Scannotation: A Suspect Screening Tool for the Rapid Pre-Annotation of the Human LC-HRMS-Based Chemical Exposome

Jade Chaker, Erwann Gilles, Christine Monfort, Cécile Chevrier, Sarah Lennon, Arthur David

Scannotation: a suspect screening tool for the rapid pre-annotation of the human LC-HRMS-based chemical exposome

Jade Chaker[‡], Erwann Gilles[‡], Christine Monfort, Cécile Chevrier, Sarah Lennon, Arthur David*

Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) – UMR_S 1085, F-35000 Rennes, France

[‡] Shared authorship (co-first authors)

*To whom correspondence should be addressed:

Tel: +33 299022885

email: arthur.david@ehesp.fr

Abstract

In an increasingly chemically polluted environment, rapidly characterizing the human chemical exposome (i.e. chemical mixtures accumulating in humans) at the population scale is critical to understand its impact on health. High-resolution mass spectrometry (HRMS) profiling of complex biological matrices can theoretically provide a comprehensive picture of chemical exposures. However, annotating the detected chemical features, particularly low-abundant ones, remains a significant obstacle to implementing such approaches at a large scale. We present Scannotation (https://github.com/scannotation/Scannotation_software), an automated and user-friendly suspect screening tool for the rapid pre-annotation of HRMS preprocessed datasets. This software tool combines several MS1 chemical predictors, i.e., m/z, experimental and predicted retention times, isotopic patterns and neutral loss patterns, to score the proximity between features and suspects, thus efficiently prioritizing tentative annotations to verify. Scannotation and MS-DIAL4 were used to annotate blood serum samples of 75 Breton adolescents. Scannotation's combination of MS1-based chemical predictors allowed annotating 89 chemically diverse environmental compounds with high confidence (confirmed by MS2 when available). These compounds included 62% of emerging and unknown molecules, for which no toxicological and/or human biomonitoring data is reported in the literature. The complementarity observed with MS-DIAL4 results demonstrates the relevance of Scannotation for the efficient pre-annotation of large-scale exposomics datasets.

Introduction

Chemical pollution is a great and growing global problem, with more than tens of thousands of very diverse chemicals currently present on the market[1]. Human chemical exposure and toxicological data are only available for a few hundreds of these chemicals, meaning that a great share of chemicals potentially associated with deleterious health outcomes have not been investigated so far[1]. Nevertheless, the emergence of the exposome paradigm[2] as well as technological advances such as hyphenated high-resolution mass spectrometry techniques (e.g., LC-HRMS) have paved the way for the use of suspect screening (SS) and non-targeted screening (NTS) approaches, and therefore offer great promises for a comprehensive characterization of exogenous substances mixtures accumulating in humans[3–7]. It is however crucial to overcome the remaining methodological obstacles before implementing large-scale non-targeted exposomics studies to population-based studies. Indeed, the annotation of the tens of thousands of signals present in HRMS datasets remains one of the main bottlenecks, as only a few percent of signals are usually annotated[4].

Annotation of complex HRMS data can be performed using NTS, which relies on the structure elucidation of features, prioritized as differential between two (or more) groups, or SS, which relies on the annotation of features prioritized for their similarity to compounds listed in a suspect library. This second methodology is particularly promising, in part because it has a strong potential for automation and allows for a very rapid prioritization of signals of interest. Furthermore, there is no restriction regarding the number of suspects that can be included, as well as the forms they are searched as (i.e., parent or metabolite, adduct). The comparison of experimental features and suspects on characteristic properties such as their mass/charge ratio or their MS2 fragmentation profile can be automatically performed, before being manually validated. Even though some bioinformatics solutions already exist to carry out MS2-based spectral library searches, allowing the automatized comparison of reference and experimental MS2 spectra (e.g., xMSannotator, msPurity, MZmine2, MS-DIAL, patRoon,

CAMERA)[8–13], the number of software tools is still limited and not necessarily suited for exposomics applications[14], which present specific challenges detailed hereafter.

One of the main challenges of using biological matrices to characterize the human internal chemical exposome is the wide dynamic range of concentrations for compounds present in the samples. Compounds of interest in an exposomics context are often present at much lower levels (i.e., tens of pg/mL[15]) than many endogenous metabolites (i.e., up to a few mg/mL). These low-abundant xenobiotics do not systematically trigger MS2 acquisition[16]. This strongly limits the annotation's confidence level according to Schymanski's scale, which is the current reference[17]. Despite this fact, other factors accessible through MS1 data, such as retention time (Rt), distinctive isotope profiles based on halogen contents (often present in exogenous compounds such as pesticides), or detection of other phase I/II metabolites could already provide reliable indications on the annotation's plausibility. Moreover, while reference data, such as Rt from a standard or exhaustive knowledge of metabolism pathway, may not be available for each suspect compound, various models exist to provide predicted values for Rt[18–20] and structures for plausible metabolites[21]. Hence, these relevant MS1-based chemical predictors (experimental or predicted) can be combined into a substantial body of evidence pointing towards specific chemical identities for features of interest in exposomics applications. However, manually comparing these predictors between tens of thousands of suspects and features is a highly time-consuming task, so there is a need to develop tools that could speed up this process.

In this context, we developed the Scannotation software. It automatically scores the comparison between suspect compounds and features obtained from any pre-processing software[10,11,22] on three MS1-based chemical predictors: mass/charge ratio, isotopic ratios, and experimental or predicted retention time. It also generates common phase II metabolites and computes theoretical isotopic profiles for parent compounds and metabolites. Scannotation calculates and displays proximity scores between features and suspects called "confidence indices" for these three chemical predictors, as well

as an overall confidence index that effectively evaluates each automatized pre-annotation's reliability. This software allows the efficient prioritization of pre-annotations, which can then be validated manually. This prioritization saves considerable time, as it allows a fast prioritization of features to be further investigated to gain a higher confidence, decreases the amount of false positive annotations, thus contributing to the wider and faster pre-annotation of HRMS datasets. The list of suspects can easily be adjusted to the study since Scannotation can rapidly generate new predictors.

The efficiency of Scannotation was demonstrated in 75 serum samples from Breton adolescents (Pélagie cohort)[23]. These annotations were compared to those of MS-DIAL[11], which performs MS2-based annotations and was previously demonstrated to adequately and efficiently annotate low-abundant compounds in serum samples[16]. The use of these two tools allowed the annotation of 89 compounds from the internal chemical exposome of Breton adolescents, and demonstrated that the use of MS1 predictors was relevant and complementary to an MS2-based approach in an exposomic application.

Experimental section

1. Biological samples

Serum samples (n=75) were obtained from 12-year-old male children from the PELAGIE cohort. This population-based mother-child cohort included 3,421 women from Brittany (France) enrolled during early pregnancy (before 19 weeks of gestation) between 2002 and 2006[23].

