass spectral behavior of chemicals can be used. Retention indices can provide a first glance into the connection between the structure of a molecule and its behavior in the chromatographic space. On the other hand, the ionization efficiency has great potential in connecting the structure of a chemical to its response in the mass spectrometer. The combination of these two metrics can provide a valuable training set to build models where the measurability of a chemical can be assessed based on its structure. A potential byproduct of such a strategy is that these models may be able to suggest the optimized experimental conditions for the analysis of a certain structure (e.g., reverse phase vs normal phase). In addition, the development and integration of models using complementary separation techniques in parallel to chromatography are needed (e.g., electrical separations such as capillary electrophoresis (CE) and ion mobility spectrometry).[101,136−139] Ultimately, modeling multiple separation spaces using data from orthogonal techniques may reduce porosity and extend the boundaries of the measured chemical space, as well as provide additional confirmation where overlaps exist (e.g., machine learning-based prediction of both retention behavior and collision cross section may increase confidence in prioritization workflows[89,140]).

To further expand our coverage and understanding of the chemical space experimentally, harmonization of both complementary and orthogonal techniques is required, for example, through application of both GC- and LC-HRMS and HILIC and RPLC separation to the same samples. We also need to better understand the boundaries of each technique and the porosity within. For example, polar analytes typically have lower ionization efficiency; hence, further development is required for both separation and ionization techniques. Integration of ion chromatography (IC) and CE coupled with both HRMS and ICP-MS techniques needs to be considered. These instruments provide unique selectivity for very polar, inorganic, and ionized compounds (e.g., metals/metalloids, low molecular weight PFAS, and disinfection byproducts); hence, for NTA of drinking water, this is a particular knowledge gap where these platforms represent excellent solutions.[88] IC and CE separation techniques are orthogonal and inherently complement each other.[141]

Mass spectrometry is not the only option available for identification, and other techniques such as Nucelar Magnetic Resonance (NMR) can be coupled with LC separation. Proton NMR in particular should be considered for determining the structure of organic molecules as it allows the ability to elucidate the connectivity of the atoms within molecules and for identifying functional groups.[142,143]

However, this draws out a larger issue that arguably requires much more focus in the NTA community moving forward. Identification frameworks for chemical residues in environmental samples usually fundamentally consider the value offered by HRMS data first, followed by evaluation of any

https://doi.org/10.1021/jacsau.4c00220
JACS Au 2024, 4, 2412−2425

increased confidence provided by supplementary chromatographic data.[144] In other fields such as forensic science, the combination of data generated by a much wider set of orthogonal/uncorrelated techniques is considered in far more depth. For example, the Scientific Working Groups for the Identification of Seized Drugs (SWGDRUG) and Fire and Explosions (SWGFEX) both categorize techniques broadly into those providing presumptive, indicative, and confirmatory evidence. The SWGFEX guidelines for postblast explosive identification are a particularly relevant example for trace chemical NTA (recommended guidelines for the forensic identification of postblast explosive residues). An array of confirmatory techniques are categorized by their ability to offer structural or elemental level detail and include Raman, FTIR, and X-ray diffraction, in addition to LC-MS and GC-MS. Energetic materials are well-known to be very challenging to measure by any single method or technique. Gaps in the measurable space that exist for methods that use confirmatory techniques are fundamentally considered. This is especially true for chromatography coupled to mass spectrometry. Given the potential undesirable outcomes of a "false negative" in this particular field, the combination of techniques is critical (e.g., to identify both inorganic and organic explosives). Even when using MS, the choice of ionization technique in LC-MS is also extremely important (e.g., ESI is normally more suitable for nitrate esters, APCI is better for nitrotoluenes, and neither are particularly effective for detection of some explosives like nitroglycerin or hexamethylene triperoxide diamine[145]). For the identification of intact/bulk drugs, SWGDRUG also considers NMR spectroscopy a confirmatory technique (Scientific Working Group for the Analysis of Seized Drugs (SWGDRUG) Recommendations, 2019). With regards to exposomics, NMR has been used for many years for the identification of biomolecules in exposome research.[146] Though less sensitive than MS generally, higher field instruments, multiple scans, different probes, or hyperpolarisation used together with sample preconcentration methods may improve its contribution for identification of new substances at trace concentrations in complex samples.[147]

As for the measured exposome chemical space, the main challenges are the raw data quality assessment, incomplete preprocessing workflows, and identification workflows. The data quality assessment must become independent from the data preprocessing workflows as they may be a major source of error into final results, for example, low quality MS/MS spectrum due to the lack of deconvolution algorithms. The current NTA workflows are set to focus on a single sample or batch of samples analyzed with one specific method. This limitation hinders large scale retrospective analysis of archive data, as most of the signals may remain unidentified. Moreover, the use of the same set of internal standards may not be adequate. Therefore, the development of alignment algorithms that are based on the raw data or the raw feature lists is a must for being able to fully take advantage of the publicly available archived data. Finally, when it comes to identification, the current approaches are based on a set number of matched fragments and hard set thresholds. This may not be the most adequate way forward, as different chemicals may need different parameters for their high confidence identification. For example, for PFAS chemicals, having two or three fragments may be enough, while for hormone-like chemicals, sometimes even 100 fragments are not enough. Moreover, depending on the levels of background signal and matrix effects, the mass accuracy of the instrument may be different, resulting in a better match with incorrect candidates.

From a regulatory point of view, knowledge on what can be measured or not is essential. Chemicals that cannot be measured or are very difficult to measure (e.g., very mobile chemicals) are very difficult to regulate. Thus, for new chemicals to be introduced, evidence of the ready measurability of the parent and the most abundant transformation products may be considered as one of the necessary criteria. The high detection frequency of chemicals in the archived data can be further integrated as one of the strategies for early detection of chemicals of emerging concern.

Overall, here, we have highlighted the most immediate scientific gaps related to mapping the exposome chemical space. It should be noted that there are many more challenges that need to be tackled. However, based on our assessment, it is clear that the current approaches do not provide the means for a pro-active chemical management. Therefore, the combination of data-driven approaches with existing strategies will be a necessary step forward to bridge these knowledge gaps.

