High Resolution Mass Spectrometry To Investigate the Human Exposome: Where Are We?

Begoña Talavera Andújar[1]* & Emma L. Schymanski[1]*

[1] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6, Avenue du Swing, L-4367, Belvaux, Luxembourg

*Corresponding authors: begona.talavera@uni.lu & emma.schymanski@uni.lu

ORCIDs: BTA: 0000-0002-3430-9255; ELS: 0000-0001-6868-8145.

Abstract

Despite the significant role of the environment in health and disease, the accurate assessment of environmental exposures remains underdeveloped compared to genetic factors. To address this, the concept of the exposome was introduced in 2005 as a complement to the genome. High-resolution mass-spectrometry (HRMS) has emerged as a key technology for the comprehensive assessment of the chemical exposome. However, non-target HRMS based exposomics still faces numerous challenges, with the majority of features detected by HRMS (the "dark matter" of the chemical exposome) remaining unannotated. The lack of standardized workflows across the field often results in poorly comparable studies. Nevertheless, many positive developments have arisen in recent years, with open data revealing interesting trends in HRMS coverage. This review will examine and discuss the entire non-target HRMS exposomics workflow, from experimental design (study design and sample preparation) to the computational analysis and biological interpretation. It will also delve into key concepts, including the sometimes blurred distinction between the metabolome and the chemical exposome and the importance of exposomics within the "omics cascade". Visualizations are used to support this discussion, including a detailed look at the chemical coverage of key categories of open exposomics resources. The review ends by exploring current challenges and strategies to advance towards harmonized exposomics studies, which are essential for greater biological insights and personalized medicine goals.

Keywords: exposomics, non-target screening, high-resolution mass-spectrometry (HRMS), liquid chromatography (LC), cheminformatics

1. Introduction

The phenotype of an individual arises from the interplay between genes and environment. Since only a small proportion of chronic diseases can be attributed solely to genetic factors, it is now hypothesized that the majority (70-90%) are influenced by environmental factors, many of which remain unknown [1, 2]. Despite the significant role of the environment in health and disease, the accurate assessment of many environmental exposures remains underdeveloped compared to genetic factors [2]. Christopher Wild proposed the concept of the exposome in 2005 as a complement of the genome, defining it as “all life-course exposures (including risk factors), from the prenatal period onwards” [3]. This concept was extended in 2014 by Miller and Jones to: “the cumulative measure of environmental influences and associated biological responses throughout the lifespan, including exposures from the environment, diet, behavior, and endogenous processes” [4]. Recent collaborative discussions in 2024 and 2025 resulted in further revised definitions of the exposome and its research goals, to foster a common understanding in the field [5–7].

Figure 1 shows the three different research domains that have been described within the exposome [8]. The *internal exposome* comprises the internal biological processes as a result of an exposure such as oxidative stress, metabolism, and microbiome changes [8–10]. The internal chemical exposome also includes environmentally derived chemicals plus their transformation products found in cells, tissues, organs or organisms [2]. The *general external exposome* includes social, economic, and environmental factors, while the *specific external exposome* encompasses the individual’s immediate local environment (such as diet, alcohol, infectious agents, pollutants) [8–10].

Unlike the genome, which remains relatively stable over time, the exposome varies on different timescales, requiring complex study designs [2]. The integration of different “omics” layers, known as multi-omics analysis, can offer a more comprehensive understanding of a biological system’s state [11, 12]. Among these “omics” (see **Figure 2A**), metabolomics and exposomics emerge as the terminal downstream outcomes of the genome, reflecting the phenotype of a cell, tissue or organism, in response to diverse genetic or environmental influences over life [13, 14]. Metabolomics and exposomics aim to investigate small molecules, typically ranging between 50 and 1,200 Da [2], although this boundary can also be constrained by analytical reality such as instrument range of *e.g.* m/z = 1000 or 2000 [15]. Given the shared analytical scope, metabolomics and exposomics can be conceptually grouped as “small molecule omics.” Some of these small molecules are closely related with the genome and proteome, leading some to call them the “canaries of the genome” [16]. This metaphor underscores that even minor alterations, such as a single base change in a gene (especially in a metabolic enzyme), can potentially lead to a 10,000 fold-change in concentrations of specific small molecules [16].

The disparity in complexity in the different “omics” layers is shown in **Figure 2B**, which helps explain why genomics and transcriptomics are the most mature omics, followed by proteomics, then small molecule omics [17, 18]. Current automated high-throughput techniques are able to (almost) completely cover the genome and proteome [17]. Gene sequencing requires a DNA sequencer, while protein characterization can be performed on a single type of high-resolution mass spectrometer [16]. In contrast, a wide range of analytical instrumentation is required to capture small molecules, explored further in section 2.4 below.

Since the border between the metabolome and exposome is not always clear-cut [2, 19, 20], a distinction can be made between environmental exposure and the resulting biological response [2] (see **Figure 3**). The metabolome refers to the complete set of small molecules (i.e., metabolites) present within a biological sample such as a cell, tissue, organ or organism at a given time [17]. As these metabolites arise from both endogenous and exogenous sources, the metabolome is considered a key measure for exposome research [19]. In this context, the metabolome can be considered as a subset of the exposome, often referred to as the internal chemical exposome [20]. Thus, while all metabolites found in a biological sample can be considered part of the exposome, not all chemicals within the exposome belong to the metabolome. In contrast, the exposome is a broader concept that encompasses not only chemicals but also all physical, biological, and psychosocial influences that may impact health, as described above and illustrated in **Figure 1.**

This review will specifically delve into common workflows for measuring the *chemical exposome*, with a particular focus on non-target high-resolution mass spectrometry (HRMS) coupled to liquid chromatography (LC). Although standardized workflows remain an area of active research (discussed further in [2, 20, 21]), they can be divided into three main components (**Figure 4**), which have be used to construct the subsequent sections of this article: *(1) experimental workflow*, which encompasses experimental design, sample collection, sample preparation and data acquisition; *(2) computational workflow*, which includes data pre-processing and compound annotation; and *(3) statistical analysis* and *biological interpretation*.

A recent and extensive exposomics review was published in 2024 by Lai et al. [21], however the present review has a different focus. Specifically, this manuscript emphasizes the conceptual, and often blurred, distinctions between the metabolome and the exposome, as well as the analytical and computational workflows employed in non-target HRMS-based exposomics studies. These workflows are illustrated with visualizations to facilitate understanding, and the manuscript also discusses key challenges and strategies to advance towards harmonized studies. This review and the references provided are intended for readers entering the exposomics field, such as PhD students and early-career researchers but will be also helpful to any researcher investigating the human exposome.

2. Experimental workflows

2.1. Experimental design

Individuals may encounter millions of chemical exposures throughout their lifetime, where some exposures may exert lifelong consequences, while childhood exposures may increase the risk of developing disease later in life [19]. The design of prospective longitudinal studies (**Figure 5**), which collect samples at various life stages including perinatal, childhood, and adulthood, holds inherent advantages over those that gather single samples from individuals who have already developed the disease [19, 22]. Although challenging, promising initiatives are currently underway to provide such longitudinal data, such as All of Us [23] and HELIX [24]. While all exposomics studies should undergo independent validation to ensure the reproducibility, this poses significant challenges due to variations in exposures across populations [22]. To ensure robust scientific conclusions, several key factors must be considered, such as the number of samples, the types of samples (matrix selection) and their storage conditions [21].

The number of samples required for an small molecule omics experiment depends on the biological variability of the studied system and the analytical variability of the technology employed [25]. Generally, hundreds or thousands of subjects need to be investigated to obtain robust results, although sample numbers of 3-20 per group can be suitable for generating preliminary and/or pilot data [25]. Such preliminary data can then be used to estimate the number of samples needed to achieve certain statistical power (e.g., 0.8), using open tools such as G*Power [26], MetaboAnalyst [27] and MultiPower [28], which estimates sample sizes for multi-omics studies.

2.2. Sample collection

The matrix selection is a critical aspect of the study design, affected by multiple factors including the cost of storage and collection devices, research question, technical feasibility and expected presence of specific chemicals, as some chemicals accumulate specifically in some tissues [21, 29]. To date, blood (plasma or serum) and urine are the most studied matrices [21, 22]. While biological samples can provide insights into potential biological responses (i.e., endogenous compounds) to toxicants or pollutants (i.e., exogenous compounds), these exogenous compounds are typically present at trace levels compared to the endogenous ones and thus stretch the dynamic range of current instrumentation [30], as detailed below. Consequently, it may be advantageous to also incorporate non-target screening of environmental samples such as dust to first build a picture of potential contaminants, before analysing biological samples.

The collection and storage of many biological samples (at -80°C) for long-term longitudinal studies (**Figure 5**) can be costly. Dried Blood Spot (DBS) samples are compact, easy to collect and can be shipped at ambient temperature. However, the analysis of DBS is not as standardized as for plasma or serum, and it is not yet clear if there is sufficient sample material for non-target analysis [31]. Urine is a well-studied matrix, easily accessible and less invasive than blood [22], while fecal analysis can yield valuable insights into gut microbiota metabolites, with relatively easy sample collection. Although cerebrospinal fluid (CSF) provides valuable information on brain metabolism, the collection of CSF samples is invasive, and the recruitment of healthy volunteers is more difficult, such that studies with large sample sizes involve significant commitments and dedication. The concentrations of exogenous chemicals in this matrix are typically very low compared with other biological samples such as blood. Other sample types that can be collected to answer specific questions include hair, teeth and nails, which can help to understand historical exposures, although determining the timing of exposure can be laborious [2]. The study of the specific external exposome can be done by collecting household dust samples with a vacuum cleaner [32], although these samples are prone to variability. Passive samplers such as silicone wristbands allow a more individual assessment of the exposure to chemicals, although the silicone used in wristbands can lead to analytical interferences if they are not properly cleaned before use and the sample preparation may be more tedious compared to household dust [33, 34].

Depending on the aim, two different approaches can be used when analysing the chemical exposome: (1) *target or targeted* studies, which focus on identifying a limited number of specific chemicals (hypothesis driven) and (2) *non-target or untargeted* studies, which are hypothesis generating and aim to identify as many compounds as possible [21, 35]. Typically, non-target studies are *semiquantitative*, which use peak height or area as direct readouts or estimate the concentrations relative to quantifiable compounds such as internal standards spiked at known concentrations. In contrast, target studies frequently employ *absolute quantitation*, which involves calculating the exact concentration via calibration curves [21]. The sections below focus on non-target HRMS studies, which can be used to generate hypotheses and insights for designing subsequent quantitative targeted studies.

2.3. Sample preparation

Sample preparation (or pretreatment) prior to instrumental analysis aims to reduce interferences, separate and concentrate analytes [36]. Due to the chemical diversity of the exposome, there is no universal method that captures all chemicals present in a sample [25] and a combination of different sample preparation approaches can cover a broader range of the chemical space [21]. However, this is limited by increased costs in resources and time.

Sample preparation methods for non-target analysis of biological/environmental samples should be [37]: (1) unselective (to cover a wide range of chemicals); (2) simple and fast (to prevent chemical loss/degradation during preparation), (3) reproducible, and, for biological samples, (4) incorporate a quenching step. The rapid stopping or quenching of metabolism is an essential step to produce a stable extract that reflects the endogenous metabolite levels present in the original biological system [37, 38]. The sample preparation method is highly dependent on the matrix (e.g., cells, plasma, or tissues), and analytical platform employed. For instance, although protein precipitation is the first step for plasma samples, tissues must be homogenized first [39]. Furthermore, while Gas Chromatography coupled with MS (GC-MS) often requires a derivatization step to make the compounds sufficiently volatile for the analysis [40], the sample preparation for Liquid Chromatography coupled with MS (LC-MS) analysis is simpler and typically does not require this step [39].

Liquid-liquid extraction (LLE), often including a protein precipitation step, and dilute and shoot (DNS) are frequently employed sample pretreatment methods in exposomics studies. Other sample preparation methods include solid-phase extraction (SPE) and dispersive solid-phase extraction, such as QuEChERS [21, 36]. While LLE and DNS offer broad analytical coverage, they are susceptible to matrix effects and interferences, which can limit reproducibility and sensitivity. In contrast, SPE and QuEChERS typically enhance sensitivity and reproducibility, albeit often at the cost of reduced chemical coverage [36]. Further information can be found in the NORMAN guidance for non-target screening [15] and other publications on sample preparation for metabolomics [25, 37, 39] and exposomics [21, 36].

2.4. Data acquisition

While HRMS has emerged as the leading technique to investigate the chemical exposome, there is no “one size fits all” analytical method and a combination of different separation and ionization platforms is needed to capture the relevant chemical space (**Figure 6**) [35]. Flow-injection is very fast, but cannot separate any isobars, while chromatographic separation introduces analytes slowly into the mass spectrometer, separating more isobars and reducing the risk of ion suppression, source fouling and coelution. A mobile phase, either liquid or gas, transports the analytes through a stationary phase-fixed system [2, 2, 21] (see top part of **Figure 6**).

GC is frequently used for the analysis of volatile and thermally stable compounds such as fatty acids, and organic compounds (see left part of **Figure 6** for examples). Due to the high temperatures, a derivatization step is often required before the analysis, which may result in compound loss [39, 41, 42] and changes the resulting mass spectra. The most common ionization technique applied in GC is Electron Ionization (EI), which provides robust, and highly reproducible fragmentation patterns [39].

Currently, LC coupled with an electrospray ionisation (ESI) source is probably the most used HRMS-based platform for non-target studies due to the soft ionization process, high dynamic range and versatility [2]. LC-HRMS does not typically require derivatization and is highly applicable to the analysis of a broad range of medium to very polar compounds

[39, 41], with several examples shown in the right part of **Figure 6**. Using both positive (+) and negative (-) ionization modes increases the coverage. Reversed Phase (RP) columns are widely employed to separate polar and medium polar compounds, providing relatively reliable, robust and reproducible results. Hydrophilic Interaction Liquid Chromatography (HILIC) columns improve the separation of very polar compounds minimally retained by RP [21, 39, 42], but can be less reproducible. Alternative ionisation techniques such as Atmospheric Pressure Photoionization (APPI) and Atmospheric Pressure Chemical Ionization (APCI) can extend the LC range almost into the GC range (**Figure 6**).

Despite recent progress, current analytical platforms still lack the dynamic range to simultaneously detect trace-level exogenous compounds and high-abundance endogenous metabolites (see **Figure 7** for an example based on LC-ESI-HRMS). The integration of Ion Mobility Spectrometry (IMS) into HRMS workflows has been gaining more attention as it can improve the dynamic range and throughput of the analysis [2, 18]. Coupling IMS with HRMS offers the ability to resolve isomers or isobars that are difficult to distinguish using only HRMS, and adds Collison Cross Section (CCS) information to the retention time, MS1 and MS2 data, providing complementary information for compound identification and/or annotation [2, 35]. Depending on the instrument, including IMS also reduces the complexity in the MS1 and MS2 spectra [21], but comes at a cost of sensitivity and more difficult data analysis [2, 43]. Typical commercial IMS instruments do not yet offer sufficient resolution to resolve isomers within measurement and prediction errors [44, 45].

Quadrupole Time-of-Flight (Q-TOF) and Orbitrap are the most commonly used mass analyzers in non-target studies, as they provide high resolution [2, 39, 42, 46]. A Q-TOF analyzer maintains the selectivity of the quadrupole and provides a mass resolution of approximately 40,000 - 60,000. Orbitrap analyzers reach resolutions between 250,000 to even 1,000,000 for ions with *m/z* below 300 [46]. Triple Quadrupole (QQQ) analyzers are often used for targeted analysis to confirm potential biomarkers due to high sensitivity and selectivity [41, 42], but are not suitable for non-target studies due to their low resolution [39].

Two acquisition methods are often used in non-target HRMS studies: Data-Dependent Acquisition (DDA), and Data-Independent Acquisition (DIA). DDA is typically more common, since MS2 spectra can be associated with a specific precursor [2, 47, 48], resulting in simpler processing compared with DIA [47] [49]. While low intensity features may not be selected for fragmentation, this can be alleviated with the creation of inclusion lists, supported by either open software solutions such as iterative exclusion lists from IE-Omics [50], or instrument software such as AcquireX from ThermoFisher. DIA mode operates in a less-selective manner, as the instrument selects all the peaks within the isolation window, regardless of the peak intensity [2, 47]. Thus, the MS2 spectra are information-rich collections including fragments from low intensity ions, but they lack the precursor-fragment information [47] and require deconvolution to link the specific precursor to the MS2 spectra [2, 48]. This can be performed by open software such as MS-DIAL [51]. All-Ion Fragmentation (AIF), and Sequential Window Acquisition of all Theoretical Fragmention Spectra (SWATH) are two commonly employed DIA methods.

In practice, method selection depends strongly on the research question, instrument availability and analytical expertise, among others. While LC-MS (e.g, QQQ) remains the preferred option for the quantification of a specific and small number of compounds (target studies), LC-HRMS and GC-HRMS (e.g., Orbitrap) are employed for non-target exposomics studies generating larger and more complex datasets that make the data preprocessing steps more challenging. Therefore, the analytical considerations detailed above, together with the references provided, can serve as practical guidance for selecting the most appropriate analytical platform and acquisition mode for a given a study.

3. Computational workflows

3.1. Non-target HRMS Data pre-processing

The main objective of data pre-processing (or feature detection) is to transform the raw data files into a format that simplifies the access to the distinct characteristics of every observed feature in each sample analyzed, i.e., a feature list [52]. A “feature” does not necessarily refer to a specific chemical, but rather it typically refers to a peak (or signal) identified at a specific retention time, and *m/z,* containing spectral information such as MS1 and/or MS2 [35]. Features are also called “*m/z* features”, “ion features” or “ion peaks” [21]. The resulting feature list (**Figure 8**, bottom right) can be then employed for various purposes, including feature prioritization, compound annotation, and statistics and typically contain the retention time, *m/z*, and peak area and/or intensity of each feature from each raw data file [21, 52–54].