2. Sample preparation

Samples were prepared using a dual sample preparation method, previously optimized to widen the visible chemical space[5]. Briefly, protein precipitation (PPT) was performed on all 75 samples using a 4:1 (v:v) ratio of cold methanol to matrix. After centrifugation, half of the supernatants was resuspended in 40 µL of injection phase (i.e., 90:10 (v:v) ultrapure water to acetonitrile ratio), while the second half was further cleaned up using a Phree (Phenomenex) protein and phospholipid removal

plate. Extracts were then evaporated to dryness under vacuum, and recovered in 40 µL of injection phase.

3. Data acquisition

Samples were analyzed using an AB SCIEX X500R QTOF-MS (Resolution > 30,000) interfaced with an AB SCIEX ExionLC AD UHPLC system. An Acquity UPLC HSS T3 C18 column (1.8 µm, 1.0 × 150 mm, Waters Corporation) maintained at 40°C was used to perform reverse phase chromatographic separation as described in Chaker *et al*. (2021)[16]. Samples were analyzed in full scan experiments in electrospray ionization (ESI) (–) and (+) modes. MS2 fragmentation data for chemical elucidation was obtained by analysis of selected samples in data-dependent acquisition experiments. Additional information regarding the chromatographic separation and ESI source parameters are available in the Supporting Information.

4. MS1 data pre-processing

Raw data acquired in full scan mode were pre-processed using vendor software MarkerView v.1.3 (AB SCIEX) and MSDIAL4 with parametrization previously optimized to detect low-abundant chemicals in blood serum samples[16]. Briefly, critical parameters values were set as: noise threshold of 10, mass tolerance of 10 ppm, retention time (Rt) tolerance of 1 min, average peak width of 12s, no isotope filtering. Feature areas in solvent blanks were systematically subtracted from corresponding feature areas in all samples.

5. Suspect screening

5.1. MS1-based suspect screening: Scannotation

Scannotation is a two-part Python program aimed at providing a prioritized list of scored pre-annotations available at https://github.com/scannotation/Scannotation_software. It was developed on Windows and tested on both Windows and on a Mac computer in a Windows virtual machine. Scannotation can process peak lists in .csv generated by any pre-processing software. The score is established by comparing the proximity between experimental features and suspects based on three

MS1-based chemical predictors (i.e., m/z, Rt, and isotopic pattern). Its operating principle is described in Figure 1.

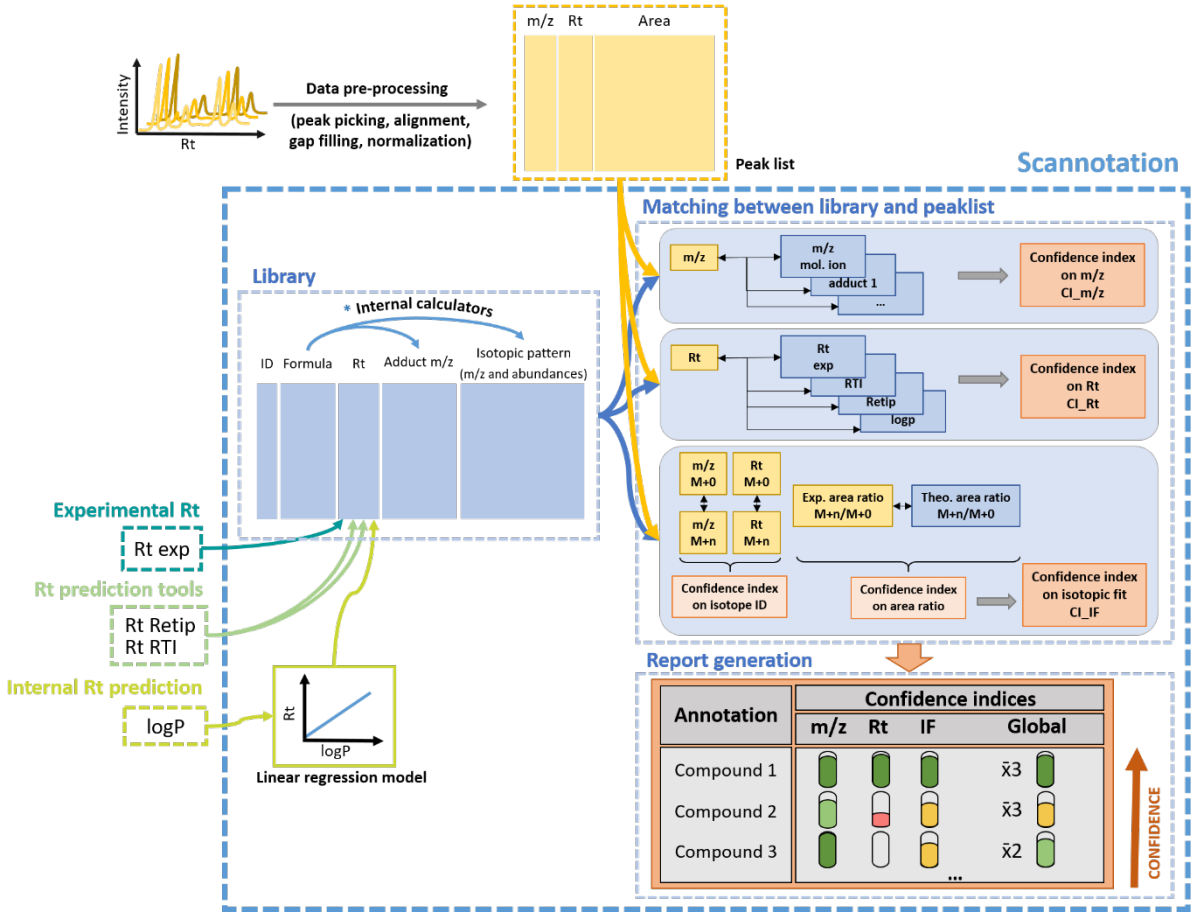


Fig. 1 - Scannotation annotation workflow relies on comparing a user-built library to a list of features. Compounds' identifiers (name and SMILES), molecular formula, experimental, predicted retention time (Rt) and logP values (when not listed by the user), allowing the software to compute molecular ion and adduct masses, theoretical isotopic pattern, and a logP-predicted Rt. Any other predicted Rt can also be added to the library and will be used by the software. The software then successively compares experimental features to the suspect library data for three predictors: m/z, Rt and isotopic fit. Scores are generated for each predictor, and combined into a global score.