POTENTIAL IMPACT

The expanded and mapped chemical space of the exposome, its predicted physiochemical properties, and biological activities will unleash new waves of developments in chemical management, toxicology, and analytical technology development. The predicted properties and biological activities will guide new chemical regulation. The transformation products added to the exposome chemical space will provide the means for the replacement of toxic chemicals with safe alternatives and thus future safe and sustainable-by-design chemicals. The measurability assessment will identify the portion of the exposome chemical space that cannot be analyzed with our current technology, stimulating the development of new analytical tools to further expand this coverage. The newly identified chemicals via a retrospective analysis of the archived data will provide insights into the connection between chemical exposure and an observed health outcome. These connections will provide insights into the mechanistic relationships between exposure and certain health outcomes.

AUTHOR INFORMATION

Corresponding Author

Saer Samanipour − *Van't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, Amsterdam 1090 GD, The Netherlands; UvA Data Science Center, University of Amsterdam, Amsterdam 1090 GD, The Netherlands; Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, Woolloongabba, Queensland 4102, Australia;* orcid.org/0000-0001-8270-6979; Email: s.samanipour@uva.nl

Authors

Leon Patrick Barron − *MRC Centre for Environment and Health, Environmental Research Group, School of Public Health, Faculty of Medicine, Imperial College London, London W12 0BZ, United Kingdom; Van't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, Amsterdam 1090 GD, The Netherlands*

Denice van Herwerden − *Van't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, Amsterdam*

https://doi.org/10.1021/jacsau.4c00220
JACS Au 2024, 4, 2412−2425

1090 GD, The Netherlands; orcid.org/0000-0003-1940-9415

Antonia Praetorius − *Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, Amsterdam 1090 GD, The Netherlands;* orcid.org/0000-0003-0197-0116

Kevin V. Thomas − *Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, Woolloongabba, Queensland 4102, Australia;* orcid.org/0000-0002-2155-100X

Jake William O'Brien − *Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, Woolloongabba, Queensland 4102, Australia; Van't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, Amsterdam 1090 GD, The Netherlands;* orcid.org/0000-0001-9336-9656

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.4c00220

Author Contributions

CRediT: **Saer Samanipour** conceptualization, funding acquisition, investigation, project administration, validation, visualization, writing-original draft, writing-review & editing; **Leon Patrick Barron** resources, validation, visualization, writing-original draft, writing-review & editing; **Denice van Herwerden** data curation, formal analysis, visualization, writing-review & editing; **Antonia Praetorius** conceptualization, investigation, writing-original draft, writing-review & editing; **Kevin V. Thomas** resources, writing-original draft, writing-review & editing; **Jake William O'Brien** formal analysis, funding acquisition, resources, validation, visualization, writing-original draft, writing-review & editing.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors express their gratitude to all the members of Environmental Modeling & Computational Mass Spectrometry (www.emcms.info). S.S. thanks the ChemistryNL TKI and the UvA Data Science Center for their funding support (projects Edified and SCOPE). J.W.O. is the recipient of a National Health and Medical Research Council (NHMRC) Investigator Grant (EL12009209) funded by the Australian Government. S.S. also acknowledges financial 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. 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. L.P.B. thanks the National Institute for Health and Care Research (NIHR) for part funding under the Health Protection Research Units in Environmental Exposures and Health and Chemical and Radiation Threats and Hazards, both partnerships between the UK Health Security Agency and Imperial College London. The views expressed are those of the authors and not necessarily those of the NIHR, UK Health Security Agency the Department of Health and Social Care.

REFERENCES

(1) Arp, H. P. H.; Aurich, D.; Schymanski, E. L.; Sims, K.; Hale, S. E. Avoiding the Next Silent Spring: Our Chemical Past, Present, and Future. *Environ. Sci. Technol.* **2023**, *57*, 6355−6359.

(2) Cousins, I. T.; Johansson, J. H.; Salter, M. E.; Sha, B.; Scheringer, M. Outside the Safe Operating Space of a New Planetary Boundary for Per- and Polyfluoroalkyl Substances (PFAS). *Environ. Sci. Technol.* **2022**, *56*, 11172−11179.

(3) Muir, D. C. G.; Getzinger, G. J.; McBride, M.; Ferguson, P. L. How Many Chemicals in Commerce Have Been Analyzed in Environmental Media? A 50 Year Bibliometric Analysis. *Environ. Sci. Technol.* **2023**, *57*, 9119.

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

(6) Schymanski, E. L.; Zhang, J.; Thiessen, P. A.; Chirsir, P.; Kondic, T.; Bolton, E. E. Per- and Polyfluoroalkyl Substances (PFAS) in PubChem: 7 Million and Growing. *Environ. Sci. Technol.* **2023**, *57*, 16918−16928.

(7) C&EN West Coast News Bureau. The Digital Data Dive. *Chem. Eng. News Archive* **2015**, *93*, 14−15.

(8) Howard, P. H.; Muir, D. C. G. Identifying New Persistent and Bioaccumulative Organics Among Chemicals in Commerce. *Environ. Sci. Technol.* **2010**, *44*, 2277−2285.

(9) 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 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2019**, *47*, D1102−D1109.

(10) 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.

(11) Kristiansson, E.; Coria, J.; Gunnarsson, L.; Gustavsson, M. Does the scientific knowledge reflect the chemical diversity of environmental pollution? − A twenty-year perspective. *Environmental Science amp; Policy* **2021**, *126*, 90−98.

(12) 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.

(13) 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*, 1510−1521.

(14) Arús-Pous, J.; Blaschke, T.; Ulander, S.; Reymond, J.-L.; Chen, H.; Engkvist, O. Exploring the GDB-13 chemical space using deep generative models. *Journal of Cheminformatics* **2019**, *11*, 20.

(15) Awale, M.; Visini, R.; Probst, D.; Arús-Pous, J.; Reymond, J.-L. Chemical Space: Big Data Challenge for Molecular Diversity. *CHIMIA* **2017**, *71*, 661−661.

(16) Black, G.; Lowe, C.; Anumol, T.; Bade, J.; Favela, K.; Feng, Y.-L.; Knolhoff, A.; Mceachran, A.; Nuñez, J.; Fisher, C.; Peter, K.; Quinete, N. S.; Sobus, J.; Sussman, E.; Watson, W.; Wickramasekara, S.; Williams, A.; Young, T. Exploring chemical space in non-targeted analysis: a proposed ChemSpace tool. *Anal Bioanal Chem.* **2023**, *415*, 35−44.