Data pre-processing steps include data conversion (if vendor software is not used), centroiding, filtering step for noise removal, generation of Extracted Ion Chromatograms (EICs), peak picking, peak grouping across samples, and retention time alignment (**Figure 8**) [2, 54]. Software approaches include vendor software such as Compound Discoverer (ThermoFisher), MassHunter Profinder (Agilent), MetaboScape (Bruker) and Progenesis QI (Waters), as well as open software including MS-DIAL [51], MZmine [55], OpenMS [56], XCMS [57], and patRoon [58]. Recent reviews provide a comprehensive overview of the start of the art of data pre-processing software [15, 20, 59, 60]. Since different instrument vendors use different formats, conversion of the raw data files into an open format such as mzML or mzXML [53] is required for open approaches, typically using ProteoWizard [61, 62], although some software now embed this conversion, including MS-DIAL [51] and patRoon [58]. Data acquisition can be done in either profile or centroid mode; the centroiding of profile data is an important data reduction step (see top part of **Figure 8**) [15, 54], often performed together with the data conversion using ProteoWizard [61]. While ProteoWizard offers both vendor-specific or general algorithms, the use of vendor-specific algorithms is recommended [15] The quality of the conversion using ProteoWizard hinges on the vendor type - while the ThermoFisher conversion yields high quality results, this is not the case for Waters [63].

Once data is centroided, a filtering step is applied to suppress or reduce the random analytical noise, which is always present in the acquired MS data [53, 65, 66]. Noise is often removed via smoothing [64, 65]. Methods include linear weighted moving average [51], Savitzky-Golay smoothing [67], moving average [67], and binomial filter [68]. MS-DIAL uses the linear weighted moving average by default, although it supports all the aforementioned approaches [51]. After the EIC generation, peak picking is performed to determine the area under the peak [52, 53, 64]. This requires the establishment of criteria (i.e., parameters) to distinguish true peaks from noise [65] such as setting a minimum intensity threshold and an estimated chromatographic peak width, establishing a maximum *m/z* error (e.g., 0.001 Da for a Q-TOF instrument and 5 ppm for an Orbitrap), and ensuring that identified peaks are present in a significant proportion of the samples [54, 65, 66]. Next, the selected peaks are grouped across samples [54], and an alignment step is needed to correct retention time differences between runs (samples) [53]. The alignment algorithm often requires a reference sample (e.g., pooled Quality Control (QC) sample) to correct the retention time differences, and this choice can have a significant impact on the results [53, 66].

The selection of parameters for data preprocessing is a critical step in the exposomics analysis [54], as inadequate parameter selection can lead to biased results. Since exposomics seeks to identify highly abundant endogenous compounds along with exogenous small molecules that are often present at trace levels (see **Figure 7**), there are no universal parameters that can effectively capture the chemical complexity of the human exposome currently [21]. While approaches such as IPO [69] can be used to assist in parameter selection, they should be employed with a degree of caution, as they tend to discard low abundant or rare peaks, which may be trace level exogenous molecules of relevance for the exposomics question. Recently proposed quality assurance/quality control (QA/QC) guidelines support the exposomics community in their data preprocessing choices [54], see further details in Section 5.

3.2. Compound Annotation

The annotation of small molecules remains a major challenge in non-target HRMS based metabolomics and exposomics studies, as the majority of features detected (around 80%) remain unknown [70], although experts continue to debate the fraction. High abundant peaks, with MS2 available, usually represent <10% of the detected features by non-target

HRMS, while the majority (60-70%) are low abundant peaks without fragmentation information [71]. *Annotation* involves associating an identified MS feature with a specific chemical identity, while *identification* is the process of verifying that the annotated compound corresponds to the proposed chemical (i.e., confirming the annotation with the reference standard) [72]. However, the lack of availability of chemical reference standards for given molecules of interest represents a major bottleneck in exposomics/metabolomics studies, complicating biological interpretations [73].

The large number of unannotated features in non-target studies is a common subject of debate. Recently, Giera et al. [74] suggested that most (>70 %) of the unannotated features in non-target LC-HRMS experiments are in-source fragments (ISFs), concluding that the dark metabolome may be smaller than previously thought. However, the results were based on a 931K compound library that is not publicly disclosed, while the formation and recognition of ISFs is highly dependent on multiple parameters including source voltages, instrument design, matrix type, extraction methods, analyte concentration and data preprocessing workflow [75]. Moreover, many of the ISFs observed in highly concentrated synthetic chemical standard mixes may not be detected in complex biological samples. Previous studies, in contrast, have reported that ISFs entail ~2-25% of all detected ions [75–77]. While ISFs can be problematic, they can also be intentionally enhanced to improve annotation confidence [78] and modern workflows are increasingly accounting for them [75]. These different estimates highlight the current lack of consensus and harmonization in the field. While high prevalence of ISFs would imply that the dark metabolome is smaller than previously thought, assuming that most of the features identified are known, lower estimates would indicate that although ISFs remain relevant, they may represent a relatively small subset of detected features.

Non-target HRMS studies generate vast amounts of data, necessitating various computational strategies such as suspect screening and non-target screening (top part of **Figure 9A**). While suspect screening approaches use lists of chemicals that could potentially be present in the samples (i.e., suspect lists of certain chemical classes or organisms/matrices) for more efficient discovery, non-target screening approaches aim to identify as many compounds as possible via tandem mass spectral libraries or database search (e.g., via in silico fragmentation software), as shown in the bottom part of **Figure 9A**. A detailed glossary of terms can be found in the 2023 NORMAN guidance [15].

Compound annotation is a fundamental step to convert raw HRMS data into meaningful biological information [35, 71]. Since the amount of information available for identification varies, it is essential that the confidence assignment of each feature is transparent [35]. Several confidence level schemes exist, including the 2007 Metabolomics Standards Initiative (MSI) [83], the 2014 guidelines for HRMS data with an environmental focus [81] (left part of **Figure 9B**), 2020 guidelines for IMS [84] and 2022 guidelines for PFAS [85] and GC-HRMS [86]. The 2014 levels shown in **Figure 9B** range from Level 1 (confirmed structure with reference standard), to Level 5 (only a *m/z* is known) [87]. According to metabolomics terminology conventions, only Level 1 can be considered identifications, while the rest (Level 2-5) are annotations. In downstream analyses, it is recommended to base biological interpretations primarily on Level 1 identifications, however, this is often hampered by the limited availability of chemical standards. Consequently, Level 2 annotations are frequently used as the next most reliable basis for interpretation, whereas Level 3-5 should be interpreted with caution, serving mainly for prioritization for future validation efforts.

Several criteria are used to assign the identification level of each feature, including the MS1, retention time, fragmentation pattern (MS2), CCS (if available) and experimental data [71]. However, although different classification systems have been proposed [38, 81, 83, 88, 89], currently there is no standardized system for compound annotation integrated in the processing workflows, making the comparison of results between studies challenging [71]. This necessitates a degree of “translation” between software outputs (see *e.g.,* **Figure 9B**, right). While the use of False Discovery Rates (FDR), as done in other omics fields, has been proposed, estimating total FDR for compound identification in small molecule omics is still a nascent challenge in the field [90, 91]. The use of identification probability was proposed recently as an alternative to the identification levels [91]. However, the probability depends on the reference library size and treats all candidates equally, such that smaller reference libraries can artificially lead to high identification probabilities, while larger libraries (such as those more applicable to exposomics) will potentially yield too many apparently equally-valid candidates with very low probabilities to support meaningful outcomes. Identification levels can be assigned automatically in some patRoon workflows [58, 92], while the NORMAN Network also trialled an automated assignment system [93] and Boatman et al recently published a checklist to facilitate automation for PFAS identification with IMS [94].

3.2.1. Compound Annotation Via Tandem Mass Spectral Libraries Search

The fastest and most accurate (and thus most common) strategy for compound annotation is to compare the experimental mass spectra (MS2), with standard mass spectral libraries [35, 47, 70]. Compound annotation via spectral library searching is based on the premise that molecules generate a reproducible “fingerprint” under specific fragmentation conditions [72, 95] (see **Figure 10**). Good matches between the experimental and the library MS2 spectra can lead to Level 2a annotations [72]. If the MS2 library is created *in-house*, i.e., the experimental and library spectra are acquired under the same conditions, this can lead to Level 1 annotations when the retention times also match. Commonly used spectral libraries include MassBank [96], MassBank of North America (MoNA) [97], GNPS [98], mzCloud [99], METLIN [100], and NIST [101]. While GC-EI mass spectra have been standardized for over 60 years, LC-MS spectra are less standardized due to instrument variability and differences in acquisition parameters such as collision energy. As a result, MS2 libraries often contain multiple entries for each compound [72, 91].

Tandem mass spectral libraries are typically generated by the analysis of chemical reference standards [72]. Therefore, compound annotation by this approach is hampered by the limited availability of mass spectra data due to lack of standards, and/or lack of open data. Of the 122 million compounds in PubChem (September 2025), only 36,242, or 0.03 % have open LC-MS/MS data available [102]. Nonetheless, there has been substantial progress in the quality and quantity of mass spectral libraries in recent years. Automated spectral library searching and matching can be performed using various open and commercial software with different algorithms. Typically, these software tools work in a two-step procedure: (1) MS1 filter, (e.g., 0.01 Da) which can remove up to 99% of false candidates and speeds up searching, and (2) similarity algorithm, which ranks the experimental MS2 spectra against the remaining library spectra and calculates a similarity score, see bottom part of **Figure 9** [47]. Ideally, scores should be able to distinguish true and false positive matches [72].

The most common spectral match score is the cosine score, which converts two MS2 spectra (observed and reference) into two equally size vectors through mass peak binning, and then calculates the dot product which ranges from 0 to 999 (or 0 to 0.999) [47, 73]. A score of 999 indicates a perfect match between the two spectra, while a score of 0 indicates no match [47]. Newer scoring approaches include the Entropy score [103] and a range of new machine learning algorithms [73]. The right part of **Figure 9B** shows how these scores can be applied to annotate compounds using different software. Several open software approaches support spectral library search, including MS-DIAL [104], MZmine [55], openMS [56], and XCMS [57], while commercial examples include Progenesis QI (Waters) and MetaboScape (Bruker). The in silico fragmentation software MetFrag also integrates a library search and calculates an Exact Spectral Similarity score (also known as “MoNA score”) [105]. However, the variability in output results, including similarity scores, across different software can make the identification levels obtained poorly comparable. For instance, MS-DIAL and patRoon provide similarity scores, dot product and MoNA scores, respectively, with a proposed minimum data requirements for annotation shown in **Figure 9B**. Importantly, even when the spectral similarity score is high (>0.9), the identity of the compound must be confirmed with a chemical reference standard to classify it as Level 1 [21]. Otherwise, the match should be considered an annotation rather than an identification (Level 2 or below), since several isomers can have very similar spectra such that only retention time or other orthogonal parameters can distinguish between them.

3.2.2. In silico Approaches for Compound Annotation

In silico fragmentation software supports compound annotation of candidates beyond those in mass spectral databases. These methods typically involve matching the experimental spectra against a selection of candidates obtained from known compound databases (discussed in the next section). Approaches such as MetFrag [105], Mass Frontier, MS-FINDER [106] and CFM-ID [107] fragment the candidates and match the resulting spectra with the experimental spectrum, while approaches such as CSI:FingerID [108] and SIRIUS use the experimental spectrum to generate fingerprints, which are matched with the fingerprints of the candidates to rank the structure candidates [70, 109]. MetFrag uses a bond dissociation approach to generate fragments for each candidate, which are compared with the experimental spectra to determine which are the best candidates [105]. Mass Frontier (Thermo) uses rule-based fragmentation prediction, complementary to the bond disconnection approach. CFM-ID [107] is a machine learning-based approach that can predict fragments and intensities, and thus can be used to generate in silico libraries of the given spectrum type used during the training [47].

In silico spectral libraries can help to overcome the limited number spectra in MS2 libraries and avoid the need for “on the fly” calculations in each workflow. In silico spectral libraries can be generated via *e.g.*, quantum chemistry, machine learning, heuristic-based, and chemical reaction-based methods. Heuristic approaches are best applied to compounds with consistent fragmentation patterns such as lipids [47]. LipidBlast [110], integrated within MS-DIAL [104] and LipidMatch [82], is an in silico library containing more than 200,000 spectra generated using a heuristic approach. LipidMatch [82] is a rule-based software that incorporates various libraries to facilitate the lipid annotation. CFM-ID [107] has been used to *e.g.* generate in silico EI-MS and MS/MS spectra of small molecules in HMDB [111]. In general, predictions should only be used for compounds within or close to the domain of the training set / rule sets used.

In silico approaches for compound annotation typically yield Level 3 annotations or below, but they can be upgraded to Level 2 with the support of a good tandem mass spectral library match, such as the combined approach with MoNA used in MetFrag. In silico annotations often serve an important early role in the elucidation process, guiding subsequent activities such as the interpretation, prioritization and even acquisition of reference standards [73].

3.2.3. Compound Databases for Compound Annotation

Due to the extreme chemical diversity of the chemical exposome, the database selection plays a critical role in the annotation process. This aims to reduce both false positives (i.e., incorrect annotations) and false negatives (absence of the correct structure in the database). While the Chemical Abstract Service (CAS) [112] database is the largest chemical registry containing over 290 million organic substances (September 2025) [15], it is not freely available or compatible with open software approaches. ChemSpider [113] and PubChem [114] contain over 128 and 122 million chemicals respectively (September 2025), making them the two largest freely available chemical databases. However, due to user quota limitations on ChemSpider, PubChem currently emerges as the most feasible large chemical database for integration into open software workflows [15].

The use of smaller subsets of chemicals helps in the annotation process as these contain known molecules specific to certain domains. For example, PubChemLite for Exposomics (PCL), a subset of PubChem containing 442,379 chemicals (version 2.0.0) [87, 115, 116] and the Blood Exposome Database [117] (67,291 compounds) aid exposomics researchers in identifying relevant chemicals. Metabolite discovery in humans is facilitated by the HMDB [111], while KEGG [118] and MetaCyc [119] establish connections with proteomics and transcriptomics disciplines [70]. Lipidomics studies are supported by LIPID MAPS Structure Database (LMSD) [120], which contains 49,790 unique lipid structures (July 2025), which is thus the largest public lipid-specific database. The Human Microbial Metabolome Database (MiMeDB) facilitates the study of small molecules produced by the human microbiome [121], while the CompTox Chemicals Dashboard is a collection of 1,254,895 chemicals relevant for computational toxicity efforts [122]. The coverage of HMDB, PCL, CompTox and the Blood Exposome database is explored in **Figure 11A,** which shows the small, polar focus of the Blood Exposome database vs the wider range of chemicals in HMDB (including lipid-based molecules) and the even wider range in CompTox and PCL that may not be detectable in humans, but may still influence health outcomes.

As discussed above (see section **2.4** and **Figure 6**), a combination of analytical techniques is necessary to capture the chemical exposome. **Figure 11B** displays the chemicals within PCL that contain LC-MS and GC-MS spectral information, as well as CCS records in PubChem. This highlights the complementarity of the techniques, while GC-MS effectively identifies part of the non-polar side of the exposome, LC-MS covers a wider range of polar compounds. Notably, there are still many areas of the exposome that have no spectral information available, including very lipophilic molecules (middle and right top of the plot) and highly polar compounds (bottom right of the plot). Interestingly, CCS values are available for a good number of small molecules, providing an additional parameter to support compound annotation and identification in non-target HRMS. Several classes have been highlighted in **Figure 11B**, showing that both endogenous (triglycerides and nucleotides, in blue) as well as exogenous (perfluorooctanesulfonate (PFOS) and pesticides, in grey) can be captured with both LC-MS and GC-MS.

Figure 11C shows the overlaid plot of PCL along with four of the major PCL categories. Interestingly, the chemical space covered by the Biopathway category (**Figure 11C**) and HMDB (**Figure 11A**) are very similar, although both contain patches with little experimental data in PubChem (mass ~750-1750, XlogP 20-30). The Disorder and Disease category also overlaps more with the DrugMedicInfo category than perhaps expected, indicating that a large portion of this section may indicate drug availability and not additional disease insights. The Agrochemical information and DrugMedicInfo categories (**Figure 11C**) cluster in the left part of the plot, representing low molecular weight molecules with medium polarity, similar to the patterns observed in the Blood Exposome Database (**Figure 11A**).

One important fraction of the chemical exposome are transformation products, which are often overlooked [124]. These products, derived from biotic or abiotic reactions, can have toxicity and persistence profiles that differ significantly from their parent compounds and may even be more toxic in some cases [125]. Currently, structural elucidation of transformation products is typically performed by a combination of experimental and in silico approaches [124]. The “Transformations” section in PubChem, which contains documented parent-transformation product relationships from ChEMBL and the NORMAN Suspect List Exchange (NORMAN-SLE) [126] is included in PubChemLite [87], while software such as patRoon can aid in identifying transformation products from databases or by in silico prediction [127]. Open source software such as BioTransformer [128] or ShinyTPs [125] help generate transformation product databases or suspect lists via in silico prediction and text mining, respectively.