The first module of this SS tool (the library) generates theoretical predictor data for a list of suspects, based on user-provided data, as illustrated in Fig. 1 by the "Internal calculators" arrows in the "Library"

box. More precisely, molecular formulas (i.e. atomic contents) and atomic masses (from MIDAs https://www.ncbi.nlm.nih.gov/CBBresearch/Yu/midas/index.html) are used to compute m/z values for molecular ions and various common adducts in both ESI (–) and (+) modes (i.e., $[M-H]^-$, $[M-H_2O-H]^-$, $[M+Cl]^-$, $[M+FA-H]^-$ for negatively charged ions, and ($[M+H]^+$, $[M+NH_4]^+$, $[M+Na]^+$, $[M+K]^+$ for positively charged ones). Moreover, these m/z values can be computed for common metabolites (i.e., glucuronide [+176.0321 m/z], sulfate [+79.9568 m/z], mercapturate [+145.0198 m/z] and cysteine conjugates [+103.0092 m/z]) of the library compounds by adding the corresponding expected mass increments[24]. The sulfate conjugate [+SO3] is added if O is present in the molecular formula and the glucuronide [+C6H8O6] conjugate is added if either N or O are present in the molecular formula. Molecular formulas are also used to generate theoretical isotopologues' m/z and relative abundance values by internally implementing the previously described MIDAs polynomial algorithm[25]. Users can provide experimental Rt values (if available), as well as predicted values from existing algorithms[18,19]. Moreover, octanol-water partition coefficient (logP) values provided by the user may be used to predict a logP-based Rt, under the condition that a sufficient number of compounds (i.e., over 20) have both an experimental Rt and a logP value. More details on this model are available in the SI. For library compounds without logP information, as well as for Scannotation-predicted metabolites, logP values are predicted through atomic composition-based model. This multiple linear regression model, adapted from Mannhold et al. (2009)[26], was built on close to 1,800 compounds with available experimental logP values ranging from -5.08 to 9.29 (list available in SI Table A1). This model was trained by randomly choosing 80 % of the dataset, then validated on the remaining 20%. The process was repeated 1,000 times. Median coefficients were used to determine the contribution of each atom. Model training determined that the most sensitive parameters to predict logP from atomic composition were the numbers of atoms of carbon, halogens, and sulfur (logP increasing with number increase), as well as nitrogen and oxygen (logP decreasing with number increase). This slightly more complex model (i.e., the original model only considers the number of carbons and the number of heteroatoms) performed better than the original model, with a root mean square error (RMSE) of 1.32

compared to 2.04. Similarly, the Bayesian Information Criteria (BIC) value was lower (i.e., better adjustment) for the more complex model, with a value of 766 compared to 4553 for the original model. Details on this model are available in SI Table A1.

A human blood library, including 6 000 compounds, was constructed using data from the literature[27] and online databases such as the Blood Exposome Database[28], Human Metabolome Database[29], Exposome Explorer[30], FoodBall[31], and the NORMAN Network Suspect List Exchange[32]. This library is mainly comprised of food intake biomarkers, pesticides (and their metabolites), industrial pollutants, cosmetic ingredients, and pharmaceuticals/drugs (and their metabolites). Scannotation's library module was used to add potential metabolites, predict logP-based Rt, and compute theoretical m/z values and isotopic patterns. Custom-made libraries can also be used.

Scannotation's second module performs the matching between the library and the peaklist. It computes confidence indices (CI), which score the proximity between suspects and features for each of the three predictors between 0 (no match) and 1 (perfect match). These scores are built as functions of the absolute difference between theoretical and experimental values of each predictor (i.e., m/z, Rt, mass difference between isotopologues and area ratio between isotopologues), and of a tolerance associated with each of these parameters[16]. These tolerances, called Δ, are either determined based on instrumental uncertainty (i.e., for m/z and m/z differences) or analytical variability (i.e., for Rt and area ratios). Additionally, for m/z and Rt, values of associated Δ (i.e., $\Delta_{m/z}$ and Δ_{Rt}) vary depending on the chemical descriptor value. Indeed, higher m/z deviations (in ppm) are expected for lower m/z values. $\Delta_{m/z}$ is set to 15 ppm for masses strictly lower than 200 Da, and 10 ppm for masses over 200 Da. Moreover, Scannotation allows the use of experimental Rt as well as Rt predicted through various tools, which all present different accuracies. Rt is also expected to vary in a non-linear manner throughout the chromatogram. Therefore, the used Δ_{Rt} value depends on both the type of Rt (i.e., experimental, predicted through RTI, predicted through Retip, or predicted through logP) and the value of the tentative annotation's theoretical Rt. In particular, Δ_{Rt} values for experimental Rt can be

computed by Scannotation from a user-provided file which includes repeated (at least four) Rt measurements for any list of compounds (e.g., internal standards). All Δ values used for this work are available in SI Table A2.

CI values for each predictor are combined into an overall score, named the global CI (CIg), as a mean of CI values to efficiently rank the pre-annotations generated by Scannotation. Moreover, the CI of annotated compounds, for which metabolites and/or neutral losses are detected, are associated to a letter from a to d (details available in SI Table A3). This leads to a better ranking in the result table and highlights the additional confidence in the tentative identification. The calculation of CI values and their adequation for suspect screening were previously described by Chaker et al. (2021)[16].

5.2. Scannotation was used to perform suspect screening on the feature tables obtained from the aforementioned MS1 data pre-processing step (see paragraph 4.MS1 data pre-processing"). Manual curation on MS1 and MS2 data (when available) was performed to confirm pre-annotations. This process included verification of absence or significantly lower presence in the blank (i.e., area ratio sample/blank > 10), signal-to-noise ratio > 10, visual examination of peak shape, and verification of accuracy of m/z, Rt, and isotopic pattern. When available, experimental MS2 spectra was compared to MS2 spectra in databases (e.g. HMDB, MassBank[29,33]) or in-silico predictions (e.g., CFM-ID, MetFrag[34,35]). When a standard was available, experimental parameters of features were compared to those of the standard. This manual curation was performed on all compounds presenting a CIg value calculated with three predictors, as well as all compounds presenting a CIg value over 70% if only two predictors were available. MS2-based suspect screening: MS-DIAL

In addition to Scannotation's screening, an MS2-based suspect screening approach was performed by processing raw data obtained from MS2 data dependent acquisitions with MS-DIAL (v.4.70)[11]. Critical parameters values were set as: minimum peak height of 10, mass tolerances of 0.0025 Da (10 ppm for a m/z of 250) in MS1 and MS2, Rt tolerance of 1 min, minimum peak width of 5 scans, and

consideration for Cl and Br elements enabled (for isotope recognition). Spectral MSP databases "All Public" available online (experimental spectra for 12,879 compounds in ESI (–) mode and 13,303 compounds in ESI (+) mode) were used for suspect screening. Manual curation (i.e., manual verifications of peak shape, m/z, Rt, isotopic pattern and MS2 spectra if available) on MS1 and MS2 data was performed to confirm annotations suggested by the software.

6. Quality assurance and quality control procedures

Several quality assurance and quality control procedures were implemented, including the systematic use of instrumental and extraction blanks, composite quality control samples, and ISTD. Details are available in SI.