(17) Brown, J. S. Psychiatric Effects of Organic Chemical Exposure. In *Effects of Persistent and Bioactive Organic Pollutants on Human Health*; John Wiley & Sons, Ltd., 2013; pp 514−531.

(18) Bailey, J. M.; Wang, L.; McDonald, J. M.; Gray, J. S.; Petrie, J. G.; Martin, E. T.; Savitz, D. A.; Karrer, T. A.; Fisher, K. A.; Geiger, M. J.; Wasilevich, E. A. Immune response to COVID-19 vaccination in a population with a history of elevated exposure to per- and

polyfluoroalkyl substances (PFAS) through drinking water. *J. Expo. Sci. Environ. Epidemiol.* **2023**, *33*, 725−736.
(19) Chung, E.; Russo, D. P.; Ciallella, H. L.; Wang, Y.-T.; Wu, M.; Aleksunes, L. M.; Zhu, H. Data-Driven Quantitative Structure−Activity Relationship Modeling for Human Carcinogenicity by Chronic Oral Exposure. *Environ. Sci. Technol.* **2023**, *57*, 6573.
(20) Goldenman, G.; Fernandes, M.; Holland, M.; Tugran, T.; Nordin, A.; Schoumacher, C.; McNeill, A. *The cost of inaction: A socioeconomic analysis of environmental and health impacts linked to exposure to PFAS*; Nordisk Ministerråd, 2019.
(21) Kwon, D.; Kwak, K.; Baek, K.; Chi, Y.; Na, S.; Park, J.-T. Association between physical hazardous agent exposure and mental health in the Korean working population: the 5th Korean Working Conditions Survey. *Annals of Occupational and Environmental Medicine* **2021**, *33*, e33.
(22) Zhang, X.; Xue, L.; Deji, Z.; Wang, X.; Liu, P.; Lu, J.; Zhou, R.; Huang, Z. Effects of exposure to per- and polyfluoroalkyl substances on vaccine antibodies: A systematic review and meta-analysis based on epidemiological studies. *Environ. Pollut.* **2022**, *306*, 119442.
(23) Guo, P.; Furnary, T.; Vasiliou, V.; Yan, Q.; Nyhan, K.; Jones, D. P.; Johnson, C. H.; Liew, Z. Non-targeted metabolomics and associations with per- and polyfluoroalkyl substances (PFAS) exposure in humans: A scoping review. *Environ. Int.* **2022**, *162*, 107159.
(24) 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.; Kruve, 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. *Environmental Sciences Europe* **2020**, *32*, 100.
(25) 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. *Journal of Cheminformatics* **2021**, *13*, 19.
(26) van Dijk, J.; Gustavsson, M.; Dekker, S. C.; van Wezel, A. P. Towards 'one substance − one assessment': An analysis of EU chemical registration and aquatic risk assessment frameworks. *Journal of Environmental Management* **2021**, *280*, 111692.
(27) Schymanski, E. L.; Williams, A. J. Open Science for Identifying "Known Unknown" Chemicals. *Environ. Sci. Technol.* **2017**, *51*, 5357−5359.
(28) Barnabas, S. J.; Böhme, T.; Boyer, S. K.; Irmer, M.; Ruttkies, C.; Wetherbee, I.; Kondić, T.; Schymanski, E. L.; Weber, L. Extraction of chemical structures from literature and patent documents using open access chemistry toolkits: a case study with PFAS. *Digital Discovery* **2022**, *1*, 490−501.
(29) 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*, 7448.
(30) Wishart, D. S.; Tian, S.; Allen, D.; Oler, E.; Peters, H.; Lui, V. W.; Gautam, V.; Djoumbou-Feunang, Y.; Greiner, R.; Metz, T. O. BioTransformer 3.0—a web server for accurately predicting metabolic transformation products. *Nucleic Acids Res.* **2022**, *50*, W115−W123.
(31) Beretsou, V. G.; Psoma, A. K.; Gago-Ferrero, P.; Aalizadeh, R.; Fenner, K.; Thomaidis, N. S. Identification of biotransformation products of citalopram formed in activated sludge. *Water Res.* **2016**, *103*, 205−214.
(32) Chibwe, L.; Titaley, I. A.; Hoh, E.; Simonich, S. L. M. Integrated Framework for Identifying Toxic Transformation Products in Complex Environmental Mixtures. *Environ. Sci. Technol. Lett.* **2017**, *4*, 32−43.
(33) Chen, X.; Li, H.-R.; Feng, X.; Wang, H.-T.; Sun, X.-H. Prediction of *OH-Initiated and *NO3-Initiated Transformation Products of Polycyclic Aromatic Hydrocarbons by Electronic Structure Approaches. *ACS Omega* **2022**, *7*, 24942−24950.
(34) Ikehata, K.; Jodeiri Naghashkar, N.; Gamal El-Din, M. Degradation of Aqueous Pharmaceuticals by Ozonation and Advanced Oxidation Processes: A Review. *Ozone: Science & Engineering* **2006**, *28*, 353−414.
(35) Satoh, H.; Hafner, J.; Hutter, J.; Fenner, K. Can AI Help Improve Water Quality? Towards the Prediction of Degradation of Micropollutants in Wastewater. *CHIMIA* **2023**, *77*, 48−48.
(36) Palm, E. H.; Chirsir, P.; Krier, J.; Thiessen, P. A.; Zhang, J.; Bolton, E. E.; Schymanski, E. L. ShinyTPs: Curating Transformation Products from Text Mining Results. *Environ. Sci. Technol. Lett.* **2023**, *10*, 865−871.
(37) 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. *Chemometrics and Intelligent Laboratory Systems* **2022**, *223*, 104515.
(38) Hulleman, T.; Turkina, V.; O'Brien, J. W.; Chojnacka, A.; Thomas, K. V.; Samanipour, S. Critical Assessment of the Chemical Space Covered by LC−HRMS Non-Targeted Analysis. *Environ. Sci. Technol.* **2023**, *57*, 14101.
(39) 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.
(40) Hollender, J.; Schymanski, E. L.; Ahrens, L.; Alygizakis, N.; Béen, F.; Bijlsma, L.; Brunner, A. M.; Celma, A.; Fildier, A.; Fu, Q.; Gago-Ferrero, P.; Gil-Solsona, R.; Haglund, P.; Hansen, M.; Kaserzon, S.; Kruve, A.; Lamoree, M.; Margoum, C.; Meijer, J.; Merel, S.; Rauert, C.; Rostkowski, P.; Samanipour, S.; Schulze, B.; Schulze, T.; Singh, R. R.; Slobodnik, J.; Steininger-Mairinger, T.; Thomaidis, N. S.; Togola, A.; Vorkamp, K.; Vulliet, E.; Zhu, L.; Krauss, M. NORMAN guidance on suspect and non-target screening in environmental monitoring. *Environ. Sci. Eur.* **2023**, *35*, 75.
(41) Schulze, B.; Youngjoon, J.; Sarit, K.; Amy, H. L.; Pradeep, D.; Jake, O.; Maria Jose, G. R.; Sara, G. G.; Jochen, M. F.; Kevin, T. V.; Saer, 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.
(42) Alygizakis, N.; Konstantakos, V.; Bouziotopoulos, G.; Kormentzas, E.; Slobodnik, J.; Thomaidis, N. S. A Multi-Label Classifier for Predicting the Most Appropriate Instrumental Method for the Analysis of Contaminants of Emerging Concern. *Metabolites* **2022**, *12*, 199.
(43) Kaserzon, S. L.; Vijayasarathy, S.; Bräunig, J.; Mueller, L.; Hawker, D. W.; Thomas, K. V.; Mueller, J. F. Calibration and validation of a novel passive sampling device for the time integrative monitoring of per- and polyfluoroalkyl substances (PFASs) and precursors in contaminated groundwater. *Journal of Hazardous Materials* **2019**, *366*, 423−431.
(44) 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.
(45) Schymanski, E. L.; Singer, H. P.; Slobodnik, J.; Ipolyi, I. M.; Oswald, P.; Krauss, M.; Schulze, T.; Haglund, P.; Letzel, T.; Grosse, S.; Thomaidis, N. S.; Bletsou, A.; Zwiener, C.; Ibáñez, M.; Portolés, T.; de Boer, R.; Reid, M. J.; Onghena, M.; Kunkel, U.; Schulz, W.; Guillon, A.; Noyon, N.; Leroy, G.; Bados, P.; Bogialli, S.; Stipaničev, D.; Rostkowski, P.; Hollender, J. Non-target screening with high-