Suspect screening approaches can be used to search for compounds of interest that are expected (“suspected”) to be in the samples, facilitated through the use of suspect lists. This can be considered a form of prioritization, as it reduces the number of compounds to be investigated and assists in discovering potentially relevant results. Although this approach was initially employed in environmental and toxicology sciences to expedite non-targe screening, it has been increasingly applied in metabolomics and exposomics studies [2, 15, 21]. Different platforms exist to exchange suspect lists, such as the NORMAN-SLE [126] and the CompTox Chemicals Dashboard [129]. Furthermore, specific lists of chemicals can be generated in PubChem by literature mining [15]. Suspect screening can be facilitated by software such as patRoon [58], which performs the automatic annotation of “suspects” based on pre-defined rules. The automatic assignment of identification levels is a key feature of patRoon, enhancing the reproducibility and leading to more transparent and comparable results [58, 130]. Other software options allowing suspect screening include MZmine [55], and Compound Discoverer (ThermoFisher). However, there is a risk in focusing too narrowly on distinct suspect classes in exposomics, depending on the study context, since recent studies have shown that significant chemicals for disease penetrance arise from many different information categories and suspect lists [32].

4. Statistics & Biological Interpretation

Non-target HRMS exposomics studies generate a large amount of data, necessitating a combination of univariate and multivariate statistical analysis to identify significant differences between groups [131]. Data pre-treatment (e.g., normalization, scaling) is essential before applying statistics, as detailed in the next subsection.

4.1. Data Pre-treatment

Data pre-treatment consists of transforming the HRMS feature list into a suitable state for subsequent statistical analysis [83]. This process aims to reduce the effects of technical and measurement errors while enhancing relevant biological variations [132]. Common pre-treatment methods include normalization, centering, scaling (e.g., Pareto scaling), and transformations (e.g., log and power) [52, 133]. Multifunctional open tools such as MetaboAnalyst [27], XCMS online [57], and MS-DIAL [51] implement some pre-treatment steps, while various statistical packages are available for data pre-treatment in C/C++, Java, R, and Python [134].

Normalization strategies are used to remove or correct unwanted systematic variations between samples, making them more comparable [53, 83, 132, 135, 136]. Normalization in metabolomics and exposomics is more challenging compared to genomics and proteomics, due to the vast complexity of the chemical space [136]. For instance, there is no standard method to measure the total amount of chemicals in a sample to normalize in the way total protein amount is used in proteomics [64, 135, 136]. Normalization can be performed either pre-acquisition or post-acquisition of HRMS data and is generally divided into sample-based and data-based approaches, reviewed recently elsewhere [135].

4.2. Univariate and Multivariate Statistical Analysis

Univariate statistical tests aim to identify changes in individual molecules and work on the assumption of statistical independence [131, 137, 138]. These approaches are commonly used to initially assess the potential relationships between exposures and disease phenotypes [21]. Different tests are available to investigate differences across groups (or changes over time) and the choice depends on the data distribution and the experimental design (number of groups and type, i.e., matched or unmatched) [64, 139]. Univariate analysis can also be employed to investigate the association between specific small molecules of interest and other variables, such as known clinical parameters, environmental exposures, or microbial species, among others. For this purpose, various similarity tests, including Pearson's correlation (parametric test) and Spearman's correlation (non-parametric test) can be used [140]. Univariate statistics are also widely used within Exposome-Wide Association Studies (ExWAS)[21]. In an ExWAS, a large number of exposures are successively and independently tested for their association with a specific health outcome, using statistical approach analogous to Genome-wide association studies (GWAS)[141]. However, although univariate statistics are widely employed to test individual chemicals, non-target exposomics studies do not generate univariate data, as chemicals are not independent of each other, necessitating the use of multivariate statistics [137].

Multivariate statistical methods encompass both supervised and unsupervised methods [83]. Unsupervised methods are generally exploratory in nature, used to find patterns and generate hypotheses, while supervised methods are more confirmatory (hypothesis testing), widely used for biomarker identification, classification, and prediction [134]. Unsupervised methods are an effective approach to explore and visualize the structure of the dataset. Principal Component Analysis (PCA) [64] stands out as one of the most widely employed methods, used to reduce the number of dimensions in the data, facilitating data exploration and visualization. It is often employed as a pre-processing step, to check the data quality, before applying a supervised method [52, 142]. Pooled QC samples that cluster tightly, ideally at the origin of the scores plot, are indicative of high-quality data [142] (**Figure 12A**).

Supervised methods are used to identify the independent variables (molecules) that best discriminate the groups under study (dependent variables). PLS-DA and orthogonal-PLS-DA (oPLS-DA) are some of the most commonly used methods. As shown in **Figure 12B**, they maximize the differences between groups. However, the main drawback of these supervised methods is their susceptibility to overfitting. Therefore, it is highly recommended to perform validation analysis to avoid finding false relationships and misinterpretation of the data [131]. Ideally, these classification models require splitting the dataset into training set, validation set, and test set. Thus, individual models are tested and evaluated on unique datasets and then applied to the entire dataset and/or to other datasets [131, 137]. This approach, however, is often limited when working with small sample sizes or cases such as rare diseases or mutations, which may prevent the dataset from being split effectively.

4.3. Special Statistical Considerations for Exposomics Studies

Missing data pose a particular challenge in exposomics studies, where exposures are often analysed jointly. As the number of included exposures (such as chemicals identified by HRMS) increases, the number of complete cases can decline rapidly. This issue is further exacerbated in longitudinal exposomics studies where participant numbers are typically highest at baseline, but dropouts accumulate over time, leading to progressively fewer observations and a greater *proportion* of missing data at later time points. Therefore, the use of imputation techniques, such as multiple imputation is recommended, as detailed by Santos et al. [141].

Exposomics studies often involve analyzing samples collected at different time points and therefore processed in different analytical batches. This can introduce batch effects, which occur when quantitative measurements differ systematically between batches due to irrelevant factors [143]. Batch effects can be caused by multiple sources at any step of the experimental workflow (**Figure 4**), from sample collection to sample preparation and data acquisition by the analytical platform (e.g., LC-HRMS), which is often the primary source of variation [143]. These batch effects are almost unavoidable when working with HRMS platforms. Thus, it is very important to identify the unwanted variations (e.g., via PCA, as shown above in **Figure 12A**) and correct them. Although several strategies have been proposed to correct batch effects (e.g., using ISs or pooled QC samples), it remains an active area of research with no standardized approach [143, 144]. A recent review discusses potential sources of batch effects in multiomics studies as well as different batch effect correction algorithms (BECAs) [145].

Exposomics data typically contains many covariates such as confounders (*e.g.*, age, sex, medication), which must be accounted in the statistical analyses. For this purpose, linear models can be employed to identify potential chemicals associated with a particular outcome (e.g., Parkinson’s disease or diabetes), while accounting for any number of covariates [146].

Network based approaches can help interpret the behaviour of chemicals or diseases that are related, and to provide insights into their mechanisms. Specifically, in exposomics, network approaches allow the identification of correlated exposures and potential biological changes associated with multiple exposures and health effects [141, 146].

Mendelian randomization (MR) has become a valuable method in exposomics for assessing causal relationships between exposures and health outcomes. MR uses genetic variants, such as single nucleotide polymorphisms (SNPs), as instrumental variables to estimate the causal effect of a modifiable exposure on an outcome [146]. MR relies on the assumptions that genetic variants are strongly associated with the exposure, are not associated with confounders, and influence the outcome only through the exposure [21, 146]. Recent studies [147–150] have successfully applied this approach in various exposomics contexts.

Since chemical exposures typically occur in complex mixtures rather than individually, various statistical methods can be applied to account for multiple exposures and their potential combined effects [21, 151]. These include, mixture analysis approaches, dimension reduction techniques or Bayesian model averaging. Further details can be found in Maitre et al. [151]. In addition, machine learning approaches such as random forest, support vector machine, and gradient boosting can be used for biomarker selection, classification, and predictive modelling [21, 146].

4.4. Biological Interpretation

Biological interpretation of exposomics data is still a major challenge [152]. This is primarily due to the fact that the majority of the detected features by non-target HRMS remain unknown or, if annotated, little additional information is available. Another key challenge is distinguishing between exogenous and endogenous chemicals that reflect the biological response to the exposure [20], where it is increasingly plausible that some detected chemicals originate from both exogenous and endogenous sources. Therefore, a multi-faceted approach is essential for interpreting small molecule omics data, encompassing different computational methods for statistical analysis, functional analysis, chemical classification, data integration, and data visualization, among others [137].

Network modelling and pathway mapping are key approaches for providing biological context to the data by enhancing the understanding of relationships between small molecules [153, 154]. MetaboAnalyst [27, 140] is a widely employed platform allowing data processing and interpretation through various functionality, including pathway, enrichment and network analysis. These approaches are valuable for data exploration and hypothesis generation, but the results require further validation. Pathway mapping is often limited by incomplete and manually curated pathway databases, leading to variability in results across different databases (e.g., KEGG and MetaCyc) [137], while various metabolites can belong to different pathways. This analysis also excludes exogenous chemicals (i.e. the chemical exposome). An alternative approach is ChemRICH [152], which uses MeSH and Tanimoto substructure chemical similarity coefficients to cluster small molecules into non-overlapping chemical groups. In contrast with pathways analysis, ChemRICH sets have a self-contained size (based on the chemicals found in a particular study), therefore p-values do not rely on the size of a background database, such as KEGG. Furthermore, the analysis can also place exposome chemicals into metabolite sets. However, results from ChemRICH cannot yet be directly integrated with genomics or proteomics results [137, 152]. Another strategy that does not require compound identification is mummichog [155]. This method uses as input peak lists, which are queried against a database to identify all the potential matches to metabolic pathways and networks [154]. Linear models are useful for complex exposomics studies as they can account for covariates like age, sex, and occupation, helping to identify chemicals associated with specific metadata of interest. A recent review summarizes different computational methods including linear models with covariate adjustment, dimensionality reduction, and neural networks, among others, that support exposomics data analysis and interpretation [146].

Multi-omics approaches offer a more comprehensive understanding of the biological system by integrating small molecule omics with other omics data acquired from the same samples. Currently, there is a wide array of tools available for the integration of multi-omics data, such as mixOmics [156]. For instance, combining metabolomics with metagenomics can help elucidate the role of bacteria derived metabolites [153]. In this context, tools like microbeMASST can help identifying the potential microbial origin of annotated chemicals by mapping known and unknown MS2 spectra to potential microbial producers [157].

5. From History to Harmonization: Addressing Challenges Between the Metabolome and Exposome

Currently, the lack of harmonization not only in the annotation process but also throughout the entire non-target HRMS workflow leads to poorly comparable metabolomics and exposomics studies. As shown in **Figure 13**, metabolomics and exposomics are still young and rapidly growing fields of research, where workflow standardization is an ongoing task. To address this need, several initiatives have emerged separately for each discipline. In 2005, the MSI was formed [158], proposing a series of minimum reporting standards for data analysis in metabolomics two years later [52]. A network of global metabolomics repositories arose in 2015 (COordination of Standards in MetabOlomicS (COSMOS)) [159]. Since exposomics is a newer field, there is no "Exposomics Standard Initiative" yet. However, in the last few years various European, American and international initiatives have been developed, including The Human Early-Life Exposome (HELIX) [24], EXPOsOMICS [160], The European Human Exposome Network (EHEN) [161]), the Network for EXposomics in the United States (NEXUS) [162], HERCULES [163], Human Health Exposure Analysis Resource (HHEAR) network [164], and the International Human Exposome Network (IHEN) [165].

QA/QC procedures should be implemented throughout the entire metabolomics and exposomics workflow, from sample preparation to data acquisition and data pre-processing, to ensure that the analysis is consistent, comparable, reproducible, precise and accurate. While well established QA/QC procedures in metabolomics serve as valuable benchmarks, such as the 2018 guidelines by Broadhurst et al. [142] (**Figure 13**), they may require adaptation for exposomics studies or research endeavours that integrate both metabolomics and exposomics. Notably, filtering features based on QC pooled samples may inadvertently exclude low-abundance small molecules, including exogenous compounds typically present at trace levels, which may be significant in certain groups or conditions. To address this issue, an alternative approach involves preparing tailored pooled QC samples for specific study groups, such as cases and controls [167]. Currently, the usage of QC pooled samples varies greatly across studies accompanied by inconsistencies in the reporting of their preparation methods, as recently highlighted by Broeckling et al. [168]. Therefore, it would be beneficial to establish guidelines for the preparation, usage, and reporting of QC samples in metabolomics and exposomics studies. A minimum set of QC measures as well as a standardized method for reporting them should be required in future studies. **Table 1** summarizes a structure summary of recommended QA/QC for each stage of the non-target HRMS metabolomics and exposomics workflow, offering a practical starting point for harmonized QA/QC reporting in future exposomics studies.

The NORMAN network released their guidance on suspect and non-target screening for environmental monitoring in 2023 [15], which has many parallel applications in the exposomics community. This guidance offers recommendations for all steps in a non-target screening experiment, from sample preparation to HRMS and data analysis and reporting. Additionally, the PARC initiative (environmental and biomonitoring communities) proposed their harmonized QA/QC procedures for data pre-processing of non-target and suspect screening LC-HRMS data in 2024 [54]. These are very good starting points to be followed by the exposomics community, and these efforts indicate that the field is moving towards standardized workflows, which will hopefully emerge in the coming years.

Finally, artificial intelligence (AI) approaches (grey area of **Figure 13**), including Machine Learning (ML) approaches, are increasingly integrated into metabolomics and exposomics fields. AI can support various steps of the non-target HRMS workflow, such as feature prioritization, compound annotation, prediction modelling and pathways analysis. Although some of these approaches are still in their early stages and not yet widely and readily integrated into small molecule omics workflows, they are key starting points for supporting the new era of omics research [178, 179]. The shift from having most of the features unannotated to actively predicting the structure and biological impact of thousands of molecules is essential for translating exposomics research findings into precision medicine.

6. Conclusions and future perspectives

Genes only explain a small fraction of chronic diseases, highlighting the need for a broader approach to better understand health and disease etiology [180]. In this context, the significant global health burden of environmental factors, such as the 9 million premature deaths attributed to pollution in 2019 [181], underscores the critical role of environmental factors. Consequently, exposomics has emerged as a complementary field to genomics to study environmental drivers of disease. However, translating these omics insights into clinical practice requires overcoming several challenges including the large number of unannotated features, the lack of harmonization, poor reproducibility across studies, and ethical, legal and social issues related to human exposomics data [182]. To address these challenges, substantial investment in exposomics research and the establishment of large-scale international consortia are urgently needed. Recent initiatives such as NEXUS are pivotal in this regard, fostering collaboration among scientists, policymakers and funders to accelerate innovation and standardize practices.

The increasing availability of open data (**Figure 11**) on relevant environmental chemicals and the matrices in which they are detected, is crucial to assess the suitability of our current analytical approaches to capture the chemical exposome. This data will enable a critical evaluation of whether existing methods can effectively capture relevant chemicals and downstream biological processes or whether additional analytical techniques, such as supercritical fluid chromatography [183] or two-dimensional chromatography, will be needed for a comprehensive characterization of the human exposome. However, while open data is important it is not always possible, particularly for human datasets, where privacy and ethical constraints may limit open access. To address these challenges, tiered or control data access, embargo periods and data use agreements can be employed.

Over the last years, the field has evolved significantly, bringing the future of exposomics and metabolomics closer to the well-established omics fields such as genomics, transcriptomics, and proteomics. The exposomics field is currently moving from proof-of-concept studies, with low sample sizes, to ExWAS, which include thousands of participants and a broad range of environmental exposures and disease endpoints [184]. A key example is the HELIX study, which focuses on a cohort of more than 1,000 mother-child pairs and demonstrated that early life exposures can lead to biological responses detectable through different omics layers [24, 185]. These findings highlight that exposomics can identify early life biomarkers of exposure, which can improve our understanding of health and disease status and promote public health policies. Another recent ExWAS explored environmental exposures associated with aging in the UK Biobank [186]. Twenty-five independent exposures were associated with mortality and proteomic aging. Notably, this study revealed a greater impact of the exposome on variation in mortality than polygenic risk scores [184, 186]. This progress is essential for a future where integrating data from different omics will allow for a more comprehensive assessment of an individual's disease risk, thereby facilitating personalized medicine. In the long term, the findings from exposomics studies will be instrumental in guiding policymakers to implement measures that protect future generations from harmful environmental exposures.

Glossary

Exposomics: Refers to the study of the exposome, i.e., characterization of small molecules environmental derived and its transformation products within an entity (cell, tissue, or organism).

Feature: Peak (or signal) identified at a specific retention time, and m/z, containing spectral information such as MS1 and/or MS2, and intensity.

Metabolomics: Systematic and comprehensive study of low molecular weight molecules in a particular biological sample [39, 40].

MS1: In Mass Spectrometry refers to the full scan information of the precursor ion (also known as parent ion) including the information of adducts, isotopic pattern and in-source fragments.

MS2: Also known as MS/MS refers to the fragmentation pattern of the precursor ion.

Non-target: This can refer to non-target *study* or non-target *compound*. *Non-target study* also known as untarget or untargeted, refers to discovery-based studies aiming to identify as many compounds as possible (known and unknown), whereas *non-target compound* refers to a compound for which no target or suspect identity can be assigned readily. The term not-target screening refers to the computational strategy searching for a broad range of compounds via tandem mass spectral libraries or database search.

Small molecule omics: Refers to the study of low molecule weight molecules in a particular biological or environmental sample. It encompasses metabolomics and exposomics.

Suspect: Suspects can refer to known *compounds* that are expected to be in the sample but with insufficient standard information to be identified. Suspect *screening* refers to a computational strategy searching for known chemicals that are expected to be in the sample.

Target: This can refer to target *study* or target *compound*. *Target study* also known as targeted, refers to validation-based studies focused on a limited number of known compounds. *Target compound* refers to known compound, preselected for the analysis, with reference standard data, including MS2 and retention time, available for the unequivocal identification.