Results and discussion

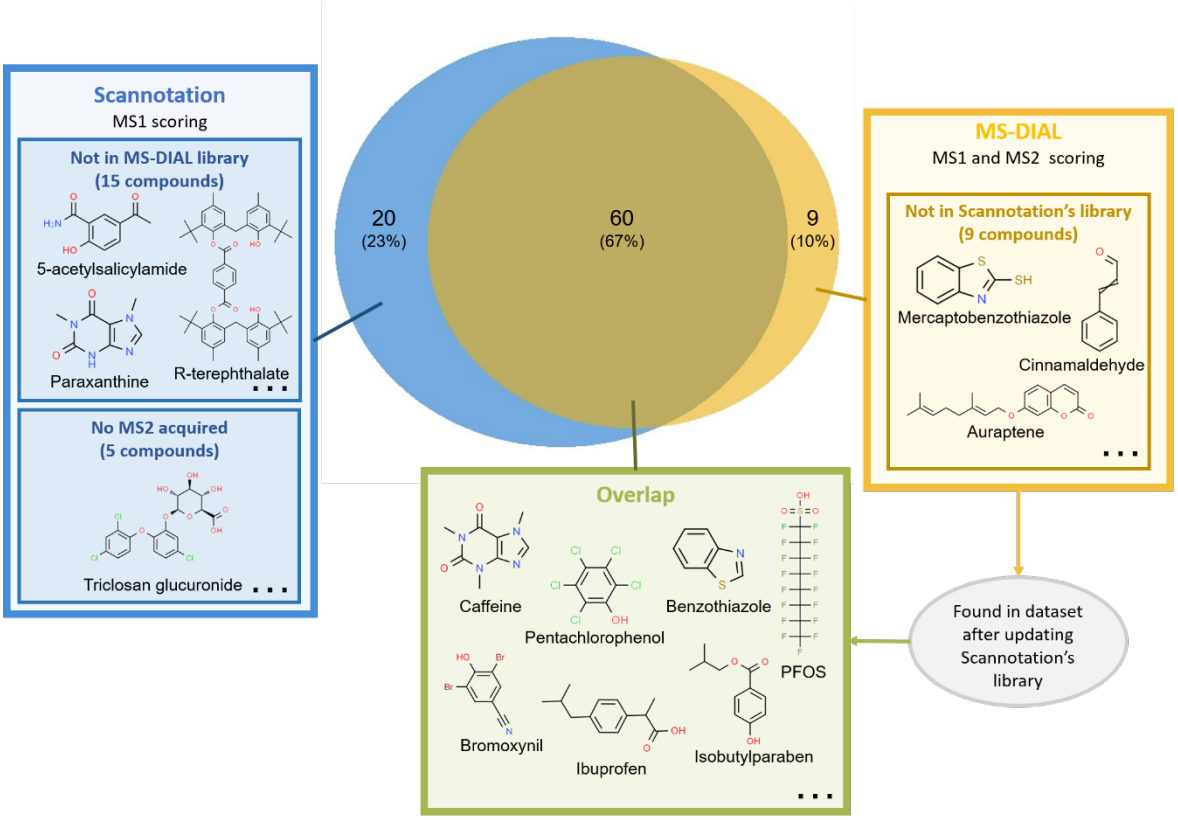
1. Comparison of MS1 and MS2-based suspect screening workflows

In total, 75 serum samples from a mother-child cohort were prepared with two different sample preparations[5] and injected on a UHPLC-ESI-QTOF. Quality control criteria were met (results presented in Supporting information table A6). Two suspect screening approaches were then implemented and compared to annotate the HRMS pre-treated datasets (i.e., approximately 50,000 and 93,000 features for ESI (−) and (+) modes respectively). Both software tools provided comparable numbers of raw pre-annotations. In total, Scannotation provided close to 33,000 pre-annotations while MS-DIAL provided close to 36,000. However, establishing a cut-off score of 70% for both tools (i.e., selecting pre-annotations with Scannotation CIg > 0.7 and MS-DIAL identification score > 70%) allowed reducing the number of tentative annotations by 78% using Scannotation while it only resulted in an 8% reduction for MS-DIAL. Reaching a similar level of prioritization for MS-DIAL can only be achieved by choosing perfect matches (i.e., scores of 100%). The discrepancy in scores is inherent to the algorithms used by both tools. Firstly, MS-DIAL incorporates a score based on MS2 spectral similarity, and Scannotation is currently only MS1-based. Secondly, MS-DIAL's Rt and m/z score calculations are significantly less

restrictive than Scannotation's, as an error of half the tolerance margin will provide an MS-DIAL score of 88%[11] and a Scannotation CI of 50%. This is because MS-DIAL considers the distribution of errors (presumed gaussian) whereas Scannotation's CI formula is linear to represent the gap between theoretical and experimental values as directly and accurately as possible. Thirdly, MS-DIAL's isotopic fit score is based on five ratios between six isotopologues, instead of one ratio between two isotopologues for Scannotation, which is arguably less stringent. These last two points clarify why Scannotation's CI values decrease more rapidly compared to MS-DIALs', thus explaining a more pronounced discriminating effect of scoring. As there are often limited resources (particularly time) that can be dedicated to manual curation, an efficient prioritization tool is essential to focus on plausible tentative annotations first. For this work, manual verifications were conducted on 8000 pre-annotations from Scannotation and 3000 pre-annotations from MS-DIAL, which approximately required 35h and 15h of work, respectively (i.e., comparable rates).

After manual curation (as detailed in the Experimental section, paragraphs 5.1. and 5.2.), 89 prioritized annotations were proposed with a confidence level ranging from 1 to 4 according to Schymanski's scale[17], with an overlap of 60 compounds for both software tools, as shown in Figure 2. The detailed prioritization process is available in SI Table A4 and the list of annotated compounds is available in SI Table A5. More specifically, of the 89 annotated compounds, 13 were attributed a confidence level of 1 (15%), 60 were attributed a level 2a (67%), 6 were attributed a level 2b (7%), and 10 were attributed a level 4 (11%) according to Schymanski's scale[17]. Despite the heterogeneity in the levels of confidence reported here (i.e., between 1 and 4 on a scale of 1 to 5), these 89 annotations are supported by multiple orthogonal elements of proof. For instance, the level 4 does not accurately reflect the confidence that can be put in these annotations since they were suggested based on a combination of several MS1 predictors, including a match with Rt from a standard (when available), identification of neutral loss, or the presence of additional metabolite from the same parent compound. This is the case for the herbicide bromoxynil, which could not be fragmented in the samples during MS2 acquisitions, but for which the match between the experimental Rt of the annotated ion and of the standard's was

298 scored at 88% by Scannotation (CIg = 0.84), or for the triclosan glucuronide and sulfate as discussed
299 below. It should be noted that features for these 89 compounds were correctly peak picked by both
300 data processing software tools (i.e. checked after manual observations); observed differences in
301 compound lists can therefore be attributed to the annotation process.