https://doi.org/10.1021/jacsau.4c00220
JACS Au 2024, 4, 2412−2425

resolution mass spectrometry: critical review using a collaborative trial on water analysis. *Anal. Bioanal. Chem.* **2015**, *407*, 6237−55.
(46) Schulze, B.; van Herwerden, D.; Allan, I.; Bijlsma, L.; Etxebarria, N.; Hansen, M.; Merel, S.; Vrana, B.; Aalizadeh, R.; Bajema, B.; Dubocq, F.; Coppola, G.; Fildier, A.; Fialová, P.; Frøkjær, E.; Grabic, R.; Gago-Ferrero, P.; Gravert, T.; Hollender, J.; Huynh, N.; Jacobs, G.; Jonkers, T.; Kaserzon, S.; Lamoree, M.; Le Roux, J.; Mairinger, T.; Margoum, C.; Mascolo, G.; Mebold, E.; Menger, F.; Miège, C.; Meijer, J.; Moilleron, R.; Murgolo, S.; Peruzzo, M.; Pijnappels, M.; Reid, M.; Roscioli, C.; Soulier, C.; Valsecchi, S.; Thomaidis, N.; Vulliet, E.; Young, R.; Samanipour, S. Inter-laboratory mass spectrometry dataset based on passive sampling of drinking water for non-target analysis. *Sci. Data* **2021**, *8*, 1−10.
(47) 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.
(48) 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.
(49) Anzardi, M. B.; Arancibia, J. A.; Olivieri, A. C. Processing multiway chromatographic data for analytical calibration, classification and discrimination: A successful marriage between separation science and chemometrics. *TrAC Trends in Analytical Chemistry* **2021**, *134*, 116128.
(50) Reymond, J.-L.; van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical space as a source for new drugs. *Med. Chem. Commun.* **2010**, *1*, 30−38.
(51) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*, 722−730.
(52) Lin, A.; Horvath, D.; Afonina, V.; Marcou, G.; Reymond, J.-L.; Varnek, A. Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. *ChemMedChem.* **2018**, *13*, 540−554.
(53) Irwin, J. J. Staring off into chemical space. *Nat. Chem. Biol.* **2009**, *5*, 536−537.
(54) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 6065−6073.
(55) Saldívar-González, F. I.; Pilón-Jiménez, B. A.; Medina-Franco, J. L. Chemical space of naturally occurring compounds. *Physical Sciences Reviews* **2019**, *4*, 1.
(56) 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. *Journal of Exposure Science & Environmental Epidemiology* **2018**, *28*, 411.
(57) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824−828.
(58) Lunghini, F.; Gilles, M.; Azam, P.; Enrici, M.-H.; Van Miert, E.; Varnek, A. *Visualization and Analysis of the REACH-chemical Space with Generative Topographic Mapping* **2021**, *40* (4), 2000232.
(59) Saldívar-González, F. I.; Medina-Franco, J. L. Approaches for enhancing the analysis of chemical space for drug discovery. *Expert Opinion on Drug Discovery* **2022**, *17*, 789−798.
(60) Kunkel, C.; Schober, C.; Oberhofer, H.; Reuter, K. Knowledge discovery through chemical space networks: the case of organic electronics. *J. Mol. Model* **2019**, *25*, 87.
(61) Osolodkin, D. I.; Radchenko, E. V.; Orlov, A. A.; Voronkov, A. E.; Palyulin, V. A.; Zefirov, N. S. Progress in visual representations of chemical space. *Expert Opinion on Drug Discovery* **2015**, *10*, 959−973.
(62) Vogt, M. Using deep neural networks to explore chemical space. *Expert Opinion on Drug Discovery* **2022**, *17*, 297−304.
(63) Vogt, M. How do we optimize chemical space navigation? *Expert Opinion on Drug Discovery* **2020**, *15*, 523−525.
(64) Arús-Pous, J.; Awale, M.; Probst, D.; Reymond, J.-L. Exploring Chemical Space with Machine Learning. *CHIMIA* **2019**, *73*, 1018−1018.
(65) Pence, H. E.; Williams, A. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87*, 1123−1124.
(66) Knox, C.; Wilson, M.; Klinger, C. M.; Franklin, M.; Oler, E.; Wilson, A.; Pon, A.; Cox, J.; Chin, N. E. L.; Strawbridge, S. A.; Garcia-Patino, M.; Kruger, R.; Sivakumaran, A.; Sanford, S.; Doshi, R.; Khetarpal, N.; Fatokun, O.; Doucet, D.; Zubkowski, A.; Rayat, D. Y.; Jackson, H.; Harford, K.; Anjum, A.; Zakir, M.; Wang, F.; Tian, S.; Lee, B.; Liigand, J.; Peters, H.; Wang, R. Q. R.; Nguyen, T.; So, D.; Sharp, M.; da Silva, R.; Gabriel, C.; Scantlebury, J.; Jasinski, M.; Ackerman, D.; Jewison, T.; Sajed, T.; Gautam, V.; Wishart, D. S. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res.* **2024**, *52*, D1265−D1275.
(67) Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B. L.; Berjanskii, M.; Mah, R.; Yamamoto, M.; Jovel, J.; Torres-Calzada, C.; Hiebert-Giesbrecht, M.; Lui, V. W.; Varshavi, D.; Varshavi, D.; Allen, D.; Arndt, D.; Khetarpal, N.; Sivakumaran, A.; Harford, K.; Sanford, S.; Yee, K.; Cao, X.; Budinski, Z.; Liigand, J.; Zhang, L.; Zheng, J.; Mandal, R.; Karu, N.; Dambrova, M.; Schiöth, H. B.; Greiner, R.; Gautam, V. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.* **2022**, *50*, D622−D631.
(68) FOR-IDENT LC; https://water.for-ident.org/#!home; accessed 03/2024.
(69) Muir, D. C. G.; Howard, P. H. Are There Other Persistent Organic Pollutants? A Challenge for Environmental Chemists. *Environ. Sci. Technol.* **2006**, *40*, 7157−7166.
(70) Howard, P. H.; Muir, D. C. G. Identifying New Persistent and Bioaccumulative Organics Among Chemicals in Commerce II: Pharmaceuticals. *Environ. Sci. Technol.* **2011**, *45*, 6938−6946.
(71) 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.; Cirka, L.; D'Agostino, L. A.; Djoumbou Feunang, Y.; Dulio, V.; Fischer, S.; Gago-Ferrero, P.; Galani, A.; Geueke, B.; Głowacka, N.; Gluge, 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.; Rudel, 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. *Environmental Sciences Europe* **2022**, *34*, 104.
(72) 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. *Environmental Sciences Europe* **2018**, *30*, 5.
(73) McEachran, A. D.; Balabin, I.; Cathey, T.; Transue, T. R.; Al-Ghoul, H.; Grulke, C.; Sobus, J. R.; Williams, A. J. Linking in silico MS/MS spectra with chemistry data to improve identification of unknowns. *Scientific Data* **2019**, *6*, 141.
(74) Samanipour, S.; Dimitriou-Christidis, P.; Nabi, D.; Arey, J. S. Elevated Concentrations of 4-Bromobiphenyl and 1,3,5-Tribromo-