Acknowledgments

The current and former members of the Environmental Cheminformatics (ECI) group at the LCB are acknowledged for their valuable inputs and discussion, some of which have guided and inspired the ideas discussed and figures presented in this manuscript. We wish to acknowledge that AI was used by BTA to assist with generating the code for a few visualizations (Figure 11 & 12), and for support to polish some writing and finding synonyms for contents of original text.

Disclosure

This manuscript is based on unpublished portions of the PhD thesis of Begoña Talavera Andújar, defended on July 2024 and publicly available at the University of Luxembourg repository ORBilu (https://orbilu.uni.lu/handle/10993/61575).

Funding

BTA acknowledges the support of the “Microbiomes in One Health” PhD training program, which is supported by the PRIDE doctoral research funding scheme (PRIDE/11823097) of the Luxembourg National Research Fund (FNR). ELS acknowledges funding support from the Luxembourg National Research Fund (FNR) for project A18/BM/12341006.

Data availability

No new data were generated or analysed in support of this research. The code associated with Figure 11 (which retrieves relevant data from open repositories) is available in the ECI GitLab repository (https://gitlab.com/uniluxembourg/lcsb/eci/exposomics-plots).

Conflict of interest

No declared.

CRediT authorship contribution statement

Begoña Talavera Andújar: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization (lead), Writing – original draft (lead), Writing – review & editing. Emma L. Schymanski: Conceptualization, Funding acquisition, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing.

References

1. Rappaport SM, Smith MT (2010) Environment and Disease Risks. Science 330:460–461. https://doi.org/10.1126/science.1192603
2. David A, Chaker J, Price EJ, Bessonneau V, Chetwynd AJ, Vitale CM, Klánová J, Walker DI, Antignac J-P, Barouki R, Miller GW (2021) Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives. Environ Int 156:106630. https://doi.org/10.1016/j.envint.2021.106630
3. Wild CP (2005) Complementing the Genome with an ``Exposome'': The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. Cancer Epidemiol Biomarkers Prev 14:1847–1850. https://doi.org/10.1158/1055-9965.EPI-05-0456
4. Miller GW, Jones DP (2014) The Nature of Nurture: Refining the Definition of the Exposome. Toxicol Sci 137:1–2. https://doi.org/10.1093/toxsci/kft251
5. Miller GW, Banbury Exposomics Consortium, Bennett LM, Balshaw D, Barouki R, Bhutani G, Dolinoy D, Gao P, Jett D, Karagas M, Klánová J, Lein P, Li S, Metz TO, Patel CJ, Pollitt K, Rajasekar A, Sillé F, Thessen A, Thuault-Restituito S, Vermeulen R, Ward-Caviness CK, Wright R (2025) Integrating exposomics into biomedicine. Science 388:356–358. https://doi.org/10.1126/science.adr0544
6. Safarlou CW, Jongsma KR, Vermeulen R (2024) Reconceptualizing and Defining Exposomics within Environmental Health: Expanding the Scope of Health Research. Environ Health Perspect 132:095001. https://doi.org/10.1289/EHP14509
7. Roel Vermeulen Human exposome research: Potential, limitations and public policy implications | Think Tank | European Parliament.

https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2025)765791. Accessed 29 Aug 2025

8. Wild CP (2012) The exposome: from concept to utility. Int J Epidemiol 41:24–32. https://doi.org/10.1093/ije/dyr236

9. Vrijheid M (2014) The exposome: a new paradigm to study the impact of environment on health. Thorax 69:876–878. https://doi.org/10.1136/thoraxjnl-2013-204949

10. Guillien A, Ghosh M, Gille T, Dumas O (2023) The exposome concept: how has it changed our understanding of environmental causes of chronic respiratory diseases? Breathe 19:230044. https://doi.org/10.1183/20734735.0044-2023

11. La Cognata V, Morello G, Cavallaro S (2021) Omics Data and Their Integrative Analysis to Support Stratified Medicine in Neurodegenerative Diseases. Int J Mol Sci 22:4820. https://doi.org/10.3390/ijms22094820

12. Babu M, Snyder M (2023) Multi-Omics Profiling for Health. Mol Cell Proteomics 22:. https://doi.org/10.1016/j.mcpro.2023.100561

13. Botas A, Campbell HM, Han X, Maletic-Savatic M (2015) Metabolomics of Neurodegenerative Diseases. In: International Review of Neurobiology. Elsevier, pp 53–80

14. Roberts LD, Souza AL, Gerszten RE, Clish CB (2012) Targeted Metabolomics. Curr Protoc Mol Biol 98:. https://doi.org/10.1002/0471142727.mb3002s98

15. Hollender J, Schymanski EL, Ahrens L, Alygizakis N, Béen F, Bijlsma L, Brunner AM, 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 RR, Slobodnik J, Steininger-Mairinger T, Thomaidis NS, Togola A, Vorkamp K, Vulliet E, Zhu L, Krauss M (2023) NORMAN guidance on suspect and non-target screening in environmental monitoring. Environ Sci Eur 35:75. https://doi.org/10.1186/s12302-023-00779-4

16. Wishart DS (2019) Metabolomics for Investigating Physiological and Pathophysiological Processes. Physiol Rev 99:1819–1875. https://doi.org/10.1152/physrev.00035.2018

17. Wishart DS (2011) Advances in metabolite identification. Bioanalysis 3:1769–1782. https://doi.org/10.4155/bio.11.155

18. Metz TO, Baker ES, Schymanski EL, Renslow RS, Thomas DG, Causon TJ, Webb IK, Hann S, Smith RD, Teeguarden JG (2017) Integrating Ion Mobility Spectrometry Into Mass Spectrometry-Based Exposome Measurements: What Can it Add and How Far Can it Go? Bioanalysis 9:81–98. https://doi.org/10.4155/bio-2016-0244

19. Walker DI, Valvi D, Rothman N, Lan Q, Miller GW, Jones DP (2019) The metabolome: A key measure for exposome research in epidemiology. Curr Epidemiol Rep 6:93–103

20. Balcells C, Xu Y, Gil-Solsona R, Maitre L, Gago-Ferrero P, Keun HC (2024) Blurred lines: Crossing the boundaries between the chemical exposome and the metabolome. Curr Opin Chem Biol 78:102407. https://doi.org/10.1016/j.cbpa.2023.102407

21. Lai Y, Koelmel JP, Walker DI, Price EJ, Papazian S, Manz KE, Castilla-Fernández D, Bowden JA, Nikiforov V, David A, Bessonneau V, Amer B, Seethapathy S, Hu X, Lin EZ, Jbebli A, McNeil BR, Barupal D, Cerasa M, Xie H, Kalia V, Nandakumar R, Singh R, Tian Z, Gao P, Zhao Y, Froment J, Rostkowski P, Dubey S, Coufalíková K, Seličová H, Hecht H, Liu S, Udhani HH, Restituito S, Tchou-Wong K-M, Lu K, Martin

JW, Warth B, Godri Pollitt KJ, Klánová J, Fiehn O, Metz TO, Pennell KD, Jones DP, Miller GW (2024) High-Resolution Mass Spectrometry for Human Exposomics: Expanding Chemical Space Coverage. Environ Sci Technol 58:12784–12822. https://doi.org/10.1021/acs.est.4c01156

22. Zhang P, Carlsten C, Chaleckis R, Hanhineva K, Huang M, Isobe T, Koistinen VM, Meister I, Papazian S, Sdougkou K, Xie H, Martin JW, Rappaport SM, Tsugawa H, Walker DI, Woodruff TJ, Wright RO, Wheelock CE (2021) Defining the Scope of Exposome Studies and Research Needs from a Multidisciplinary Perspective. Environ Sci Technol Lett 8:839–852. https://doi.org/10.1021/acs.estlett.1c00648

23. (2020) All of Us Research Program | National Institutes of Health (NIH). In: Us Res. Program NIH. https://allofus.nih.gov/future-health-begins-all-us. Accessed 10 May 2024

24. Vrijheid M, Slama R, Robinson O, Chatzi L, Coen M, van den Hazel P, Thomsen C, Wright J, Athersuch TJ, Avellana N, Basagaña X, Brochot C, Bucchini L, Bustamante M, Carracedo A, Casas M, Estivill X, Fairley L, van Gent D, Gonzalez JR, Granum B, Gražulevičienė R, Gutzkow KB, Julvez J, Keun HC, Kogevinas M, McEachan RRC, Meltzer HM, Sabidó E, Schwarze PE, Siroux V, Sunyer J, Want EJ, Zeman F, Nieuwenhuijsen MJ (2014) The human early-life exposome (HELIX): project rationale and design. Environ Health Perspect 122:535–544. https://doi.org/10.1289/ehp.1307204

25. Barnes S, Benton HP, Casazza K, Cooper SJ, Cui X, Du X, Engler J, Kabarowski JH, Li S, Pathmasiri W, Prasain JK, Renfrow MB, Tiwari HK (2016) Training in metabolomics research. I. Designing the experiment, collecting and extracting samples and generating metabolomics data. J Mass Spectrom 51:461–475. https://doi.org/10.1002/jms.3782

26. Faul F, Erdfelder E, Lang A-G, Buchner A (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39:175–191. https://doi.org/10.3758/BF03193146

27. Pang Z, Lu Y, Zhou G, Hui F, Xu L, Viau C, Spigelman AF, MacDonald PE, Wishart DS, Li S, Xia J (2024) MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. Nucleic Acids Res gkae253. https://doi.org/10.1093/nar/gkae253

28. Tarazona S, Balzano-Nogueira L, Gómez-Cabrero D, Schmidt A, Imhof A, Hankemeier T, Tegnér J, Westerhuis JA, Conesa A (2020) Harmonization of quality metrics and power calculation in multi-omic studies. Nat Commun 11:3092. https://doi.org/10.1038/s41467-020-16937-8

29. Dennis KK, Marder E, Balshaw DM, Cui Y, Lynes MA, Patti GJ, Rappaport SM, Shaughnessy DT, Vrijheid M, Barr DB (2017) Biomonitoring in the Era of the Exposome. Environ Health Perspect 125:502–510. https://doi.org/10.1289/EHP474

30. Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A (2014) The Blood Exposome and Its Role in Discovering Causes of Disease. Environ Health Perspect 122:769–774. https://doi.org/10.1289/ehp.1308015

31. Jacobson TA, Kler JS, Bae Y, Chen J, Ladror DT, Iyer R, Nunes DA, Montgomery ND, Pleil JD, Funk WE (2023) A state-of-the-science review and guide for measuring environmental exposure biomarkers in dried blood spots. J Expo Sci Environ Epidemiol 33:505–523. https://doi.org/10.1038/s41370-022-00460-7

32. Talavera Andújar B, Pereira SL, Bhanu Busi S, Usnich T, Borsche M, Ertan S, Bauer P, Rolfs A, Hezzaz S, Ghelfi J, Brüggemann N, Antony P, Wilmes P, Klein C, Grünewald A, Schymanski EL (2024) Exploring environmental modifiers of LRRK2-associated Parkinson's disease penetrance: An exposomics and metagenomics pilot study on household dust. Environ Int 109151. https://doi.org/10.1016/j.envint.2024.109151

33. Running LS, Kordas K, Aga DS (2023) Use of wristbands to measure exposure to environmental pollutants in children: Recent advances and future directions. Curr Opin Environ Sci Health 32:100450. https://doi.org/10.1016/j.coesh.2023.100450

34. Wacławik M, Rodzaj W, Wielgomas B (2022) Silicone Wristbands in Exposure Assessment: Analytical Considerations and Comparison with Other Approaches. Int J Environ Res Public Health 19:1935. https://doi.org/10.3390/ijerph19041935

35. Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA (2016) Untargeted Metabolomics Strategies—Challenges and Emerging Directions. J Am Soc Mass Spectrom 27:1897–1905. https://doi.org/10.1007/s13361-016-1469-y

36. Gu Y, Peach JT, Warth B (2023) Sample preparation strategies for mass spectrometry analysis in human exposome research: Current status and future perspectives. TrAC Trends Anal Chem 166:117151. https://doi.org/10.1016/j.trac.2023.117151

37. Vuckovic D (2012) Current trends and challenges in sample preparation for global metabolomics using liquid chromatography–mass spectrometry. Anal Bioanal Chem 403:1523–1548. https://doi.org/10.1007/s00216-012-6039-y

38. Alseekh S, Aharoni A, Brotman Y, Contrepois K, D'Auria J, Ewald J, C. Ewald J, Fraser PD, Giavalisco P, Hall RD, Heinemann M, Link H, Luo J, Neumann S, Nielsen J, Perez de Souza L, Saito K, Sauer U, Schroeder FC, Schuster S, Siuzdak G, Skirycz A, Sumner LW, Snyder MP, Tang H, Tohge T, Wang Y, Wen W, Wu S, Xu G, Zamboni N, Fernie AR (2021) Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. Nat Methods 18:747–756. https://doi.org/10.1038/s41592-021-01197-1

39. Zeki ÖC, Eylem CC, Reçber T, Kır S, Nemutlu E (2020) Integration of GC–MS and LC–MS for untargeted metabolomics profiling. J Pharm Biomed Anal 190:113509. https://doi.org/10.1016/j.jpba.2020.113509

40. Fiehn O (2016) Metabolomics by Gas Chromatography–Mass Spectrometry: Combined Targeted and Untargeted Profiling. Curr Protoc Mol Biol 114:. https://doi.org/10.1002/0471142727.mb3004s114

41. Zhang A, Sun H, Wang P, Han Y, Wang X (2012) Modern analytical techniques in metabolomics analysis. The Analyst 137:293–300. https://doi.org/10.1039/C1AN15605E

42. Wang JH, Byun J, Pennathur S (2010) Analytical Approaches to Metabolomics and Applications to Systems Biology. Semin Nephrol 30:500–511. https://doi.org/10.1016/j.semnephrol.2010.07.007

43. Zhang X, Quinn K, Cruickshank-Quinn C, Reisdorph R, Reisdorph N (2018) The application of ion mobility mass spectrometry to metabolomics. Curr Opin Chem Biol 42:60–66. https://doi.org/10.1016/j.cbpa.2017.11.001

44. Izquierdo-Sandoval D, Fabregat-Safont D, Lacalle-Bergeron L, Sancho JV, Hernández F, Portoles T (2022) Benefits of Ion Mobility Separation in GC-APCI-

HRMS Screening: From the Construction of a CCS Library to the Application to Real-World Samples. Anal Chem 94:9040–9047. https://doi.org/10.1021/acs.analchem.2c01118

45. Ropartz D, Fanuel M, Ujma J, Palmer M, Giles K, Rogniaux H (2019) Structure Determination of Large Isomeric Oligosaccharides of Natural Origin through Multipass and Multistage Cyclic Traveling-Wave Ion Mobility Mass Spectrometry. Anal Chem 91:12030–12037. https://doi.org/10.1021/acs.analchem.9b03036

46. Collins SL, Koo I, Peters JM, Smith PB, Patterson AD (2021) Current Challenges and Recent Developments in Mass Spectrometry–Based Metabolomics. Annu Rev Anal Chem 14:467–487. https://doi.org/10.1146/annurev-anchem-091620-015205

47. Kind T, Tsugawa H, Cajka T, Ma Y, Lai Z, Mehta SS, Wohlgemuth G, Barupal DK, Showalter MR, Arita M, Fiehn O (2018) Identification of small molecules using accurate mass MS/MS search. Mass Spectrom Rev 37:513–532. https://doi.org/10.1002/mas.21535

48. Guo J, Huan T (2020) Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics. Anal Chem 92:8072–8080. https://doi.org/10.1021/acs.analchem.9b05135

49. Yang Y, Yang L, Zheng M, Cao D, Liu G (2023) Data acquisition methods for non-targeted screening in environmental analysis. TrAC Trends Anal Chem 160:116966. https://doi.org/10.1016/j.trac.2023.116966

50. Koelmel JP, Kroeger NM, Gill EL, Ulmer CZ, Bowden JA, Patterson RE, Yost RA, Garrett TJ (2017) Expanding Lipidome Coverage Using LC-MS/MS Data-Dependent Acquisition with Automated Exclusion List Generation. J Am Soc Mass Spectrom 28:908–917. https://doi.org/10.1007/s13361-017-1608-0

51. Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, Kanazawa M, VanderGheynst J, Fiehn O, Arita M (2015) MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods 12:523–526. https://doi.org/10.1038/nmeth.3393

52. Goodacre R, Broadhurst D, Smilde AK, Kristal BS, Baker JD, Beger R, Bessant C, Connor S, Capuani G, Craig A, Ebbels T, Kell DB, Manetti C, Newton J, Paternostro G, Somorjai R, Sjöström M, Trygg J, Wulfert F (2007) Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics 3:231–241. https://doi.org/10.1007/s11306-007-0081-3

53. Katajamaa M, Orešič M (2007) Data processing for mass spectrometry-based metabolomics. J Chromatogr A 1158:318–328. https://doi.org/10.1016/j.chroma.2007.04.021

54. Lennon S, Chaker J, Price EJ, Hollender J, Huber C, Schulze T, Ahrens L, Béen F, Creusot N, Debrauwer L, Dervilly G, Gabriel C, Guérin T, Habchi B, Jamin EL, Klánová J, Kosjek T, Le Bizec B, Meijer J, Mol H, Nijssen R, Oberacher H, Papaioannou N, Parinet J, Sarigiannis D, Stravs MA, Tkalec Ž, Schymanski EL, Lamoree M, Antignac J-P, David A (2024) Harmonized quality assurance/quality control provisions to assess completeness and robustness of MS1 data preprocessing for LC-HRMS-based suspect screening and non-targeted analysis. TrAC Trends Anal Chem 174:117674. https://doi.org/10.1016/j.trac.2024.117674