302

303 *Fig. 2 - Overview of the data generated by two suspect screening tools, based on either MS1 or both*
304 *MS1 and MS2 predictors (Scannotation and MS-DIAL respectively).*

305 Scannotation was able to annotate 90% of these 89 compounds, including 22% that were exclusively
306 identified through this software. Among the compounds only annotated by Scannotation, five (e.g.,
307 triclosan sulfate and glucuronide or 2-chlorophenol) were either too low abundant or not fragmented,
308 and did not present reliable MS2 spectra. Since MS-DIAL mainly relies on matching MS2 spectra, these
309 compounds could therefore be missed and/or not prioritized by this solution. Even without MS2 data,
310 Scannotation was able to provide solid annotation for triclosan sulfate (CIg = 0.93) and glucuronide
311 (CIg = 0.92) based on Rt match with pure standards, specific isotopic profiles match (3 atoms of Cl),

identification of neutral loss (i.e., triclosan glucuronide and sulfate conjugates) and the presence of 2 phase II conjugates from the same parent with a coherent elution order. These elements of proof are presented in SI Fig.S1. Moreover, for the 15 remaining compounds not annotated by MS-DIAL, there was no adequate reference spectra in the library. No adequate reference spectra may either mean no spectra at all, as for 5-acetylsalicylamide, or spectra acquired in the other ionization mode, or with different MS2 acquisition modes and/or collision energies, leading to very poor spectra matching (and thus elimination from the pool of suggested annotations), as for paraxanthine.

MS-DIAL annotated 77% of the 89 compounds, including 10% of them solely identified through this software. This was explained by the absence of these compounds in Scannotation's initial library. Subsequent addition of these 9 compounds to the library (i.e., molecular formula, identifiers, predicted Rt and logP values when available) allowed their annotation by MS1-based predictors, resulting in a mean CIg of 0.86 (all above 0.79). Many other highly scored putative annotations were suggested by MS-DIAL; however, they were mainly endogenous compounds. This was expected since a large proportion of the available MS2 spectra in MS-DIAL's databases are of endogenous compounds that were out of the scope of this work.

The Schymanski's scale is an undeniably useful tool to efficiently communicate confidence of annotations in a harmonized and easy-to-read way. However, in the context of exposomics applications with low-abundant compounds (particularly in complex matrices such as biological matrices), using MS2 predictors may face various critical obstacles. For instance, it is fairly frequent that there is either no MS2 data acquisition (i.e., acquisition not triggered due to low abundance) or no reliable MS2 spectra (i.e., acquisition triggered, but collision energy too low or compound too diluted to produce useful fragmentation data) for ions of interest. This lack of reliable MS2 data leads to the attribution of low confidence levels to compounds that may be relevant in a public health context due to their high toxicity and/or large prevalence in the population. When these issues arise, the use of other predictors based on MS1 could be relevant to differentiate between level 4

compounds ranging from bare formula matches to annotations supported by other elements such as coherent Rt and presence of related biotransformation products, and therefore to efficiently prioritize the massive number of suggested annotations for manual curation. This can also be helpful to create inclusion lists of ions of interest for possible further MS2 acquisitions.

In the end, we demonstrate that Scannotation is complementary to existing MS2-based suspect screening since both Scannotation and MS-DIAL were able to annotate a large number of exogenous chemicals in these 75 serum samples with a large majority (67%) being detected by both. We also demonstrate the importance of having a relevant library dedicated to HRMS-based exposomics studies.

2. Description of chemical exposure profiles in the Pélagie cohort

2.1. Environmental chemical exposures in the cohort

The data collected on the 75 samples injected in both ESI (-) and ESI (+) modes allowed annotating 89 compounds from the internal chemical exposome. In these chemicals, several “everyday pollutants” commonly found in human biological samples (at varying detection frequencies and levels) were annotated, such as phthalates, paraben derivatives, pesticides, and per- and poly-fluorinated alkyl substances[36,37]. Interestingly, bromoxynil, a well-characterized herbicide and established endocrine disruptor[38] found in 64% of samples, had previously been reported in the urine of 22% of pregnant women from the same cohort (i.e., during the prenatal period of these 75 teenagers)[39]. This suggests that some individuals have been chronically exposed (or at repeated occasions) to this compound, including during this specific period of vulnerability. Moreover, lidocaine (local anesthetic) was surprisingly found in more than 90% of samples. After further investigation, it was determined that anesthetic patches containing lidocaine were used 1hour prior to the blood draw, thus confirming the relevance of the presented workflow to detect environmental exposures.

These annotations of commonly detected chemicals also demonstrate the sensitivity and the relevance of this HRMS-based profiling method and SS strategy to identify chemicals usually detected at trace levels using conventional targeted MS2 method. Levels of insecticide metabolite fipronil sulfone (detected in 35% of samples) were previously reported in human blood from the general population at concentrations varying between 0.1 and 4 ng/mL[40]. Similarly, bromoxynil levels in plasma samples from teenagers residing in rural areas were previously reported from trace levels to 140 ng/mL[41]. Likewise, previously reported levels of perfluorooctanesulfonic acid (PFOS) and perfluorohexanesulfonic acid (PFHxS) (detection frequencies of 100% and 95% respectively) in the German general population from 2009 to 2019 ranged from 0.9-9.9 ng/mL and from 0-4.6 ng/mL respectively[42]. As the MS1-based suspect screening approach allowed the annotation of all of these compounds, it appears that exogenous chemicals present at low levels in complex matrices may still be identified with a suspect screening strategy using Scannotation.

To provide a global overview of the exposure profiles, the 89 detected compounds were classified in four general categories: gut microbiota metabolites (including those from exogenous dietary substrates), food compounds (natural and artificial), medication and personal care compounds (e.g., pain management, surfactants), and industrial compounds (e.g., synthesis intermediates used in the manufacturing of dyes, pesticides or plasticizers). Most of the 89 annotated compounds have multiple sources, such as ferulic acid, which is both a natural compound and a food preservative, or di(ethylhexyl)phthalate, which is a plasticizer present in many plastic products including flooring and upholstering, everyday household items, and food packaging. However, for illustrating purposes, primary uses (according to production volume) were considered for the proposed classifications. Gut microbiota, food compounds, medication and personal care compounds, and industrial compounds represented, respectively, 7%, 46%, 16%, and 31% of the overall number of identified compounds. Taken together, food compounds and medical and personal care products represented close to two-thirds (62%) of annotated compounds. The first category includes natural compounds (e.g. caffeine, piperine), flavoring agents (e.g. sweeteners aspartame and sucralose), and food contact chemicals (e.g.