https://doi.org/10.1021/jacsau.4c00220
JACS Au 2024, 4, 2412−2425

benzene Found in Deep Water of Lake Geneva Based on GC× GC-ENCI-TOFMS and GC× GC- ECD. *ACS Omega* **2017**, *2*, 641−652.
(75) Egli, M.; Rapp-Wright, H.; Oloyede, O.; Francis, W.; Preston-Allen, R.; Friedman, S.; Woodward, G.; Piel, F. B.; Barron, L. P. A One-Health environmental risk assessment of contaminants of emerging concern in London's waterways throughout the SARS-CoV-2 pandemic. *Environ. Int.* **2023**, *180*, 108210.
(76) Samanipour, S.; Langford, K.; Reid, M. J.; Thomas, K. V. A two stage algorithm for target and suspect analysis of produced water via gas chromatography coupled with high resolution time of flight mass spectrometry. *Journal of Chromatography A* **2016**, *1463*, 153−161.
(77) 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.
(78) Fenner, K.; Scheringer, M.; Macleod, M.; Matthies, M.; McKone, T.; Stroebe, M.; Beyer, A.; Bonnell, M.; Le Gall, A. C.; Klasmeier, J.; Mackay, D.; Van De Meent, D.; Pennington, D.; Scharenberg, B.; Suzuki, N.; Wania, F. Comparing estimates o persistence and long-range transport potential among multimedia models. *Environ. Sci. Technol.* **2005**, *39*, 1932−1942.
(79) Wicker, J.; Lorsbach, T.; Gütlein, M.; Schmid, E.; Latino, D.; Kramer, S.; Fenner, K. enviPath − The environmental contaminant biotransformation pathway resource. *Nucleic Acids Res.* **2016**, *44*, D502−D508.
(80) Escher, B. I.; Fenner, K. Recent Advances in Environmental Risk Assessment of Transformation Products. *Environ. Sci. Technol.* **2011**, *45*, 3835−3847.
(81) Richardson, S. D. Invited Perspective: Existing Rules for Disinfection By-Products Are Good, but They Are Not Enough. *Environ. Health Perspect.* **2022**, *130*, 081302.
(82) Justen, P. T.; Kilpatrick, M. L.; Soto, J. L.; Richardson, S. D. Low Parts Per Trillion Detection of Iodinated Disinfection By-products in Drinking Water and Urine using Vacuum-Assisted Sorbent Extraction and GC−MS/MS. *Environ. Sci. Technol.* **2024**, *58*, 1321−1328.
(83) Samy, M.; Mensah, K.; Gar Alalm, M. A review on photodegradation mechanism of bio-resistant pollutants: Analytical methods, transformation products, and toxicity assessment. *Journal of Water Process Engineering* **2022**, *49*, 103151.
(84) Helbling, D. E.; Hollender, J.; Kohler, H.-P. E.; Singer, H.; Fenner, K. High-Throughput Identification of Microbial Transformation Products of Organic Micropollutants. *Environ. Sci. Technol.* **2010**, *44*, 6621−6627.
(85) Li, D.; Liang, W.; Feng, X.; Ruan, T.; Jiang, G. Recent advances in data-mining techniques for measuring transformation products by high-resolution mass spectrometry. *TrAC Trends in Analytical Chemistry* **2021**, *143*, 116409.
(86) Bletsou, A. A.; Jeon, J.; Hollender, J.; Archontaki, E.; Thomaidis, N. S. Targeted and non-targeted liquid chromatography-mass spectrometric workflows for identification of transformation products of emerging pollutants in the aquatic environment. *TrAC Trends in Analytical Chemistry* **2015**, *66*, 32−44.
(87) Manz, K. E.; Feerick, A.; Braun, J. M.; Feng, Y.-L.; Hall, A.; Koelmel, J.; Manzano, C.; Newton, S. R.; Pennell, K. D.; Place, B. J.; Godri Pollitt, K. J.; Prasse, C.; Young, J. A. Non-targeted analysis (NTA) and suspect screening analysis (SSA): a review of examining the chemical exposome. *J. Expo Sci. Environ. Epidemiol* **2023**, *33*, 524−536.
(88) Ciccarelli, D.; Christopher Braddock, D.; Surman, A. J.; Arenas, B. I. V.; Salal, T.; Marczylo, T.; Vineis, P.; Barron, L. P. Enhanced selectivity for acidic contaminants in drinking water: From suspect screening to toxicity prediction. *Journal of Hazardous Materials* **2023**, *448*, 130906.
(89) Mollerup, C. B.; Mardal, M.; Dalsgaard, P. W.; Linnet, K.; Barron, L. P. Prediction of collision cross section and retention time for broad scope screening in gradient reversed-phase liquid chromatography-ion mobility-high resolution accurate mass spectrometry. *Journal of Chromatography A* **2018**, *1542*, 82−88.
(90) 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.
(91) McEachran, A. D.; Mansouri, K.; Newton, S. R.; Beverly, B. E. J.; Sobus, J. R.; Williams, A. J. A comparison of three liquid chromatography (LC) retention time prediction models. *Talanta* **2018**, *182*, 371−379.
(92) Stanstrup, J.; Neumann, S.; Vrhovšek, U. PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. *Anal. Chem.* **2015**, *87*, 9421−9428.
(93) den Uijl, M. J.; Schoenmakers, P. J.; Schulte, G. K.; Stoll, D. R.; van Bommel, M. R.; Pirok, B. W. J. Measuring and using scanning-gradient data for use in method optimization for liquid chromatography. *Journal of Chromatography A* **2021**, *1636*, 461780.
(94) den Uijl, M. J.; Schoenmakers, P. J.; Pirok, B. W.; van Bommel, M. R. Recent applications of retention modelling in liquid chromatography. *J. Sep. Sci.* **2021**, *44*, 88−114.