55. Schmid R, Heuckeroth S, Korf A, Smirnov A, Myers O, Dyrlund TS, Bushuiev R, Murray KJ, Hoffmann N, Lu M, Sarvepalli A, Zhang Z, Fleischauer M, Dührkop K, Wesner M, Hoogstra SJ, Rudt E, Mokshyna O, Brungs C, Ponomarov K, Mutabdžija L, Damiani T, Pudney CJ, Earll M, Helmer PO, Fallon TR, Schulze T, Rivas-Ubach A, Bilbao A, Richter H, Nothias L-F, Wang M, Orešič M, Weng J-K, Böcker S, Jeibmann A, Hayen H, Karst U, Dorrestein PC, Petras D, Du X, Pluskal T (2023) Integrative analysis of multimodal mass spectrometry data in MZmine 3. Nat Biotechnol 41:447–449. https://doi.org/10.1038/s41587-023-01690-2

56. Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, Kohlbacher O (2008) OpenMS – An open-source software framework for mass spectrometry. BMC Bioinformatics 9:163. https://doi.org/10.1186/1471-2105-9-163

57. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G (2012) XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. Anal Chem 84:5035–5039. https://doi.org/10.1021/ac300698c

58. Helmus R, Ter Laak TL, Van Wezel AP, De Voogt P, Schymanski EL (2021) patRoon: open source software platform for environmental mass spectrometry based non-target screening. J Cheminformatics 13:1. https://doi.org/10.1186/s13321-020-00477-w

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

60. Renner G, Reuschenbach M (2023) Critical review on data processing algorithms in non-target screening: challenges and opportunities to improve result comparability. Anal Bioanal Chem 415:4111–4123. https://doi.org/10.1007/s00216-023-04776-7

61. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M-Y, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P (2012) A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30:918–920. https://doi.org/10.1038/nbt.2377

62. ProteoWizard: Home. https://proteowizard.sourceforge.io/index.html. Accessed 2 Apr 2024

63. Holman JD, Tabb DL, Mallick P (2014) Employing ProteoWizard to Convert Raw Mass Spectrometry Data. Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al 46:13.24.1-13.24.9. https://doi.org/10.1002/0471250953.bi1324s46

64. Santamaria G, Pinto FR (2024) Bioinformatic Analysis of Metabolomic Data: From Raw Spectra to Biological Insight. BioChem 4:90–114. https://doi.org/10.3390/biochem4020005

65. Karaman I, Climaco Pinto R, Graça G (2018) Metabolomics Data Preprocessing: From Raw Data to Features for Statistical Analysis. In: Comprehensive Analytical Chemistry. Elsevier, pp 197–225

66. Boccard J, Veuthey J, Rudaz S (2010) Knowledge discovery in metabolomics: An overview of MS data handling. J Sep Sci 33:290–304. https://doi.org/10.1002/jssc.200900609

67. Savitzky Abraham, Golay MJE (1964) Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal Chem 36:1627–1639. https://doi.org/10.1021/ac60214a047

68. Lommen A (2009) MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. Anal Chem 81:3079–3086. https://doi.org/10.1021/ac900036d

69. Libiseller G, Dvorzak M, Kleb U, Gander E, Eisenberg T, Madeo F, Neumann S, Trausinger G, Sinner F, Pieber T, Magnes C (2015) IPO: a tool for automated optimization of XCMS parameters. BMC Bioinformatics 16:118. https://doi.org/10.1186/s12859-015-0562-8

70. Blaženović I, Kind T, Ji J, Fiehn O (2018) Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. Metabolites 8:31. https://doi.org/10.3390/metabo8020031

71. Chaleckis R, Meister I, Zhang P, Wheelock CE (2019) Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. Curr Opin Biotechnol 55:44–50. https://doi.org/10.1016/j.copbio.2018.07.010

72. Oberacher H, Sasse M, Antignac J-P, Guitton Y, Debrauwer L, Jamin EL, Schulze T, Krauss M, Covaci A, Caballero-Casero N, Rousseau K, Damont A, Fenaille F, Lamoree M, Schymanski EL (2020) A European proposal for quality control and quality assurance of tandem mass spectral libraries. Environ Sci Eur 32:43. https://doi.org/10.1186/s12302-020-00314-9

73. de Jonge NF, Mildau K, Meijer D, Louwen JJR, Bueschl C, Huber F, van der Hooft JJJ (2022) Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. Metabolomics 18:103. https://doi.org/10.1007/s11306-022-01963-y

74. Giera M, Aisporna A, Uritboonthai W, Siuzdak G (2024) The hidden impact of in-source fragmentation in metabolic and chemical mass spectrometry data interpretation. Nat Metab 6:1647–1648. https://doi.org/10.1038/s42255-024-01076-x

75. El Abiead Y, Rutz A, Zuffa S, Amer B, Xing S, Brungs C, Schmid R, Correia MSP, Caraballo-Rodriguez AM, Zarrinpar A, Mannochio-Russo H, Witting M, Mohanty I, Pluskal T, Bittremieux W, Knight R, Patterson AD, van der Hooft JJJ, Böcker S, Dunn WB, Linington RG, Wishart DS, Wolfender J-L, Fiehn O, Zamboni N, Dorrestein PC (2025) Discovery of metabolites prevails amid in-source fragmentation. Nat Metab 7:435–437. https://doi.org/10.1038/s42255-025-01239-4

76. Schmid R, Petras D, Nothias L-F, Wang M, Aron AT, Jagels A, Tsugawa H, Rainer J, Garcia-Aloy M, Dührkop K, Korf A, Pluskal T, Kameník Z, Jarmusch AK, Caraballo-Rodríguez AM, Weldon KC, Nothias-Esposito M, Aksenov AA, Bauermeister A, Albarracin Orio A, Grundmann CO, Vargas F, Koester I, Gauglitz JM, Gentry EC, Hövelmann Y, Kalinina SA, Pendergraft MA, Panitchpakdi M, Tehan R, Le Gouellec A, Aleti G, Mannochio Russo H, Arndt B, Hübner F, Hayen H, Zhi H, Raffatellu M, Prather KA, Aluwihare LI, Böcker S, McPhail KL, Humpf H-U, Karst U, Dorrestein PC (2021) Ion identity molecular networking for mass spectrometry-based metabolomics

in the GNPS environment. Nat Commun 12:3832. https://doi.org/10.1038/s41467-021-23953-9

77. Xu Y-F, Lu W, Rabinowitz JD (2015) Avoiding misannotation of in-source fragmentation products as cellular metabolites in liquid chromatography-mass spectrometry-based metabolomics. Anal Chem 87:2273–2281. https://doi.org/10.1021/ac504118y

78. Xue J, Domingo-Almenara X, Guijas C, Palermo A, Rinschen MM, Isbell J, Benton HP, Siuzdak G (2020) Enhanced in-Source Fragmentation Annotation Enables Novel Data Independent Acquisition and Autonomous METLIN Molecular Identification. Anal Chem 92:6051–6059. https://doi.org/10.1021/acs.analchem.0c00409

79. Krier J, Singh RR, Kondić T, Lai A, Diderich P, Zhang J, Thiessen PA, Bolton EE, Schymanski EL (2022) Discovering pesticides and their TPs in Luxembourg waters using open cheminformatics approaches. Environ Int 158:106885. https://doi.org/10.1016/j.envint.2021.106885

80. Hollender J, Bourgin M, Fenner KB, Longrée P, Mcardell CS, Moschet C, Ruff M, Schymanski EL, Singer HP (2014) Exploring the Behaviour of Emerging Contaminants in the Water Cycle using the Capabilities of High Resolution Mass Spectrometry. CHIMIA 68:793. https://doi.org/10.2533/chimia.2014.793

81. Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP, Hollender J (2014) Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. Environ Sci Technol 48:2097–2098. https://doi.org/10.1021/es5002105

82. Koelmel JP, Kroeger NM, Ulmer CZ, Bowden JA, Patterson RE, Cochran JA, Beecher CWW, Garrett TJ, Yost RA (2017) LipidMatch: an automated workflow for rule-based lipid identification using untargeted high-resolution tandem mass spectrometry data. BMC Bioinformatics 18:331. https://doi.org/10.1186/s12859-017-1744-3

83. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TW-M, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR (2007) Proposed minimum reporting standards for chemical analysis. Metabolomics 3:211–221. https://doi.org/10.1007/s11306-007-0082-2

84. Celma A, Sancho JV, Schymanski EL, Fabregat-Safont D, Ibáñez M, Goshawk J, Barknowitz G, Hernández F, Bijlsma L (2020) Improving Target and Suspect Screening High-Resolution Mass Spectrometry Workflows in Environmental Analysis by Ion Mobility Separation. Environ Sci Technol 54:15120–15131. https://doi.org/10.1021/acs.est.0c05713

85. Charbonnet JA, McDonough CA, Xiao F, Schwichtenberg T, Cao D, Kaserzon S, Thomas KV, Dewapriya P, Place BJ, Schymanski EL, Field JA, Helbling DE, Higgins CP (2022) Communicating Confidence of Per- and Polyfluoroalkyl Substance Identification via High-Resolution Mass Spectrometry. Environ Sci Technol Lett 9:473–481. https://doi.org/10.1021/acs.estlett.2c00206

86. Koelmel JP, Xie H, Price EJ, Lin EZ, Manz KE, Stelben P, Paige MK, Papazian S, Okeme J, Jones DP, Barupal D, Bowden JA, Rostkowski P, Pennell KD, Nikiforov V, Wang T, Hu X, Lai Y, Miller GW, Walker DI, Martin JW, Godri Pollitt KJ (2022) An actionable annotation scoring framework for gas chromatography-high-resolution

mass spectrometry. Exposome 2:osac007. https://doi.org/10.1093/exposome/osac007

87. Schymanski EL, Kondić T, Neumann S, Thiessen PA, Zhang J, Bolton EE (2021) Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag. J Cheminformatics 13:19. https://doi.org/10.1186/s13321-021-00489-0

88. Creek DJ, Dunn WB, Fiehn O, Griffin JL, Hall RD, Lei Z, Mistrik R, Neumann S, Schymanski EL, Sumner LW, Trengove R, Wolfender J-L (2014) Metabolite identification: are you sure? And how do your peers gauge your confidence? Metabolomics 10:350–353. https://doi.org/10.1007/s11306-014-0656-8

89. Liebisch G, Vizcaíno JA, Köfeler H, Trötzmüller M, Griffiths WJ, Schmitz G, Spener F, Wakelam MJO (2013) Shorthand notation for lipid structures derived from mass spectrometry. J Lipid Res 54:1523–1530. https://doi.org/10.1194/jlr.M033506

90. Scheubert K, Hufsky F, Petras D, Wang M, Nothias L-F, Dührkop K, Bandeira N, Dorrestein PC, Böcker S (2017) Significance estimation for large scale metabolomics annotations by spectral matching. Nat Commun 8:1494. https://doi.org/10.1038/s41467-017-01318-5

91. Metz TO, Chang CH, Gautam V, Anjum A, Tian S, Wang F, Colby SM, Nunez JR, Blumer MR, Edison AS, Fiehn O, Jones DP, Li S, Morgan ET, Patti GJ, Ross DH, Shapiro MR, Williams AJ, Wishart DS (2025) Introducing "Identification Probability" for Automated and Transferable Assessment of Metabolite Identification Confidence in Metabolomics and Related Studies. Anal Chem 97:1–11. https://doi.org/10.1021/acs.analchem.4c04060

92. Talavera Andújar B, Aurich D, Aho VTE, Singh RR, Cheng T, Zaslavsky L, Bolton EE, Mollenhauer B, Wilmes P, Schymanski EL (2022) Studying the Parkinson's disease metabolome and exposome in biological samples through different analytical and cheminformatics approaches: a pilot study. Anal Bioanal Chem 414:7399–7419. https://doi.org/10.1007/s00216-022-04207-z

93. Alygizakis N, Lestremau F, Gago-Ferrero P, Gil-Solsona R, Arturi K, Hollender J, Schymanski EL, Dulio V, Slobodnik J, Thomaidis NS (2023) Towards a harmonized identification scoring system in LC-HRMS/MS based non-target screening (NTS) of emerging contaminants. TrAC Trends Anal Chem 159:116944. https://doi.org/10.1016/j.trac.2023.116944

94. Boatman AK, Chappel JR, Kirkwood-Donelson KI, Fleming JF, Reif DM, Schymanski EL, Rager JE, Baker ES (2025) Updated Guidance for Communicating PFAS Identification Confidence with Ion Mobility Spectrometry. Environ Sci Technol 59:17711–17721. https://doi.org/10.1021/acs.est.5c01354

95. Bittremieux W, Wang M, Dorrestein PC (2022) The critical role that spectral libraries play in capturing the metabolomics community knowledge. Metabolomics 18:94. https://doi.org/10.1007/s11306-022-01947-y

96. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Tanaka K, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T (2010) MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom JMS 45:703–714. https://doi.org/10.1002/jms.1777

97. MassBank of North America. https://mona.fiehnlab.ucdavis.edu/. Accessed 11 Sept 2023

98. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, P CAB, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837. https://doi.org/10.1038/nbt.3597

99. mzCloud – Advanced Mass Spectral Database. https://www.mzcloud.org/. Accessed 28 Apr 2024

100. Guijas C, Montenegro-Burke JR, Domingo-Almenara X, Palermo A, Warth B, Hermann G, Koellensperger G, Huan T, Uritboonthai W, Aisporna AE, Wolan DW, Spilker ME, Benton HP, Siuzdak G (2018) METLIN: A Technology Platform for Identifying Knowns and Unknowns. Anal Chem 90:3156–3164. https://doi.org/10.1021/acs.analchem.7b04424

101. Stein SE, Scott DR (1994) Optimization and testing of mass spectral library search algorithms for compound identification. J Am Soc Mass Spectrom 5:859–866. https://doi.org/10.1016/1044-0305(94)87009-8

102. PubChem Classification Browser. https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72. Accessed 19 May 2024

103. Li Y, Kind T, Folz J, Vaniya A, Mehta SS, Fiehn O (2021) Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. Nat Methods 18:1524–1531. https://doi.org/10.1038/s41592-021-01331-z

104. Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, Kanazawa M, VanderGheynst J, Fiehn O, Arita M (2015) MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods 12:523–526. https://doi.org/10.1038/nmeth.3393

105. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminformatics 8:3. https://doi.org/10.1186/s13321-016-0115-9

106. Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, Cajka T, Saito K, Fiehn O, Arita M (2016) Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. Anal Chem 88:7946–7958. https://doi.org/10.1021/acs.analchem.6b00770

107. Allen F, Pon A, Wilson M, Greiner R, Wishart D (2014) CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. Nucleic Acids Res 42:W94–W99. https://doi.org/10.1093/nar/gku436

108. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci 112:12580–12585. https://doi.org/10.1073/pnas.1509788112

109. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, Dorrestein PC, Rousu J, Böcker S (2019) SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods 16:299–302. https://doi.org/10.1038/s41592-019-0344-8

110. Kind T, Liu K-H, Lee DY, DeFelice B, Meissen JK, Fiehn O (2013) LipidBlast in silico tandem mass spectrometry database for lipid identification. Nat Methods 10:755–758. https://doi.org/10.1038/nmeth.2551

111. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, Sajed T, Johnson D, Li C, Karu N, Sayeeda Z, Lo E, Assempour N, Berjanskii M, Singhal S, Arndt D, Liang Y, Badran H, Grant J, Serra-Cayuela A, Liu Y, Mandal R, Neveu V, Pon A, Knox C, Wilson M, Manach C, Scalbert A (2018) HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res 46:D608–D617. https://doi.org/10.1093/nar/gkx1089

112. CAS REGISTRY | CAS. https://www.cas.org/cas-data/cas-registry. Accessed 30 Apr 2024

113. Pence HE, Williams A (2010) ChemSpider: An Online Chemical Information Resource. J Chem Educ 87:1123–1124. https://doi.org/10.1021/ed100697w

114. Bolton EE, Wang Y, Thiessen PA, Bryant SH (2008) Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities. In: Wheeler RA, Spellmeyer DC (eds) Annual Reports in Computational Chemistry. Elsevier, pp 217–241

115. Bolton E, Schymanski E, Kondic T, Thiessen P, Zhang J (2020) PubChemLite for Exposomics

116. Elapavalore A, Ross DH, Grouès V, Aurich D, Krinsky AM, Kim S, Thiessen PA, Zhang J, Dodds JN, Baker ES, Bolton EE, Xu L, Schymanski EL (2025) PubChemLite Plus Collision Cross Section (CCS) Values for Enhanced Interpretation of Nontarget Environmental Data. Environ Sci Technol Lett 12:166–174. https://doi.org/10.1021/acs.estlett.4c01003

117. Barupal DK, Fiehn O (2019) Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach. Environ Health Perspect 127:097008. https://doi.org/10.1289/EHP4713

118. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34:D354-357. https://doi.org/10.1093/nar/gkj102

119. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 36:D623-631. https://doi.org/10.1093/nar/gkm900

120. Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, Merrill AH, Murphy RC, Raetz CRH, Russell DW, Subramaniam S (2007) LMSD: LIPID MAPS structure database. Nucleic Acids Res 35:D527-532. https://doi.org/10.1093/nar/gkl838

121. Wishart DS, Oler E, Peters H, Guo A, Girod S, Han S, Saha S, Lui VW, LeVatte M, Gautam V, Kaddurah-Daouk R, Karu N (2023) MiMeDB: the Human Microbial Metabolome Database. Nucleic Acids Res 51:D611–D620. https://doi.org/10.1093/nar/gkac868

122. Williams AJ, Grulke CM, Edwards J, McEachran AD, Mansouri K, Baker NC, Patlewicz G, Shah I, Wambaugh JF, Judson RS, Richard AM (2017) The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J Cheminformatics 9:61. https://doi.org/10.1186/s13321-017-0247-6