1,3,5-tris(2,2-dimethylpropionylamino)benzene), while the second includes medication and metabolites (e.g. acetaminophen phase II metabolites), additives from personal care products (e.g. shampoo and shower gel surfactant cocamidopropyl betaine), and preservatives (e.g. isobutyl- and isopropylparaben). Detecting many compounds from these categories was expected, as their concentrations in blood can be up to 10^6 times higher than some industrial pollutants (such as pesticides or plasticizers)[15], thus leading to easier detection and characterization (in particular, easier acquisition of MS2 spectra).

2.2. Inter-individual chemical exposure variability

In order to study the inter-individual variability that could be observed in terms of chemical exposure profiles, the presence or absence of each annotated compound in analyzed samples was assessed. The number of annotated compounds from each category and sub-category is represented in Figure 3.

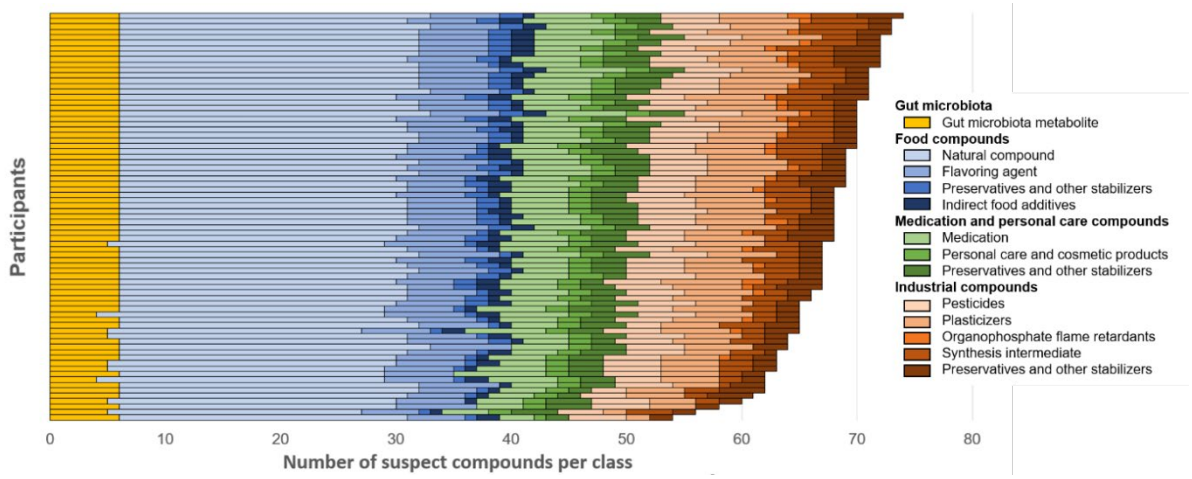


Fig. 3 - Detection of suspect compounds in each participant. Compounds were classified in four main categories: gut microbiota metabolites, food compounds, medication and personal care compounds, and industrial compounds.

It should be noted that most compounds have multiple possible uses. The classification presented here is based on primary use according to PubChem[43]. Secondary uses for each compound are also presented in the SI, table A4.

A median of 66 compounds were detected per sample. At the scale of the four main categories, food compounds are the most represented with a median of 52% (CV=4%) of annotated compounds per individual. The less represented category was gut microbiota-related chemicals, with a median of 9% of annotated compounds per individual (CV=8%). Industrial compounds represented a median of a quarter of all annotated compounds per individual, with a CV of 10%.

At the scale of subcategories, natural food compounds regrouped the largest proportion of annotated compounds (median of 38% of annotated compounds per individual). On the other hand, the less represented chemical class is organophosphate flame retardants (e.g., tris(2-butoxyethyl)phosphate) (median of 0%, average of 1%). Exposure profiles (i.e., combining exposures to different compounds) may be indicative of an individual's lifestyle. For example, 12 participants presented a co-exposure to acesulfame, aspartame and sucralose (all artificial sweeteners), which may indicate an overall more processed diet. It should however be noted that the high number of annotated compounds in comparison to the number of participants significantly limits the statistical power necessary to establish such profiles. It is also impossible to identify the source of these exposures with certainty, as this would require analyses of environmental samples (i.e., indoor air, items from diet, personal care products, etc.) and/or data obtained from questionnaires in addition to the data acquired from these biological samples.

Our results also highlight the capacity of suspect screening approaches to discover new relevant biomarkers of exposure to known toxicants. Bromoxynil metabolite 3,5-dibromo-4-hydroxybenzoic acid was detected in 97% of samples with areas 3 to 8 times higher than areas for bromoxynil (detected in 64% of samples). However, despite its apparent easier detection, this metabolite was not reported (as a biomarker of bromoxynil exposure or otherwise) in human biomonitoring studies in blood or urine before. These new biomarkers may subsequently be used either to retrospectively assess exposure if non-targeted data was acquired or may be included in new lists of targeted compounds of interest.

These exploratory approaches allow to significantly increase our knowledge of the chemical exposome, and particularly to investigate emerging and unknown compounds (without toxicological and/or human biomonitoring data available). Detection frequencies for all compounds were computed. Detailed results are available in SI Table A5. Overall, 66% of all annotated compounds were found in more than 80% of samples. Of these ubiquitous compounds, 17% (10 compounds) are not documented in the extensive NORMAN Network's SUSDAT list, which combines close to 110,000 structures from 98 suspect lists provided by the scientific community[32]. These compounds include 3 phase II metabolites (sulfated forms), highlighting the need to include biotransformation products (known or predicted) in suspect lists. Furthermore, 20% (11 compounds) of these ubiquitous compounds have no reported toxicological data according to the CompTox chemistry dashboard[44] (Figure 4). For instance, a phthalate found in 92% of samples (i.e., Bis(2-(tert-butyl)-6-(3-(tert-butyl)-2-hydroxy-5-methylbenzyl)-4-methylphenyl) terephthalate) and only annotated with Scannotation has no reported toxicological data, even though phthalates are known endocrine and metabolic disruptors[45]. This again demonstrates that MS1-based suspect screening approaches can be of great use to uncover previously unknown or poorly known exposures to chemicals of potential concern. These compounds, while largely predominant in number in reality (more than 110 million compounds registered on PubChem[43] compared to less than a thousand biomonitored chemical species in some of the biggest human biomonitoring initiatives[46,47]), are not investigated by targeted approaches. The use of exploratory approaches focused on low-abundant chemicals, in terms of data acquisition (i.e., non-targeted analyses), data processing (i.e., adequate peak picking parameters), and annotation (i.e., annotation even without MS2 acquisitions), may help starting to bridge this gap in knowledge by uncovering emerging and unknown compounds in the studied population (Figure 4).