(95) van Herwerden, D.; Nikolopoulos, A.; Barron, L.; O'Brien, J.; Pirok, B.; Thomas, K.; Samanipour, S. Exploring the Chemical Subspace of RPLC: a Data Driven Approach. *ChemRxiv*, 2023; DOI: 10.26434/chemrxiv-2023-bdwh0.
(96) Kruve, A. Strategies for Drawing Quantitative Conclusions from Nontargeted Liquid Chromatography−High-Resolution Mass Spectrometry Analysis. *Anal. Chem.* **2020**, *92*, 4691−4699.
(97) Yang, L.; Wu, J.; Zheng, M.; Cao, Z.; Li, C.; Shi, M.; Liu, G. Non-target screening of organic pollutants and target analysis of halogenated polycyclic aromatic hydrocarbons in the atmosphere around metallurgical plants by high-resolution GC/Q-TOF-MS. *Environmental Sciences Europe* **2020**, *32*, 96.
(98) Karasawa, K.; Duchoslav, E.; Baba, T. Fast Electron Detachment Dissociation of Oligonucleotides in Electron-Nitrogen Plasma Stored in Magneto Radio-Frequency Ion Traps. *Anal. Chem.* **2022**, *94*, 15510−15517.
(99) Paseiro-Cerrato, R.; Ackerman, L.; de Jager, L.; Begley, T. Brominated flame retardants (BFRs) in contaminated food contact articles: identification using DART-HRMS and GC-MS. *Food Additives & Contaminants: Part A* **2021**, *38*, 350−359.
(100) Puype, F.; Ackerman, L. K.; Samsonek, J. Evaluation of Direct Analysis in Real Time − High Resolution Mass Spectrometry (DART-HRMS) for WEEE specific substance determination in polymeric samples. *Chemosphere* **2019**, *232*, 481−488.
(101) Genangeli, M.; Heeren, R. M. A.; Porta Siegel, T. Tissue classification by rapid evaporative ionization mass spectrometry (REIMS): comparison between a diathermic knife and CO2 laser sampling on classification performance. *Anal Bioanal Chem.* **2019**, *411*, 7943−7955.
(102) David, A.; Chaker, J.; Price, E. J.; Bessonneau, V.; Chetwynd, A. J.; Vitale, C. M.; Klánová, J.; Walker, D. I.; Antignac, J.-P.; Barouki, R.; Miller, G. W. Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives. *Environ. Int.* **2021**, *156*, 106630.
(103) Gentry, E. C.; Collins, S. L.; Panitchpakdi, M.; Belda-Ferre, P.; Stewart, A. K.; Carrillo Terrazas, M.; Lu, H.-h.; Zuffa, S.; Yan, T.; Avila-Pacheco, J.; Plichta, D. R.; Aron, A. T.; Wang, M.; Jarmusch, A. K.; Hao, F.; Syrkin-Nikolau, M.; Vlamakis, H.; Ananthakrishnan, A. N.; Boland, B.; Hemperly, A.; Vande Casteele, N.; Gonzalez, F. J.; Clish, C. B.; Xavier, R. J.; Chu, H.; Baker, E. S.; Patterson, A. D.; Knight, R.; Siegel, D.; Dorrestein, P. C. Reverse metabolomics for the discovery of chemical structures from humans. *Nature* **2024**, *626*, 419−426.
(104) Rostkowski, P.; Haglund, P.; Aalizadeh, R.; Alygizakis, N.; Thomaidis, N.; Arandes, J. B.; Nizzetto, P. B.; Booij, 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*, 1957−1977.
(105) 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.
(106) 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. Cheminform* **2017**, *9*, 22.
(107) Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. *TrAC Trends in Analytical Chemistry* **2016**, *82*, 425−442.
(108) Helmus, R.; ter Laak, T. L.; van Wezel, A. P.; de Voogt, P.; Schymanski, E. L. patRoon: open source software platform for environmental mass spectrometry based non-target screening. *Journal of Cheminformatics* **2021**, *13*, 1.
(109) Samanipour, S.; Kaserzon, S.; Vijayasarathy, S.; Jiang, H.; Choi, P.; Reid, M. J.; Mueller, J. F.; Thomas, K. V. Machine learning combined with non-targeted LC-HRMS analysis for a risk warning system of chemical hazards in drinking water: A proof of concept. *Talanta* **2019**, *195*, 426−432.
(110) Schollée, J. E.; Schymanski, E. L.; Stravs, M. A.; Gulde, R.; Thomaidis, N. S.; Hollender, J. Similarity of High-Resolution Tandem Mass Spectrometry Spectra of Structurally Related Micropollutants and Transformation Products. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 2692−2704.
(111) Albergamo, V.; Schollée, J. E.; Schymanski, E. L.; Helmus, R.; Timmer, H.; Hollender, J.; de Voogt, P. Nontarget Screening Reveals Time Trends of Polar Micropollutants in a Riverbank Filtration System. *Environ. Sci. Technol.* **2019**, *53*, 7584−7594.
(112) Codrean, S.; Kruit, B.; Meekel, N.; Vughs, D.; Béen, F. Predicting the Diagnostic Information of Tandem Mass Spectra of Environmentally Relevant Compounds Using Machine Learning. *Anal. Chem.* **2023**, *95*, 15810−15817.
(113) Kind, T.; Fiehn, O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal Rev.* **2010**, *2*, 23−60.
(114) Andra, S. S.; Austin, C.; Patel, D.; Dolios, G.; Awawda, M.; Arora, M. Trends in the application of high-resolution mass spectrometry for human biomonitoring: An analytical primer to studying the environmental chemical space of the human exposome. *Environ. Int.* **2017**, *100*, 32−61.