123. (2025) uniluxembourg / LCSB / Environmental Cheminformatics / Exposomics plots · GitLab. In: GitLab. https://gitlab.com/uniluxembourg/lcsb/eci/exposomics-plots. Accessed 21 July 2025

124. Samanipour S, Barron LP, Van Herwerden D, Praetorius A, Thomas KV, O'Brien JW (2024) Exploring the Chemical Space of the Exposome: How Far Have We Gone? JACS Au 4:2412–2425. https://doi.org/10.1021/jacsau.4c00220

125. Palm EH, Chirsir P, Krier J, Thiessen PA, Zhang J, Bolton EE, Schymanski EL (2023) ShinyTPs: Curating Transformation Products from Text Mining Results. Environ Sci Technol Lett 10:865–871. https://doi.org/10.1021/acs.estlett.3c00537

126. NORMAN-SLE. https://www.norman-network.com/nds/SLE/

127. Helmus R, Van De Velde B, Brunner AM, Ter Laak TL, Van Wezel AP, Schymanski EL (2022) patRoon 2.0: Improved non-target analysis workflowsincluding automated transformation product screening. J Open Source Softw 7:4029. https://doi.org/10.21105/joss.04029

128. Djoumbou-Feunang Y, Fiamoncini J, Gil-de-la-Fuente A, Greiner R, Manach C, Wishart DS (2019) BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. J Cheminformatics 11:2. https://doi.org/10.1186/s13321-018-0324-5

129. CompTox Chemicals Dashboard Chemical Lists. https://comptox.epa.gov/dashboard/chemical-lists. Accessed 30 Apr 2024

130. Helmus R, Van De Velde B, Brunner AM, Ter Laak TL, Van Wezel AP, Schymanski EL (2022) patRoon 2.0: Improved non-target analysis workflowsincluding automated transformation product screening. J Open Source Softw 7:4029. https://doi.org/10.21105/joss.04029

131. Gertsman I, Barshop BA (2018) Promises and pitfalls of untargeted metabolomics. J Inherit Metab Dis 41:355–366. https://doi.org/10.1007/s10545-017-0130-7

132. Karaman I (2017) Preprocessing and Pretreatment of Metabolomics Data for Statistical Analysis. In: Sussulini A (ed) Metabolomics: From Fundamentals to Clinical Applications. Springer International Publishing, Cham, pp 145–161

133. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7:142. https://doi.org/10.1186/1471-2164-7-142

134. Ren S, Hinzman AA, Kang EL, Szczesniak RD, Lu LJ (2015) Computational and statistical analysis of metabolomics data. Metabolomics 11:1492–1513. https://doi.org/10.1007/s11306-015-0823-6

135. Misra BB (2020) Data normalization strategies in metabolomics: Current challenges, approaches, and tools. Eur J Mass Spectrom 26:165–174. https://doi.org/10.1177/1469066720918446

136. Wu Y, Li L (2016) Sample normalization methods in quantitative metabolomics. J Chromatogr A 1430:80–95. https://doi.org/10.1016/j.chroma.2015.12.007

137. Barupal DK, Fan S, Fiehn O (2018) Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets. Curr Opin Biotechnol 54:1–9. https://doi.org/10.1016/j.copbio.2018.01.010

138. Chen Y, Li E-M, Xu L-Y (2022) Guide to Metabolomics Analysis: A Bioinformatics Workflow. Metabolites 12:357. https://doi.org/10.3390/metabo12040357

139. Vinaixa M, Samino S, Saez I, Duran J, Guinovart JJ, Yanes O (2012) A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. Metabolites 2:775–795. https://doi.org/10.3390/metabo2040775

140. Pang Z, Zhou G, Ewald J, Chang L, Hacariz O, Basu N, Xia J (2022) Using MetaboAnalyst 5.0 for LC–HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data. Nat Protoc 17:1735–1761. https://doi.org/10.1038/s41596-022-00710-w

141. Santos S, Maitre L, Warembourg C, Agier L, Richiardi L, Basagaña X, Vrijheid M (2020) Applying the exposome concept in birth cohort research: a review of statistical approaches. Eur J Epidemiol 35:193–204. https://doi.org/10.1007/s10654-020-00625-4

142. Broadhurst D, Goodacre R, Reinke SN, Kuligowski J, Wilson ID, Lewis MR, Dunn WB (2018) Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. Metabolomics Off J Metabolomic Soc 14:72. https://doi.org/10.1007/s11306-018-1367-3

143. Han W, Li L (2022) Evaluating and minimizing batch effects in metabolomics. Mass Spectrom Rev 41:421–442. https://doi.org/10.1002/mas.21672

144. Fan S, Kind T, Cajka T, Hazen SL, Tang WHW, Kaddurah-Daouk R, Irvin MR, Arnett DK, Barupal DK, Fiehn O (2019) Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data. Anal Chem 91:3590–3596. https://doi.org/10.1021/acs.analchem.8b05592

145. Yu Y, Mai Y, Zheng Y, Shi L (2024) Assessing and mitigating batch effects in large-scale omics studies. Genome Biol 25:254. https://doi.org/10.1186/s13059-024-03401-9

146. Chang L, Ewald J, Hui F, Bayen S, Xia J (2024) A data-centric perspective on exposomics data analysis. Exposome 4:osae005. https://doi.org/10.1093/exposome/osae005

147. Huang S-Y, Yang Y-X, Chen S-D, Li H-Q, Zhang X-Q, Kuo K, Tan L, Feng L, Dong Q, Zhang C, Yu J-T (2021) Investigating causal relationships between exposome and human longevity: a Mendelian randomization analysis. BMC Med 19:150. https://doi.org/10.1186/s12916-021-02030-4

148. Domenighetti C, Sugier P-E, Ashok Kumar Sreelatha A, Schulte C, Grover S, Mohamed O, Portugal B, May P, Bobbili DR, Radivojkov-Blagojevic M, Lichtner P, Singleton AB, Hernandez DG, Edsall C, Mellick GD, Zimprich A, Pirker W, Rogaeva E, Lang AE, Koks S, Taba P, Lesage S, Brice A, Corvol J-C, Chartier-Harlin M-C, Mutez E, Brockmann K, Deutschländer AB, Hadjigeorgiou GM, Dardiotis E, Stefanis L, Simitsi AM, Valente EM, Petrucci S, Duga S, Straniero L, Zecchinelli A, Pezzoli G, Brighina L, Ferrarese C, Annesi G, Quattrone A, Gagliardi M, Matsuo H, Kawamura Y, Hattori N, Nishioka K, Chung SJ, Kim YJ, Kolber P, van de Warrenburg BPC, Bloem BR, Aasly J, Toft M, Pihlstrøm L, Correia Guedes L, Ferreira JJ, Bardien S, Carr J, Tolosa E, Ezquerra M, Pastor P, Diez-Fairen M, Wirdefeldt K, Pedersen NL, Ran C, Belin AC, Puschmann A, Hellberg C, Clarke CE, Morrison KE, Tan M, Krainc D, Burbulla LF, Farrer MJ, Krüger R, Gasser T, Sharma M, Elbaz A, Comprehensive Unbiased Risk Factor Assessment for Genetics and Environment in Parkinson's Disease (Courage-PD) Consortium (2022) Dairy Intake and Parkinson's Disease: A Mendelian Randomization Study. Mov Disord Off J Mov Disord Soc 37:857–864. https://doi.org/10.1002/mds.28902

149. Li D, Zhou L, Cao Z, Wang J, Yang H, Lyu M, Zhang Y, Yang R, Wang J, Bian Y, Xu W, Wang Y (2024) Associations of environmental factors with neurodegeneration: An exposome-wide Mendelian randomization investigation. Ageing Res Rev 95:102254. https://doi.org/10.1016/j.arr.2024.102254

150. Zhao Y-L, Hao Y-N, Ge Y-J, Zhang Y, Huang L-Y, Fu Y, Zhang D-D, Ou Y-N, Cao X-P, Feng J-F, Cheng W, Tan L, Yu J-T (2025) Variables associated with cognitive function: an exposome-wide and mendelian randomization analysis. Alzheimers Res Ther 17:13. https://doi.org/10.1186/s13195-025-01670-5

151. Maitre L, Guimbaud J-B, Warembourg C, Güil-Oumrait N, Petrone PM, Chadeau-Hyam M, Vrijheid M, Basagaña X, Gonzalez JR (2022) State-of-the-art methods for exposure-health studies: Results from the exposome data challenge event. Environ Int 168:107422. https://doi.org/10.1016/j.envint.2022.107422

152. Barupal DK, Fiehn O (2017) Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. Sci Rep 7:14567. https://doi.org/10.1038/s41598-017-15231-w

153. Johnson CH, Ivanisevic J, Siuzdak G (2016) Metabolomics: beyond biomarkers and towards mechanisms. Nat Rev Mol Cell Biol 17:451–459. https://doi.org/10.1038/nrm.2016.25

154. Xia J (2017) Computational Strategies for Biological Interpretation of Metabolomics Data. In: Sussulini A (ed) Metabolomics: From Fundamentals to Clinical Applications. Springer International Publishing, Cham, pp 191–206

155. Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, Jones DP, Pulendran B (2013) Predicting Network Activity from High Throughput Metabolomics. PLOS Comput Biol 9:e1003123. https://doi.org/10.1371/journal.pcbi.1003123

156. Rohart F, Gautier B, Singh A, Cao K-AL (2017) mixOmics: An R package for ‘omics feature selection and multiple data integration. PLOS Comput Biol 13:e1005752. https://doi.org/10.1371/journal.pcbi.1005752

157. Zuffa S, Schmid R, Bauermeister A, P. Gomes PW, Caraballo-Rodriguez AM, El Abiead Y, Aron AT, Gentry EC, Zemlin J, Meehan MJ, Avalon NE, Cichewicz RH, Buzun E, Terrazas MC, Hsu C-Y, Oles R, Ayala AV, Zhao J, Chu H, Kuijpers MCM, Jackrel SL, Tugizimana F, Nephali LP, Dubery IA, Madala NE, Moreira EA, Costa-Lotufo LV, Lopes NP, Rezende-Teixeira P, Jimenez PC, Rimal B, Patterson AD, Traxler MF, Pessotti R de C, Alvarado-Villalobos D, Tamayo-Castillo G, Chaverri P, Escudero-Leyva E, Quiros-Guerrero L-M, Bory AJ, Joubert J, Rutz A, Wolfender J-L, Allard P-M, Sichert A, Pontrelli S, Pullman BS, Bandeira N, Gerwick WH, Gindro K, Massana-Codina J, Wagner BC, Forchhammer K, Petras D, Aiosa N, Garg N, Liebeke M, Bourceau P, Kang KB, Gadhavi H, de Carvalho LPS, Silva dos Santos M, Pérez-Lorente AI, Molina-Santiago C, Romero D, Franke R, Brönstrup M, Vera Ponce de León A, Pope PB, La Rosa SL, La Barbera G, Roager HM, Laursen MF, Hammerle F, Siewert B, Peintner U, Licona-Cassani C, Rodriguez-Orduña L, Rampler E, Hildebrand F, Koellensperger G, Schoeny H, Hohenwallner K, Panzenboeck L, Gregor R, O’Neill EC, Roxborough ET, Odoi J, Bale NJ, Ding S, Sinninghe Damsté JS, Guan XL, Cui JJ, Ju K-S, Silva DB, Silva FMR, da Silva GF, Koolen HHF, Grundmann C, Clement JA, Mohimani H, Broders K, McPhail KL, Ober-Singleton SE, Rath CM, McDonald D, Knight R, Wang M, Dorrestein PC (2024) microbeMASST: a taxonomically informed mass spectrometry search tool for microbial metabolomics data. Nat Microbiol 9:336–345. https://doi.org/10.1038/s41564-023-01575-9

158. Fiehn O, Robertson D, Griffin J, van der Werf M, Nikolau B, Morrison N, Sumner LW, Goodacre R, Hardy NW, Taylor C, Fostel J, Kristal B, Kaddurah-Daouk R, Mendes P, van Ommen B, Lindon JC, Sansone S-A (2007) The metabolomics standards initiative (MSI). Metabolomics 3:175–178. https://doi.org/10.1007/s11306-007-0070-6

159. Salek RM, Neumann S, Schober D, Hummel J, Billiau K, Kopka J, Correa E, Reijmers T, Rosato A, Tenori L, Turano P, Marin S, Deborde C, Jacob D, Rolin D, Dartigues B, Conesa P, Haug K, Rocca-Serra P, O’Hagan S, Hao J, van Vliet M, Sysi-Aho M, Ludwig C, Bouwman J, Cascante M, Ebbels T, Griffin JL, Moing A, Nikolski M, Oresic M, Sansone S-A, Viant MR, Goodacre R, Günther UL, Hankemeier T, Luchinat C, Walther D, Steinbeck C (2015) COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. Metabolomics 11:1587–1597. https://doi.org/10.1007/s11306-015-0810-y

160. Vineis P, Chadeau-Hyam M, Gmuender H, Gulliver J, Herceg Z, Kleinjans J, Kogevinas M, Kyrtopoulos S, Nieuwenhuijsen M, Phillips DH, Probst-Hensch N, Scalbert A, Vermeulen R, Wild CP, EXPOsOMICS Consortium (2017) The exposome in practice: Design of the EXPOsOMICS project. Int J Hyg Environ Health 220:142–151. https://doi.org/10.1016/j.ijheh.2016.08.001

161. Home - The European Human Exposome Network (EHEN). https://www.humanexposome.eu/. Accessed 3 May 2024

162. NEXUS. https://www.nexus-exposomics.org/

163. Niedzwiecki MM, Miller GW (2019) HERCULES: An Academic Center to Support Exposome Research. In: Dagnino S, Macherone A (eds) Unraveling the Exposome: A Practical View. Springer International Publishing, Cham, pp 339–348

164. Human Health Exposure Analysis Resource (HHEAR). https://hhearprogram.org/. Accessed 15 July 2025

165. IHEN - The International Human Exposome Network. In: IHEN. https://humanexposome.net/. Accessed 15 July 2025

166. Aharoni A, Goodacre R, Fernie AR (2023) Plant and microbial sciences as key drivers in the development of metabolomics research. Proc Natl Acad Sci 120:e2217383120. https://doi.org/10.1073/pnas.2217383120

167. Frigerio G, Moruzzi C, Mercadante R, Schymanski EL, Fustinoni S (2022) Development and Application of an LC-MS/MS Untargeted Exposomics Method with a Separated Pooled Quality Control Strategy. Molecules 27:2580. https://doi.org/10.3390/molecules27082580

168. Broeckling CD, Beger RD, Cheng LL, Cumeras R, Cuthbertson DJ, Dasari S, Davis WC, Dunn WB, Evans AM, Fernández-Ochoa A, Gika H, Goodacre R, Goodman KD, Gouveia GJ, Hsu P-C, Kirwan JA, Kodra D, Kuligowski J, Lan RS-L, Monge ME, Moussa LW, Nair SG, Reisdorph N, Sherrod SD, Ulmer Holland C, Vuckovic D, Yu L-R, Zhang B, Theodoridis G, Mosley JD (2023) Current Practices in LC-MS Untargeted Metabolomics: A Scoping Review on the Use of Pooled Quality Control Samples. Anal Chem 95:18645–18654. https://doi.org/10.1021/acs.analchem.3c02924

169. Caballero-Casero N, Belova L, Vervliet P, Antignac J-P, Castaño A, Debrauwer L, López ME, Huber C, Klanova J, Krauss M, Lommen A, Mol HGJ, Oberacher H, Pardo O, Price EJ, Reinstadler V, Vitale CM, van Nuijs ALN, Covaci A (2021) Towards harmonised criteria in quality assurance and quality control of suspect and non-target LC-HRMS analytical workflows for screening of emerging contaminants in human biomonitoring. TrAC Trends Anal Chem 136:116201. https://doi.org/10.1016/j.trac.2021.116201

170. Broadhurst D, Goodacre R, Reinke SN, Kuligowski J, Wilson ID, Lewis MR, Dunn WB (2018) Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. Metabolomics 14:72. https://doi.org/10.1007/s11306-018-1367-3

171. David A, Chaker J, Price EJ, Bessonneau V, Chetwynd AJ, Vitale CM, Klánová J, Walker DI, Antignac J-P, Barouki R, Miller GW (2021) Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives. Environ Int 156:106630. https://doi.org/10.1016/j.envint.2021.106630

172. Viant MR, Ebbels TMD, Beger RD, Ekman DR, Epps DJT, Kamp H, Leonards PEG, Loizou GD, MacRae JI, van Ravenzwaay B, Rocca-Serra P, Salek RM, Walk T, Weber RJM (2019) Use cases, best practice and reporting standards for metabolomics in regulatory toxicology. Nat Commun 10:3041. https://doi.org/10.1038/s41467-019-10900-y

173. Kirwan JA, Gika H, Beger RD, Bearden D, Dunn WB, Goodacre R, Theodoridis G, Witting M, Yu L-R, Wilson ID (2022) Quality assurance and quality control reporting in untargeted metabolic phenotyping: mQACC recommendations for analytical quality management. Metabolomics 18:70. https://doi.org/10.1007/s11306-022-01926-3

174. Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, O'Donovan C (2020) MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Res 48:D440–D444. https://doi.org/10.1093/nar/gkz1019

175. Leao TF, Clark CM, Bauermeister A, Elijah EO, Gentry EC, Husband M, Oliveira MF, Bandeira N, Wang M, Dorrestein PC (2021) Quick-start infrastructure for untargeted metabolomics analysis in GNPS. Nat Metab 3:880–882. https://doi.org/10.1038/s42255-021-00429-0

176. Chetnik K, Petrick L, Pandey G (2020) MetaClean: a machine learning-based classifier for reduced false positive peak detection in untargeted LC–MS metabolomics data. Metabolomics 16:117. https://doi.org/10.1007/s11306-020-01738-3