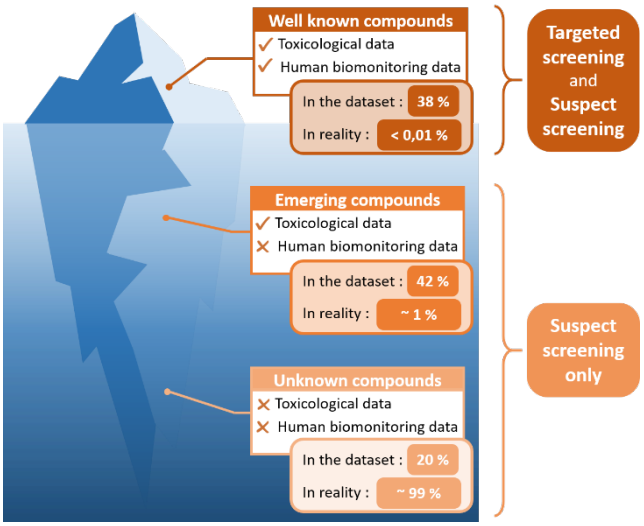


Fig. 4 – Contribution of suspect screening approaches to the number of exogenous compounds detected in human serum samples from the Pélagie cohort. Proportion of compounds in reality are based on the total number of Pubchem entries (>110M), the number of CompTox chemistry dashboard entries (~1M), and the number of chemicals usually biomonitored by the largest human biomonitoring initiatives to date (<1000).

Conclusion

Here, we demonstrate the efficiency of Scannotation to investigate the internal chemical exposome of 75 Breton adolescents using a MS1-based strategy to score the proximity between features obtained from any pre-processing software and suspects, and providing an easy-to-read indicator of each pre-annotation's reliability. This strategy was complementary to the one used by MS-DIAL based on MS2, and Scannotation provided thousands of scored pre-annotations that led to the annotation of 89 environmental chemical compounds (confirmed with manual curation) with various uses including pesticides, medication, preservatives and synthesis intermediates. It has also uncovered the relevance of Scannotation's SS strategy to identify low-abundant and/or less documented compounds (not annotated by MS-DIAL) or to detect new metabolites of known contaminants. We demonstrate that this approach will help bridging a gap in knowledge by documenting the prevalence of some emerging and unknown compounds in a given population. The chemical fingerprints acquired, and the list of

annotated compounds could be further used in association to contextual data from the cohort, to further describe the chemical exposome of the Breton teenage population, and to investigate determinants of these exposures.

Acknowledgment

This research was supported by a research chair of excellence (2016-52/IdeX Université of Sorbonne Paris Cité) awarded to AD and a grant from the Brittany council (SAD). J.C. was funded by the Réseau Doctoral en Santé Publique. J.C., S.L. and A.D. acknowledge the research infrastructure France Exposome.

Supporting information

The Supporting Information is available free of charge at http://pubs.acs.org.

Description of sample preparation method, data acquisition parameters, quality control procedures, and elements of proof supporting the annotation of triclosan glucuronide (.docx)

Predictive Rt model, parameter tolerance values, codification of global confidence index, and list of annotations performed on cohort samples (.xlsx)

References

1. Fuller R, Landrigan PJ, Balakrishnan K, et al. Pollution and health: a progress update. The Lancet Planetary Health 2022;6(6):e535–e547; doi: 10.1016/S2542-5196(22)00090-0.

2. Wild CP. Complementing the Genome with an “Exposome”: The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. Cancer Epidemiology, Biomarkers & Prevention 2005;14(8):1847–1850; doi: 10.1158/1055-9965.EPI-05-0456.

3. Tkalec Ž, Codling G, Klánová J, et al. LC-HRMS based method for suspect/non-targeted screening for biomarkers of chemical exposure in human urine. Chemosphere 2022;300:134550; doi: 10.1016/j.chemosphere.2022.134550.

4. Panagopoulos Abrahamsson D, Wang A, Jiang T, et al. A Comprehensive Non-targeted Analysis Study of the Prenatal Exposome. Environ Sci Technol 2021;55(15):10542–10557; doi: 10.1021/acs.est.1c01010.

5. Chaker J, Kristensen DM, Halldorsson TI, et al. Comprehensive Evaluation of Blood Plasma and Serum Sample Preparations for HRMS-Based Chemical Exposomics: Overlaps and Specificities. Anal Chem 2022;94(2):866–874; doi: 10.1021/acs.analchem.1c03638.

6. Al-Salhi R, Monfort C, Bonvallot N, et al. Analytical strategies to profile the internal chemical exposome and the metabolome of human placenta. Analytica Chimica Acta 2022;1219:339983; doi: 10.1016/j.aca.2022.339983.

7. Pourchet M, Debrauwer L, Klanova J, et al. Suspect and non-targeted screening of chemicals of emerging concern for human biomonitoring, environmental health studies and support to risk assessment: From promises to challenges and harmonisation issues. Environment International 2020;139:105545; doi: 10.1016/j.envint.2020.105545.

8. Uppal K, Walker DI, Jones DP. xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data. Anal Chem 2017;89(2):1063–1067; doi: 10.1021/acs.analchem.6b01214.

9. Lawson TN, Weber RJM, Jones MR, et al. msPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics. Anal Chem 2017;89(4):2432–2439; doi: 10.1021/acs.analchem.6b04358.

10. Pluskal T, Castillo S, Villar-Briones A, et al. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 2010;11(1):395; doi: 10.1186/1471-2105-11-395.

11. Tsugawa H, Cajka T, Kind T, et al. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods 2015;12(6):523–526; doi: 10.1038/nmeth.3393.

12. Helmus R, ter Laak TL, van Wezel AP, et al. patRoon: open source software platform for environmental mass spectrometry based non-target screening. Journal of Cheminformatics 2021;13(1):1; doi: 10.1186/s13321-020-00477-w.

13. Kuhl C, Tautenhahn R, Böttcher C, et al. CAMERA: An integrated strategy for compound spectra extraction and annotation of LC/MS data sets. Anal Chem 2012;84(1):283–289; doi: 10.1021/ac202450g.

14. Misra BB. New software tools, databases, and resources in metabolomics: updates from 2020. Metabolomics 2021;17(5):49; doi: 10.1007/s11306-021-01796-1.

15. David A, Chaker J, Price EJ, et al. Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives. Environment International 2021;156:106630; doi: 10.1016/j.envint.2021.106630.

16. Chaker J, Gilles E, Léger T, et al. From Metabolomics to HRMS-Based Exposomics: Adapting Peak Picking and Developing Scoring for MS1 Suspect Screening. Analytical Chemistry 2021;93(3):1792–1800; doi: 10.1021/acs.analchem.0c04660.

17. Schymanski EL, Jeon J, Gulde R, et al. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. Environ Sci Technol 2014;48(4):2097–2098; doi: 10.1021/es5002105.

18. Bonini P, Kind T, Tsugawa H, et al. Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. Anal Chem 2020;92(11):7515–7522; doi: 10.1021/acs.analchem.9b05765.

19. Aalizadeh R, Thomaidis NS, Bletsou AA, et al. Quantitative Structure–Retention Relationship Models To Support Nontarget High-Resolution Mass Spectrometric Screening of Emerging Contaminants in Environmental Samples. J Chem Inf Model 2016;56(7):1384–1398; doi: 10.1021/acs.jcim.5b00752.