(115) Neumann, S.; Böcker, S. Computational mass spectrometry for metabolomics: Identification of metabolites and small molecules. *Anal. Bioanal. Chem.* **2010**, *398*, 2779−2788.
(116) Rocca-Serra, P.; Salek, R. M.; Arita, M.; Correa, E.; Dayalan, S.; Gonzalez-Beltran, A.; Ebbels, T.; Goodacre, R.; Hastings, J.; Haug, K.; Koulman, A.; Nikolski, M.; Oresic, M.; Sansone, S.-A.; Schober, D.; Smith, J.; Steinbeck, C.; Viant, M. R.; Neumann, S. Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics* **2016**, *12*, 14.
(117) Hall, L. M.; Hill, D. W.; Menikarachchi, L. C.; Chen, M.-H.; Hall, L. H.; Grant, D. F. Optimizing artificial neural network models for metabolomics and systems biology: an example using HPLC retention index data. *Bioanalysis* **2015**, *7*, 939−955.
(118) Aalizadeh, R.; Alygizakis, N. A.; Schymanski, E. L.; Krauss, M.; Schulze, T.; Ibáñez, M.; McEachran, A. D.; Chao, A.; Williams, A. J.; Gago-Ferrero, P.; Covaci, A.; Moschet, C.; Young, T. M.; Hollender, J.; Slobodnik, J.; Thomaidis, N. S. Development and Application of Liquid Chromatographic Retention Time Indices in HRMS-Based Suspect and Nontarget Screening. *Anal. Chem.* **2021**, *93*, 11601.
(119) Aalizadeh, R.; Nika, M.-C.; Thomaidis, N. S. Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants. *Journal of Hazardous Materials* **2019**, *363*, 277−285.
(120) Kind, T.; Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **2007**, *8*, 105.
(121) Li, Y.; Kind, T.; Folz, J.; Vaniya, A.; Mehta, S. S.; Fiehn, O. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nat. Methods* **2021**, *18*, 1524−1531.
(122) Samanipour, S.; Reid, M. J.; Bæk, K.; Thomas, K. V. Combining a Deconvolution and a Universal Library Search Algorithm for the Nontarget Analysis of Data-Independent Acquisition Mode Liquid ChromatographyHigh-Resolution Mass Spectrometry Results. *Environ. Sci. Technol.* **2018**, *52*, 4694−4701.
(123) Woldegebriel, M.; Zomer, P.; Mol, H. G. J.; Vivó-Truyols, G. Application of Fragment Ion Information as Further Evidence in Probabilistic Compound Screening Using Bayesian Statistics and Machine Learning: A Leap Toward Automation. *Anal. Chem.* **2016**, *88*, 7705−7714.
(124) Alygizakis, N.; Lestremau, F.; Gago-Ferrero, P.; Gil-Solsona, R.; Arturi, K.; Hollender, J.; Schymanski, E. L.; Dulio, V.; Slobodnik, J.; Thomaidis, N. S. Towards a harmonized identification scoring system in LC-HRMS/MS based non-target screening (NTS) of emerging contaminants. *TrAC Trends in Analytical Chemistry* **2023**, *159*, 116944.
(125) Frainay, C.; Schymanski, E. L.; Neumann, S.; Merlet, B.; Salek, R. M.; Jourdan, F.; Yanes, O. Mind the Gap: Mapping Mass Spectral Databases in Genome-Scale Metabolic Networks Reveals Poorly Covered Areas. *Metabolites* **2018**, *8*, 51.
(126) Ruttkies, C.; Neumann, S.; Posch, S. Improving MetFrag with statistical learning of fragment annotations. *BMC Bioinformatics* **2019**, *20*, 376.
(127) 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*, 299−302.
(128) Allen, F.; Greiner, R.; Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **2015**, *11*, 98−110.
(129) Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. MSNovelist: de novo structure generation from mass spectra. *Nat. Methods* **2022**, *19*, 865−870.
(130) Wandy, J.; Zhu, Y.; van der Hooft, J. J. J.; Daly, R.; Barrett, M. P.; Rogers, S. Ms2lda.org: web-based topic modelling for substructure discovery in mass spectrometry. *Bioinformatics* **2018**, *34*, 317−318.
(131) Aron, A. T.; Gentry, E. C.; McPhail, K. L.; Nothias, L.-F.; Nothias-Esposito, M.; Bouslimani, A.; Petras, D.; Gauglitz, J. M.; Sikora, N.; Vargas, F.; van der Hooft, J. J. J.; Ernst, M.; Kang, K. B.; Aceves, C. M.; Caraballo-Rodríguez, A. M.; Koester, I.; Weldon, K. C.; Bertrand, S.; Roullier, C.; Sun, K.; Tehan, R. M.; P, C. A. B.; Christian, M. H.; Gutiérrez, M.; Ulloa, A. M.; Tejeda Mora, J. A.; Mojica-Flores, R.; Lakey-Beitia, J.; Vásquez-Chaves, V.; Zhang, Y.; Calderón, A. I.; Tayler, N.; Keyzers, R. A.; Tugizimana, F.; Ndlovu, N.; Aksenov, A. A.; Jarmusch, A. K.; Schmid, R.; Truman, A. W.; Bandeira, N.; Wang, M.; Dorrestein, P. C. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc.* **2020**, *15*, 1954−1991.
(132) Zhang, B.; Vogt, M.; Maggiora, G. M.; Bajorath, J. Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures. *J. Comput. Aided Mol. Des* **2015**, *29*, 937−950.
(133) Zhu, W.; Zhang, Y.; Zhao, D.; Xu, J.; Wang, L. HiGNN: A Hierarchical Informative Graph Neural Network for Molecular Property Prediction Equipped with Feature-Wise Attention. *J. Chem. Inf. Model.* **2023**, *63*, 43−55.
(134) Quinn, R. A.; Nothias, L.-F.; Vining, O.; Meehan, M.; Esquenazi, E.; Dorrestein, P. C. Molecular Networking As a Drug Discovery, Drug Metabolism, and Precision Medicine Strategy. *Trends Pharmacol. Sci.* **2017**, *38*, 143−154.