177. Mosley JD, Schock TB, Beecher CW, Dunn WB, Kuligowski J, Lewis MR, Theodoridis G, Ulmer Holland CZ, Vuckovic D, Wilson ID, Zanetti KA (2024) Establishing a framework for best practices for quality assurance and quality control in untargeted metabolomics. Metabolomics 20:20. https://doi.org/10.1007/s11306-023-02080-0

178. Petrick LM, Shomron N (2022) AI/ML-driven advances in untargeted metabolomics and exposomics for biomedical applications. Cell Rep Phys Sci 3:100978. https://doi.org/10.1016/j.xcrp.2022.100978

179. Wan M, Simonin EM, Johnson MM, Zhang X, Lin X, Gao P, Patel CJ, Yousuf A, Snyder MP, Hong X, Wang X, Sampath V, Nadeau KC (2025) Exposomics: a review of methodologies, applications, and future directions in molecular medicine. EMBO Mol Med 17:599–608. https://doi.org/10.1038/s44321-025-00191-w

180. Rappaport SM (2016) Genetic Factors Are Not the Major Causes of Chronic Diseases. PLoS ONE 11:e0154387. https://doi.org/10.1371/journal.pone.0154387

181. Fuller R, Landrigan PJ, Balakrishnan K, Bathan G, Bose-O'Reilly S, Brauer M, Caravanos J, Chiles T, Cohen A, Corra L, Cropper M, Ferraro G, Hanna J, Hanrahan D, Hu H, Hunter D, Janata G, Kupka R, Lanphear B, Lichtveld M, Martin K, Mustapha A, Sanchez-Triana E, Sandilya K, Schaefli L, Shaw J, Seddon J, Suk W, Téllez-Rojo MM, Yan C (2022) Pollution and health: a progress update. Lancet Planet Health 6:e535–e547. https://doi.org/10.1016/S2542-5196(22)00090-0

182. Wan M, Simonin EM, Johnson MM, Zhang X, Lin X, Gao P, Patel CJ, Yousuf A, Snyder MP, Hong X, Wang X, Sampath V, Nadeau KC (2025) Exposomics: a review of methodologies, applications, and future directions in molecular medicine. EMBO Mol Med 17:599–608. https://doi.org/10.1038/s44321-025-00191-w

183. Tisler S, Savvidou P, Jørgensen MB, Castro M, Christensen JH (2023) Supercritical Fluid Chromatography Coupled to High-Resolution Mass Spectrometry Reveals Persistent Mobile Organic Compounds with Unknown Toxicity in Wastewater Effluents. Environ Sci Technol 57:9287–9297. https://doi.org/10.1021/acs.est.3c00120

184. Wild CP (2025) The exposome at twenty: a personal account. Exposome 5:osaf003. https://doi.org/10.1093/exposome/osaf003

185. Maitre L, Bustamante M, Hernández-Ferrer C, Thiel D, Lau C-HE, Siskos AP, Vives-Usano M, Ruiz-Arenas C, Pelegrí-Sisó D, Robinson O, Mason D, Wright J, Cadiou S, Slama R, Heude B, Casas M, Sunyer J, Papadopoulou EZ, Gutzkow KB,

Andrusaityte S, Grazuleviciene R, Vafeiadi M, Chatzi L, Sakhi AK, Thomsen C, Tamayo I, Nieuwenhuijsen M, Urquiza J, Borràs E, Sabidó E, Quintela I, Carracedo Á, Estivill X, Coen M, González JR, Keun HC, Vrijheid M (2022) Multi-omics signatures of the human early life exposome. Nat Commun 13:7024. https://doi.org/10.1038/s41467-022-34422-2

186. Argentieri MA, Amin N, Nevado-Holgado AJ, Sproviero W, Collister JA, Keestra SM, Kuilman MM, Ginos BNR, Ghanbari M, Doherty A, Hunter DJ, Alvergne A, van Duijn CM (2025) Integrating the environmental and genetic architectures of aging and mortality. Nat Med 31:1016–1025. https://doi.org/10.1038/s41591-024-03483-9

Figure captions

Figure 1. The exposome concept and how the specific external, general external and internal exposome contribute towards health impacts. Modified from [9, 10].

Figure 2. A) The omics cascade from genome onwards, adapted from [11, 12]. B) The differences in chemical complexity of the different omics, adapted from [16, 17]. Note that the colours represent the different “omes” (genes, proteins, metabolites).

Figure 3. The chemicals part of the metabolome, exposome and the overlap. Adapted from [2, 20]. Note that gut microbiota is illustrated as a major microbial contributor to metabolic processes, however other microbiota such as saliva, nasal and skin, among others, can also contribute.

Figure 4. Common workflow steps to investigate the human chemical exposome.

Figure 5. Human life timeline (top) and proposed longitudinal study for an exposomics study where both environmental and biological samples are collected. AD; Alzheimer's disease, MCI; Mild Cognitive Impairment, PD; Parkinson's disease, RBD; REM- sleep behavior disorder. Note that while neurodegenerative diseases were chosen for illustrative purposes, this conceptual framework can be applied to other diseases such as cancer.

Figure 6. Separation analytical methods and their applicability range based on the polarity of the chemicals are displayed on the top part of the figure, while the different ionization techniques and their applicability range are displayed on the bottom. Examples of potentially endogenous (blue) and exogenous (gray) compounds are displayed. Adapted from Zeki et. al. [39] and Hollender et al. [15]. Abbreviations: GC, gas chromatography; HILIC, Hydrophilic interaction chromatography; LC, liquid chromatography; RP, Reversed Phase; EI, electron ionization; ESI, electrospray ionization; APCI, atmospheric pressure chemical ionization; APPI, atmospheric pressure photoionization; VOCs, Volatile Organic Compounds; PFAS, Perfluoroalkyl and Polyfluoroalkyl Substances; PAHs, Polycyclic Aromatic Hydrocarbons; IC, Ion Exchange; CE, Capillary Electrophoresis.

Figure 7. Concentration range of the different components of the human metabolome and internal chemical exposome as well as coverage by LC-ESI-HRMS. Adapted from [2, 17].

Figure 8. Data preprocessing steps for LC-HRMS raw data. First, data is centroided and noise is removed. Next, EICs are generated and a peak-picking algorithm is applied to detect true peaks. Finally, peaks are grouped across samples, and retention time alignment is performed. After that, a gap filling step can be performed to reduce the number of missing values. Adapted from [54, 60, 64]. Abbreviations: RT, Retention Time; m/z, mass-to-charge ratio.

Figure 9. A) Generic computational workflow for target and non-target exposomics studies [15, 35, 79, 80]. B) Identification confidence levels by Schymanski et al. [81] (left), and proposed minimum data requirements for level 2a and 3a annotations using MS-DIAL and patRoon software (right). Similar approaches could be done with other software such as **MZmine [55] and LipidMatch [82].**

Figure 10. Exemplified workflow of MS2 library search software [47]. For a given experimental MS1, first the precursor filter is applied to remove all the candidates outside the tolerance window (e.g., 0.01 Da). Subsequently, the similarity algorithm ranks the experimental MS2 spectra against the remaining library spectra candidates (four in this example) and calculates a similarity score. Note that the adducts shown ([M+H]+ and

[M+Na]+) are illustrative examples and other adducts (e.g., [M+K]+, [M+NH4]+ in positive mode, [M-H]-, [M+Cl]- in negative mode) can occur based on the matrix and acquisition settings, among others. Abbreviations: RT, retention time; m/z, mass-to-charge ratio.

Figure 11. A) Overlaid dot plot showing the coverage of the Blood Exposome Database, CompTox, the Human Metabolome Database (HMDB) and PubChemLite for Exposomics (PCL). B) Overlaid dot plot of PubChemLite (PCL) displaying the coverage of the compounds with LC-MS, GC-MS as well as CCS information. C) Overlaid plot of PCL displaying the compounds with Pathway information (Biopathway), associated disorders and diseases (DisorderDisease), agrochemical information and drug and medication information (DrugMedicInfo). R code for data visualization can be found in the GitLab repository (https://gitlab.com/uniluxembourg/lcsb/eci/exposomics-plots) [123].

Figure 12. Multivariate statistical approaches applied in exposomics studies. A) PCA scores plot of where all QC samples cluster tightly near the origin, which is indicative of quality data, affirming that the instrument variation was effectively corrected. PCA is widely used as a QA/QC tool to detect outliers and assess batch effects. B) PLS-DA scores plot of the same dataset, included here to illustrate how supervised methods maximize the differences between predefined groups. Unlike PCA, PLS-DA is not used for QA/QC but is applied in downstream analyses such as classification, biomarker discovery, and hypothesis testing. Further details regarding QA/QC measures are discussed in Broadhurst et al. [142].

Figure 13. The metabolome and exposome timeline. Only the peak heights of the Metabolome (1998), Exposome (2014) and last peak (?) are intentionally emphasized for significance. The sizes of the other peaks were adjusted for aesthetic reasons and do not reflect their importance. A representative selection of tools is displayed here for illustrative purposes, and the authors acknowledge the presence of many other numerous tools and initiatives that have emerged in recent years. Adapted from [166]. Abbreviations: Artificial Intelligence; AI.

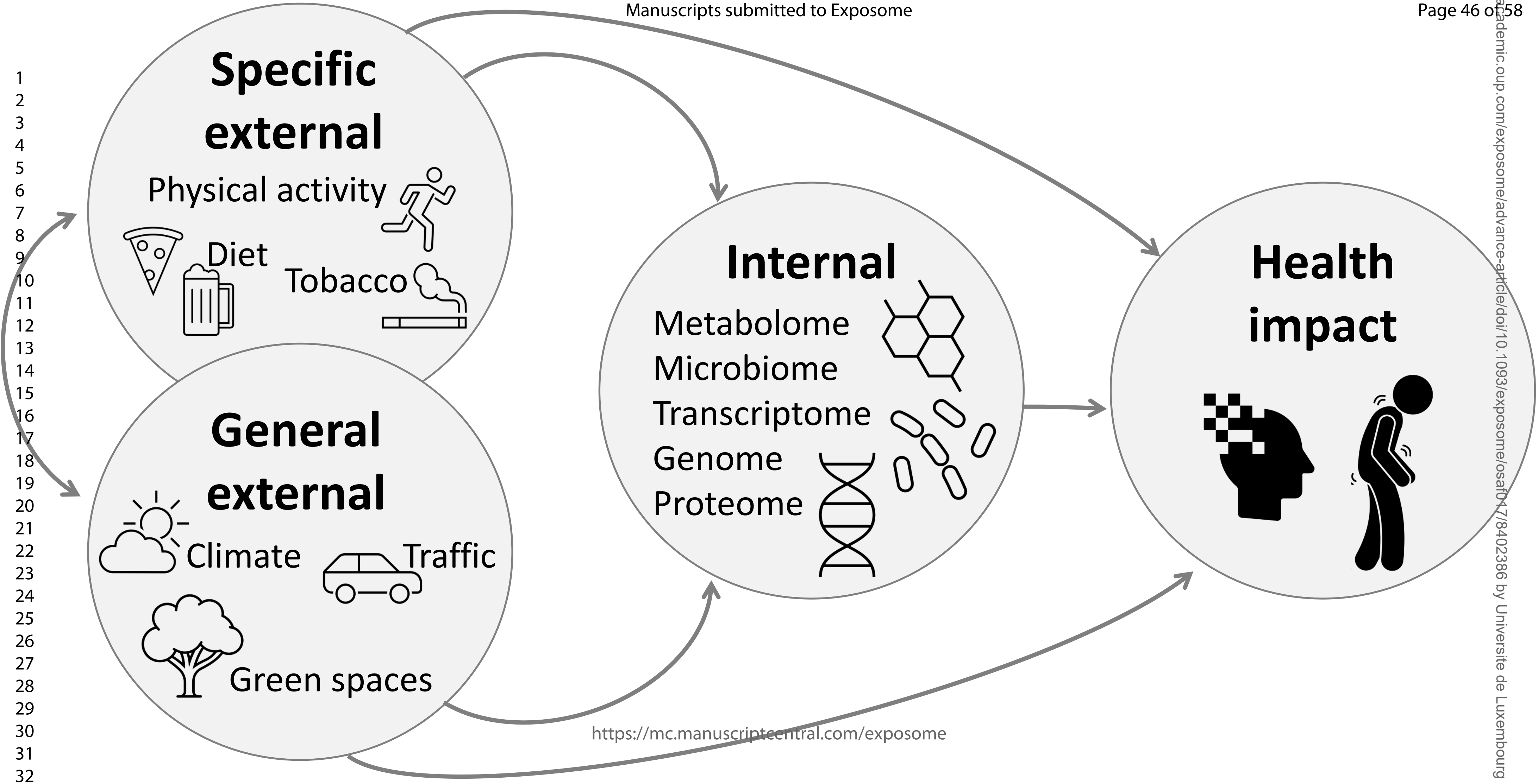
Tables

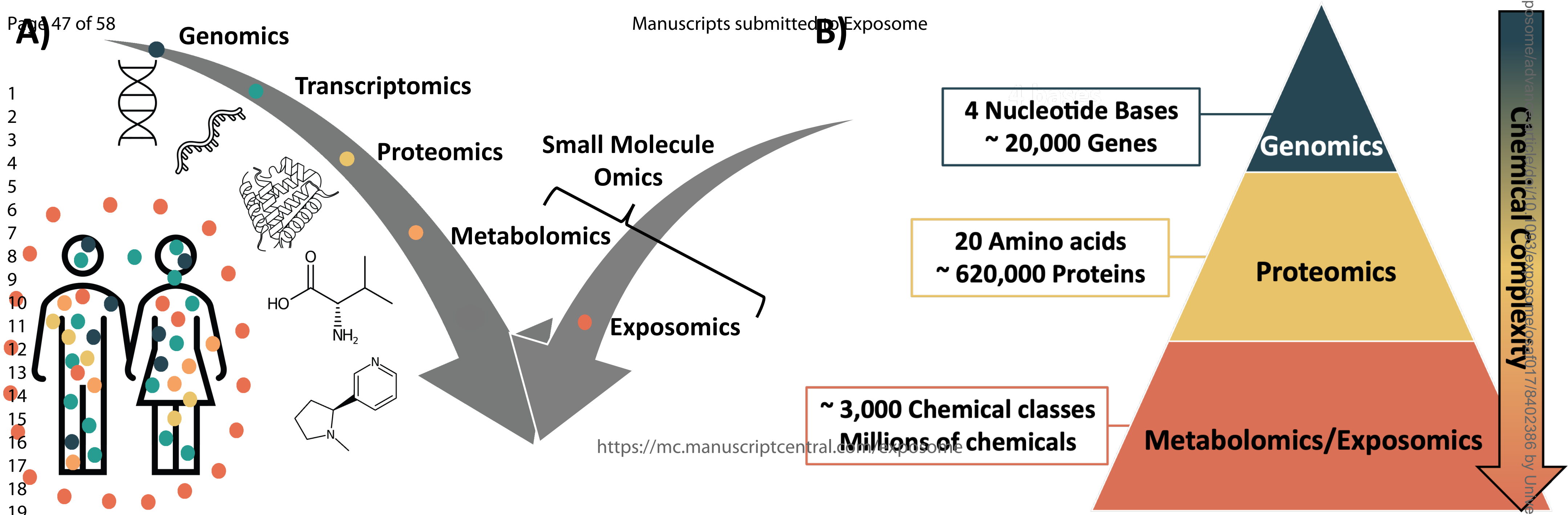
Table 1. *Summary of some recommended QA/QC procedures for non-target-HRMS exposomics and metabolomics.*

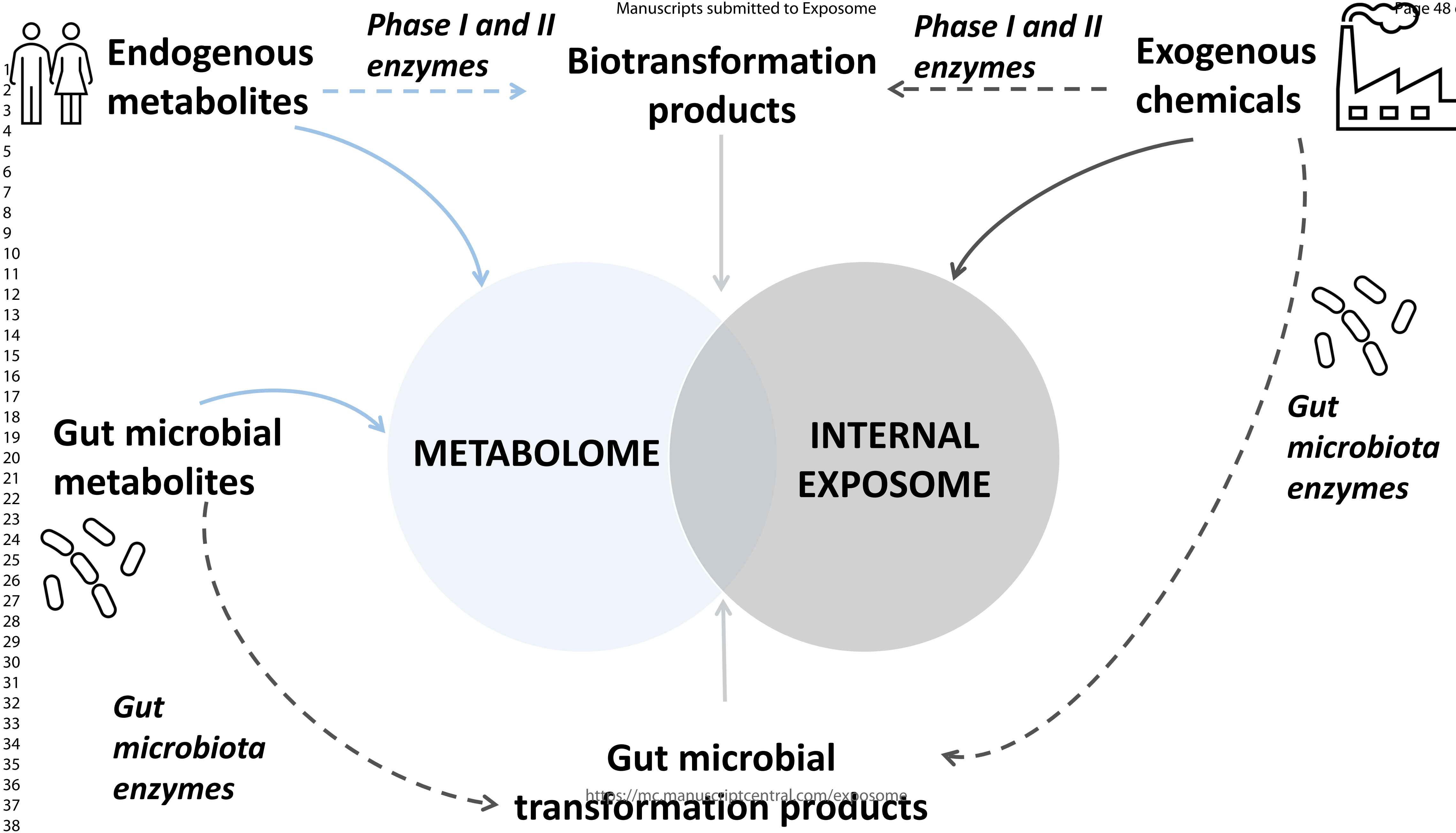
Workflow step	**QA/QC procedures**
Sample collection	• Develop detailed Standard Operational Protocols (SOPs) for sampling and storage. • Careful selection of appropriate sampling materials (e.g., avoiding tubes with components such as phthalates and Polyethylene Glycol (PEG) that may interfere with the analysis). • Train personnel performing the sample collection [169, 170].