20. Stanstrup J, Neumann S, Vrhovšek U. PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. Anal Chem 2015;87(18):9421–9428; doi: 10.1021/acs.analchem.5b02287.

21. Wishart DS, Tian S, Allen D, et al. BioTransformer 3.0—a web server for accurately predicting metabolic transformation products. Nucleic Acids Research 2022;50(W1):W115–W123; doi: 10.1093/nar/gkac313.

22. Smith CA, Want EJ, O'Maille G, et al. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Anal Chem 2006;78(3):779–787; doi: 10.1021/ac051437y.

23. Garlantezec R, Monfort C, Rouget F, et al. Maternal occupational exposure to solvents and congenital malformations: a prospective study in the general population. Occupational and Environmental Medicine 2009;66(7):456–463; doi: 10.1136/oem.2008.041772.

24. Huber C, Nijssen R, Mol H, et al. A large scale multi-laboratory suspect screening of pesticide metabolites in human biomonitoring: From tentative annotations to verified occurrences. Environment International 2022;168:107452; doi: 10.1016/j.envint.2022.107452.

25. Alves G, Ogurtsov AY, Yu Y-K. Molecular Isotopic Distribution Analysis (MIDAs) with Adjustable Mass Accuracy. J Am Soc Mass Spectrom 2014;25(1):57–70; doi: 10.1007/s13361-013-0733-7.

26. Mannhold R, Poda GI, Ostermann C, et al. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. J Pharm Sci 2009;98(3):861–893; doi: 10.1002/jps.21494.

27. Rappaport SM, Barupal DK, Wishart D, et al. The blood exposome and its role in discovering causes of disease. Environ Health Perspect 2014;122(8):769–774; doi: 10.1289/ehp.1308015.

28. Barupal DK, Fiehn O. Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach. Environmental Health Perspectives n.d.;127(9):097008; doi: 10.1289/EHP4713.

29. Wishart DS, Feunang YD, Marcu A, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Research 2018;46(D1):D608–D617; doi: 10.1093/nar/gkx1089.

30. Neveu V, Moussy A, Rouaix H, et al. Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. Nucleic Acids Res 2017;45(D1):D979–D984; doi: 10.1093/nar/gkw980.

31. Fiamoncini J, Weinert C, Dragsted LO, et al. The FoodBAll Online Resources to Support Discovery of Novel Dietary Biomarkers with Metabolomics. 2015.

32. Mohammed Taha H, Aalizadeh R, Alygizakis N, et al. 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(1):104; doi: 10.1186/s12302-022-00680-6.

33. Horai H, Arita M, Kanaya S, et al. MassBank: a public repository for sharing mass spectral data for life sciences. Journal of Mass Spectrometry 2010;45(7):703–714; doi: 10.1002/jms.1777.

34. Wang F, Allen D, Tian S, et al. CFM-ID 4.0 – a web server for accurate MS-based metabolite identification. Nucleic Acids Research 2022;50(W1):W165–W174; doi: 10.1093/nar/gkac383.

35. Ruttkies C, Schymanski EL, Wolf S, et al. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. Journal of Cheminformatics 2016;8(1):3; doi: 10.1186/s13321-016-0115-9.

36. Richterová D, Govarts E, Fábelová L, et al. PFAS levels and determinants of variability in exposure in European teenagers – Results from the HBM4EU aligned studies (2014–2021). International Journal of Hygiene and Environmental Health 2023;247:114057; doi: 10.1016/j.ijheh.2022.114057.

37. Hartmann C, Jamnik T, Weiss S, et al. Results of the Austrian Children's Biomonitoring Survey 2020 part A: Per- and polyfluorinated alkylated substances, bisphenols, parabens and other xenobiotics. International Journal of Hygiene and Environmental Health 2023;249:114123; doi: 10.1016/j.ijheh.2023.114123.

38. Arena M, Auteri D, Barmaz S, et al. Peer review of the pesticide risk assessment of the active substance bromoxynil (variant evaluated bromoxynil octanoate). EFSA J 2017;15(6):e04790; doi: 10.2903/j.efsa.2017.4790.

39. Bonvallot N, Jamin EL, Regnaut L, et al. Suspect screening and targeted analyses: Two complementary approaches to characterize human exposure to pesticides. Science of The Total Environment 2021;786:147499; doi: 10.1016/j.scitotenv.2021.147499.

40. McMahen RL, Strynar MJ, Dagnino S, et al. Identification of fipronil metabolites by time-of-flight mass spectrometry for application in a human exposure study. Environ Int 2015;78:16–23; doi: 10.1016/j.envint.2015.01.016.

41. Semchuk K, McDuffie H, Senthilselvan A, et al. Body mass index and bromoxynil exposure in a sample of rural residents during spring herbicide application. J Toxicol Environ Health A 2004;67(17):1321–1352; doi: 10.1080/15287390490471424.

42. Göckener B, Weber T, Rüdel H, et al. Human biomonitoring of per- and polyfluoroalkyl substances in German blood plasma samples from 1982 to 2019. Environ Int 2020;145:106123; doi: 10.1016/j.envint.2020.106123.

43. Kim S, Chen J, Cheng T, et al. PubChem 2023 update. Nucleic Acids Research 2023;51(D1):D1373–D1380; doi: 10.1093/nar/gkac956.

44. Williams AJ, Grulke CM, Edwards J, et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. Journal of Cheminformatics 2017;9(1):61; doi: 10.1186/s13321-017-0247-6.

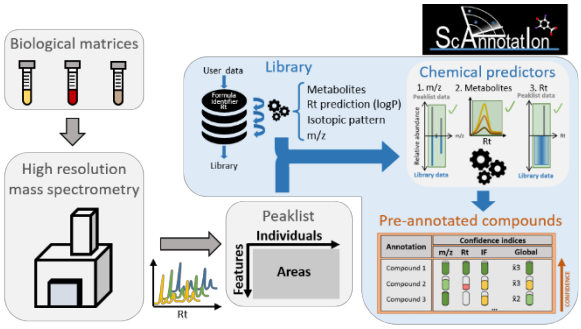
45. Hlisníková H, Petrovičová I, Kolena B, et al. Effects and Mechanisms of Phthalates' Action on Reproductive Processes and Reproductive Health: A Literature Review. Int J Environ Res Public Health 2020;17(18):6811; doi: 10.3390/ijerph17186811.

46. Vorkamp K, Castaño A, Antignac J-P, et al. Biomarkers, matrices and analytical methods targeting human exposure to chemicals selected for a European human biomonitoring initiative. Environment International 2021;146:106082; doi: 10.1016/j.envint.2020.106082.

47. CDC. Fourth National Report on Human Exposure to Environmental Chemicals. 2022.

637

638 For Table of Contents Only



639

640