https://doi.org/10.1021/jacsau.4c00220
JACS Au 2024, 4, 2412−2425

(135) Yu, J.; Wang, J.; Zhao, H.; Gao, J.; Kang, Y.; Cao, D.; Wang, Z.; Hou, T. Organic Compound Synthetic Accessibility Prediction Based on the Graph Attention Mechanism. *J. Chem. Inf. Model.* **2022**, *62*, 2973−2986.
(136) Wójtowicz, A.; Wietecha-Posłuszny, R. DESI-MS analysis of human fluids and tissues for forensic applications. *Appl. Phys. A: Mater. Sci. Process.* **2019**, *125*, 312.
(137) Schäfer, K.-C.; Szaniszló, T.; Günther, S.; Balog, J.; Dénes, J.; Keserű, M.; Dezső, B.; Tóth, M.; Spengler, B.; Takáts, Z. In Situ, Real-Time Identification of Biological Tissues by Ultraviolet and Infrared Laser Desorption Ionization Mass Spectrometry. *Anal. Chem.* **2011**, *83*, 1632−1640.
(138) Wang, X.; Jiang, Q.; Li, H.; Chen, D. D. Y. Rapid determination of chemical composition in the particulate matter of cigarette mainstream smoke. *Talanta* **2020**, *217*, 121060.
(139) Gallidabino, M. D.; Hamdan, L.; Murphy, B.; Barron, L. P. Suspect screening of halogenated carboxylic acids in drinking water using ion exchange chromatography − high resolution (Orbitrap) mass spectrometry (IC-HRMS). *Talanta* **2018**, *178*, 57−68.
(140) Pereira, K. L.; Ward, M. W.; Wilkinson, J. L.; Sallach, J. B.; Bryant, D. J.; Dixon, W. J.; Hamilton, J. F.; Lewis, A. C. An Automated Methodology for Non-targeted Compositional Analysis of Small Molecules in High Complexity Environmental Matrices Using Coupled Ultra Performance Liquid Chromatography Orbitrap Mass Spectrometry. *Environ. Sci. Technol.* **2021**, *55*, 7365−7375.
(141) Yin, X.-B.; Li, Y.; Yan, X.-P. CE-ICP-MS for studying interactions between metals and biomolecules. *TrAC Trends in Analytical Chemistry* **2008**, *27*, 554−565.
(142) Camdzic, D.; Dickman, R. A.; Aga, D. S. Total and class-specific analysis of per- and polyfluoroalkyl substances in environmental samples using nuclear magnetic resonance spectroscopy. *Journal of Hazardous Materials Letters* **2021**, *2*, 100023.
(143) Camdzic, D.; Dickman, R. A.; Joyce, A. S.; Wallace, J. S.; Ferguson, P. L.; Aga, D. S. Quantitation of Total PFAS Including Trifluoroacetic Acid with Fluorine Nuclear Magnetic Resonance Spectroscopy. *Anal. Chem.* **2023**, *95*, 5484−5488.
(144) 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.
(145) McEneff, G. L.; Murphy, B.; Webb, T.; Wood, D.; Irlam, R.; Mills, J.; Green, D.; Barron, L. P. Sorbent Film-Coated Passive Samplers for Explosives Vapour Detection Part A: Materials Optimisation and Integration with Analytical Technologies. *Sci. Rep* **2018**, *8*, 5815.
(146) Emwas, A.-H.; Roy, R.; McKay, R. T.; Tenori, L.; Saccenti, E.; Gowda, G. A. N.; Raftery, D.; Alahmari, F.; Jaremko, L.; Jaremko, M.; Wishart, D. S. NMR Spectroscopy for Metabolomics Research. *Metabolites* **2019**, *9*, 123.
(147) Gathungu, R. M.; Kautz, R.; Kristal, B. S.; Bird, S. S.; Vouros, P. The integration of LC-MS and NMR for the analysis of low molecular weight trace analytes in complex matrices. *Mass Spectrom. Rev.* **2020**, *39*, 35−54.

https://doi.org/10.1021/jacsau.4c00220
JACS Au 2024, 4, 2412−2425