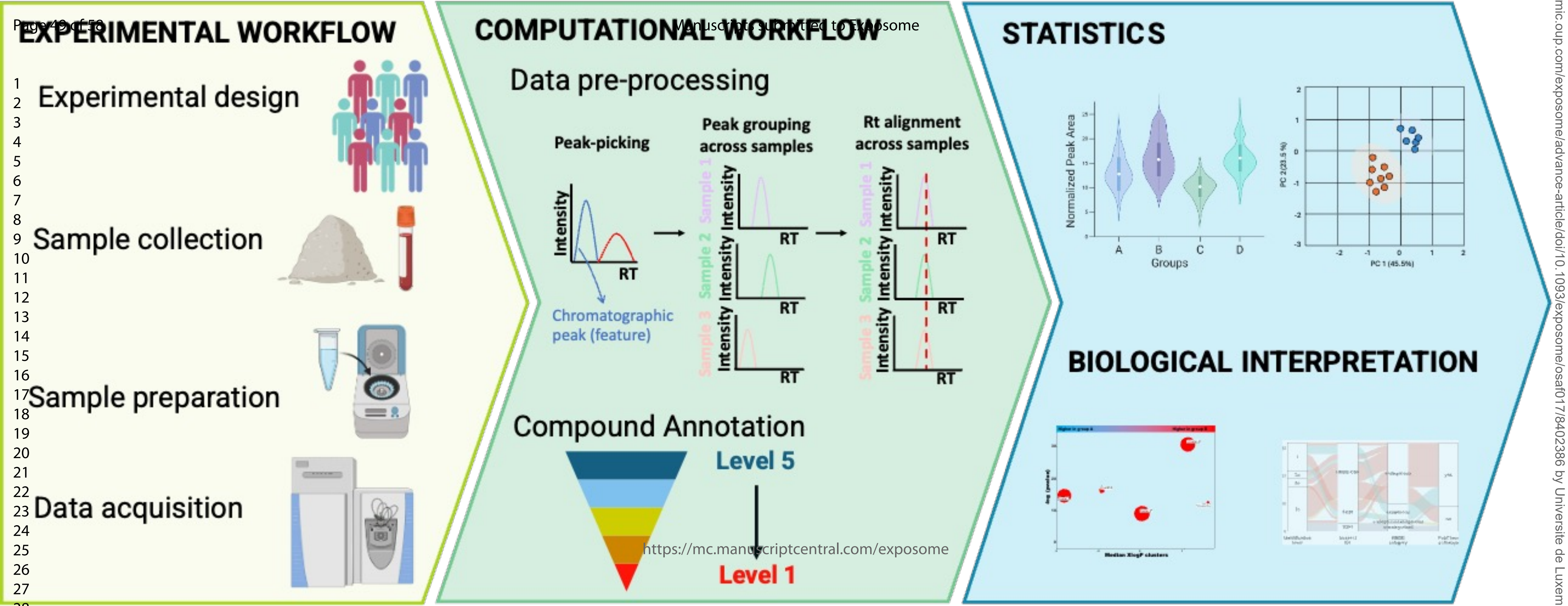
	• Define specific acceptance/rejection criteria for samples upon arrival at the laboratory [169]. • Report collection method, storage temperature and number of freeze-thaw cycles, if applicable.
Sample preparation	• Use Internal Standards (IS). Ideally a mixture of multiple IS covering the range of the chemical space to be investigated should be added to each sample, at predetermined concentrations [15, 142]. • Use pooled QC samples, prepared by taking a small aliquot of each study sample and mixing it into homogenous pooled sample. Pooled QC samples are used to condition the analytical platform [170], assess the analytical performance [171], correct batch effects [168, 170, 171], support metabolite identification [168], and filter low quality data [170], although this may result in the loss of low abundant features such as pollutants. Thus, it is not recommended for exposomics studies, where low-abundant features may be relevant. Report how pooled samples were prepared and used. • Include blank samples to prevent false positives. Different blanks can be prepared such as extraction blanks (or process banks) and system suitability blanks (instrument blanks). Further details can be found in the 2023 NORMAN guidance [15]. • Report sample preparation protocol including the reconstitution solvent and volume, IS employed and concentrations [172].
Data acquisition	• Ensure instrument is calibrated before starting the analysis [172]. • Carefully plan the injection order to ensure the quality and precision of the analysis. A graphical example of a sequence is given in [142]. System suitability blanks (e.g., MilliQ water) are injected to check that the instrument is working properly. System conditioning pooled QC samples can be injected to equilibrate the instrument before sample analysis. Injecting pooled QC samples throughout the sequence (e.g., every 5 or 10 samples) helps measure the precision of the system (e.g., stable retention times), correct for systematic bias, and support the later data pre-processing (e.g., feature filtering) [15, 142]. Randomization of all the samples across the sequence reduces systematic errors due to carryover [15]. • Report the instrument configuration including the LC system, ionization source (ESI, APPI…), and MS analyzer (e.g., Orbitrap). • Report the LC-HRMS method details including mobile phase, gradient, flow rate, column temperature, sample volume injected, acquisition mode (e.g., DDA), polarity (e.g., positive), and *m/z* range, among others. Further details can be found in Viant et al. [172].
Data pre-processing	• Following existing reporting guidelines for data pre-processing such as the MEtabolomics standaRds Initiative in Toxicology

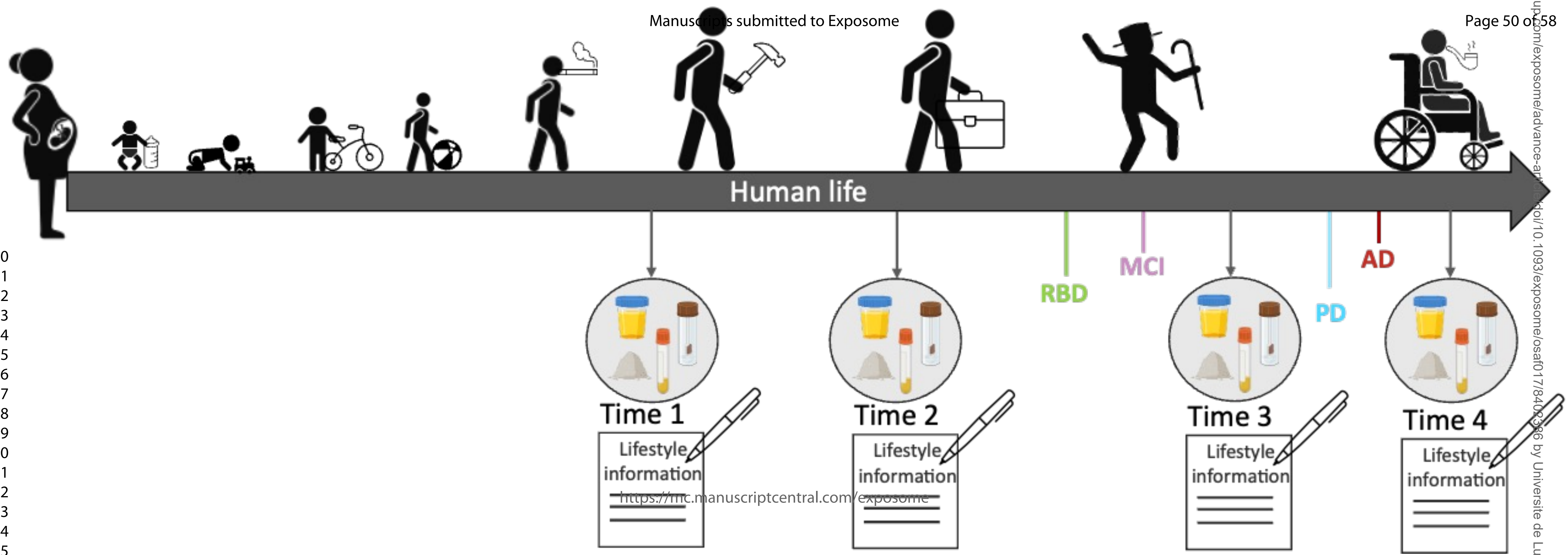
	(MERIT) [172] and the Metabolomics Quality Assurance and Quality Control Consortium (mQACC) [173]. • Share raw data in public repositories (e.g. MetaboLights [174] and GNPS [175]) to ensure data reproducibility and reusability [54]. • Use open software computational workflows integrating data pre-processing, compound annotation, and statistical analysis, to enhance reproducibility by minimizing manual data curation [54]. • Use benchmark datasets to evaluate pre-processing algorithms [54]. Packages such as IPO [69] and Meta-Clean [176] help optimize data pre-processing parameters. • Consider data pre-processing QA/QC guidelines proposed in [54].
Compound Annotation	• Assigning identification levels to the detected features is not trivial process and requires appropriate QA/QC procedures. Document the confidence level assigned to each feature and the annotation system employed [72]. • Manual curation of all the annotations, coupled with the provision of evidence supporting the confidence level (e.g., m/z, retention time, and MS2) is important for mitigating false positives and enhancing the reliability of the results [177] . • Report software employed, and parameters (e.g., mass tolerance) used for compound annotation. The mass spectral library, suspect lists, and/or chemical databases utilized should be clearly stated, along with their respective versions [15]. Spectra included in MS2 libraries should be curated by filtering, noise removal, and recalibration to ensure the quality of the reference spectra [72].
Statistical analysis	• Report all statistical methods transparently; the use of open source-software are preferred. • In-house generated code, such as R scripts, should be shared in public repositories (e.g., GitHub) fostering transparency and reproducibility [172]. • Univariate statistics: report median and Relative Standard Deviation (RSD) of each feature across the pooled QC samples. • Multivariate statistics: performing and reporting PCA is recommended to confirm that the pooled QC samples cluster tightly, indicating high quality data.

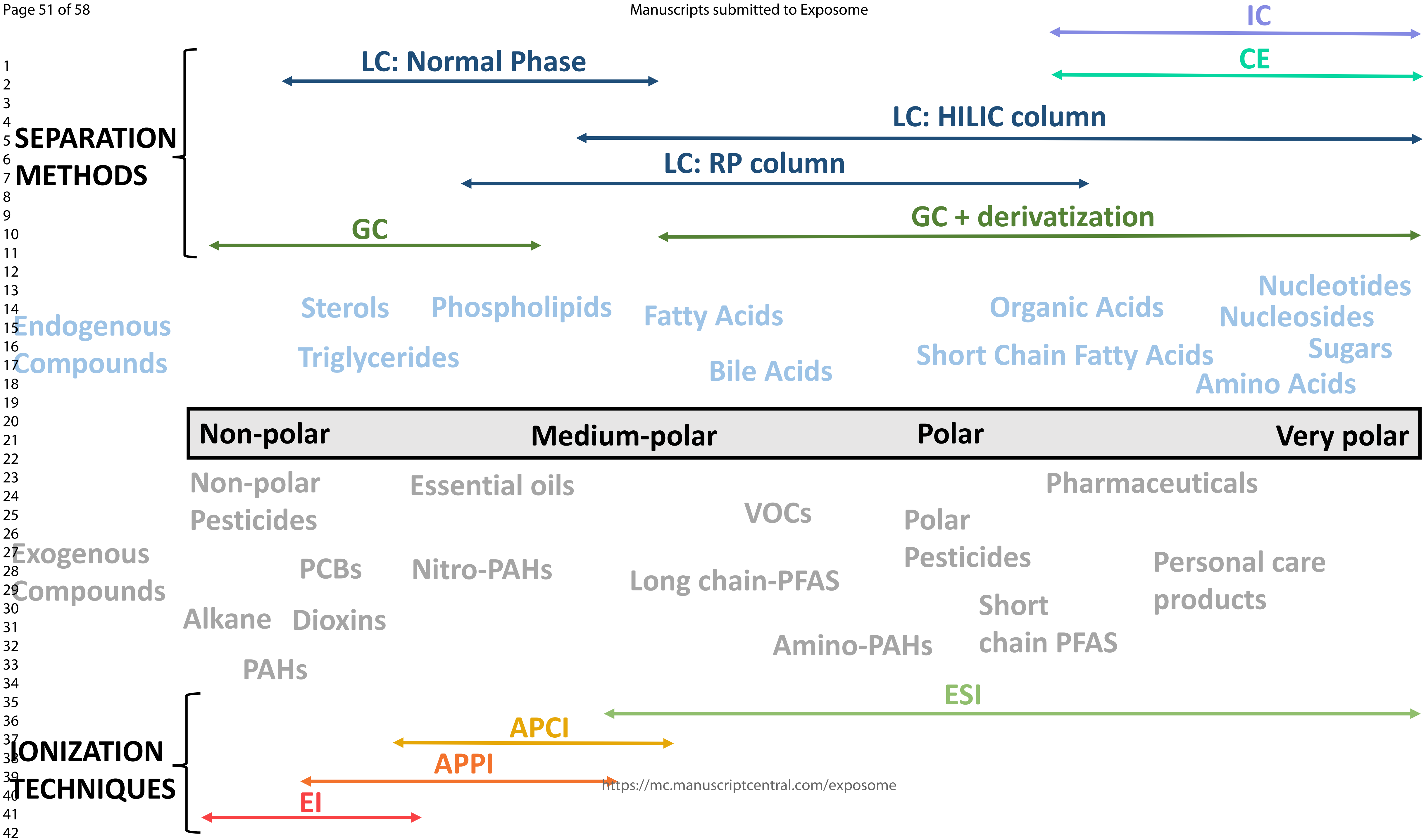

Specific external
Physical activity
Diet
Tobacco
General external
Climate
Traffic
Green spaces
Internal
Metabolome
Microbiome
Transcriptome
Genome
Proteome
Health impact


A)
Genomics
Transcriptomics
Proteomics
Small Molecule Omics
Metabolomics
Exposomics
O
HO
NH_2
N
N
B)
4 Nucleotide Bases
~ 20,000 Genes
Genomics
20 Amino acids
~ 620,000 Proteins
Proteomics
~ 3,000 Chemical classes
Millions of chemicals
Metabolomics/Exposomics
Chemical Complexity


Endogenous metabolites
Phase I and II enzymes
Biotransformation products
Phase I and II enzymes
Exogenous chemicals
METABOLOME
INTERNAL EXPOSOME
Gut microbial metabolites
Gut microbiota enzymes
Gut microbiota enzymes
Gut microbial transformation products


EXPERIMENTAL WORKFLOW
Experimental design
Sample collection
Sample preparation
Data acquisition
COMPUTATIONAL WORKFLOW
Data pre-processing
Peak-picking
Intensity
RT
Chromatographic peak (feature)
Peak grouping across samples
Rt alignment across samples
Sample 1
Sample 2
Sample 3
Intensity
RT
Compound Annotation
Level 5
Level 1
STATISTICS
Normalized Peak Area
A
B
C
D
Groups
PC 2 (23.5 %)
PC 1 (45.5%)
BIOLOGICAL INTERPRETATION
Median XlogP clusters


Human life
Time 1
Lifestyle information
Time 2
Lifestyle information
RBD
MCI
Time 3
Lifestyle information
PD
AD
Time 4
Lifestyle information



SEPARATION METHODS

IC

CE

LC: Normal Phase

LC: HILIC column

LC: RP column

GC

GC + derivatization

Endogenous Compounds

Sterols

Phospholipids

Triglycerides

Fatty Acids

Bile Acids

Organic Acids

Short Chain Fatty Acids

Nucleotides

Nucleosides

Sugars

Amino Acids

Non-polar | Medium-polar | Polar | Very polar

Exogenous Compounds

Non-polar Pesticides

Essential oils

PCBs

Nitro-PAHs

Alkane

Dioxins

PAHs

VOCs

Long chain-PFAS

Amino-PAHs

Polar Pesticides

Short chain PFAS

Pharmaceuticals

Personal care products

IONIZATION TECHNIQUES

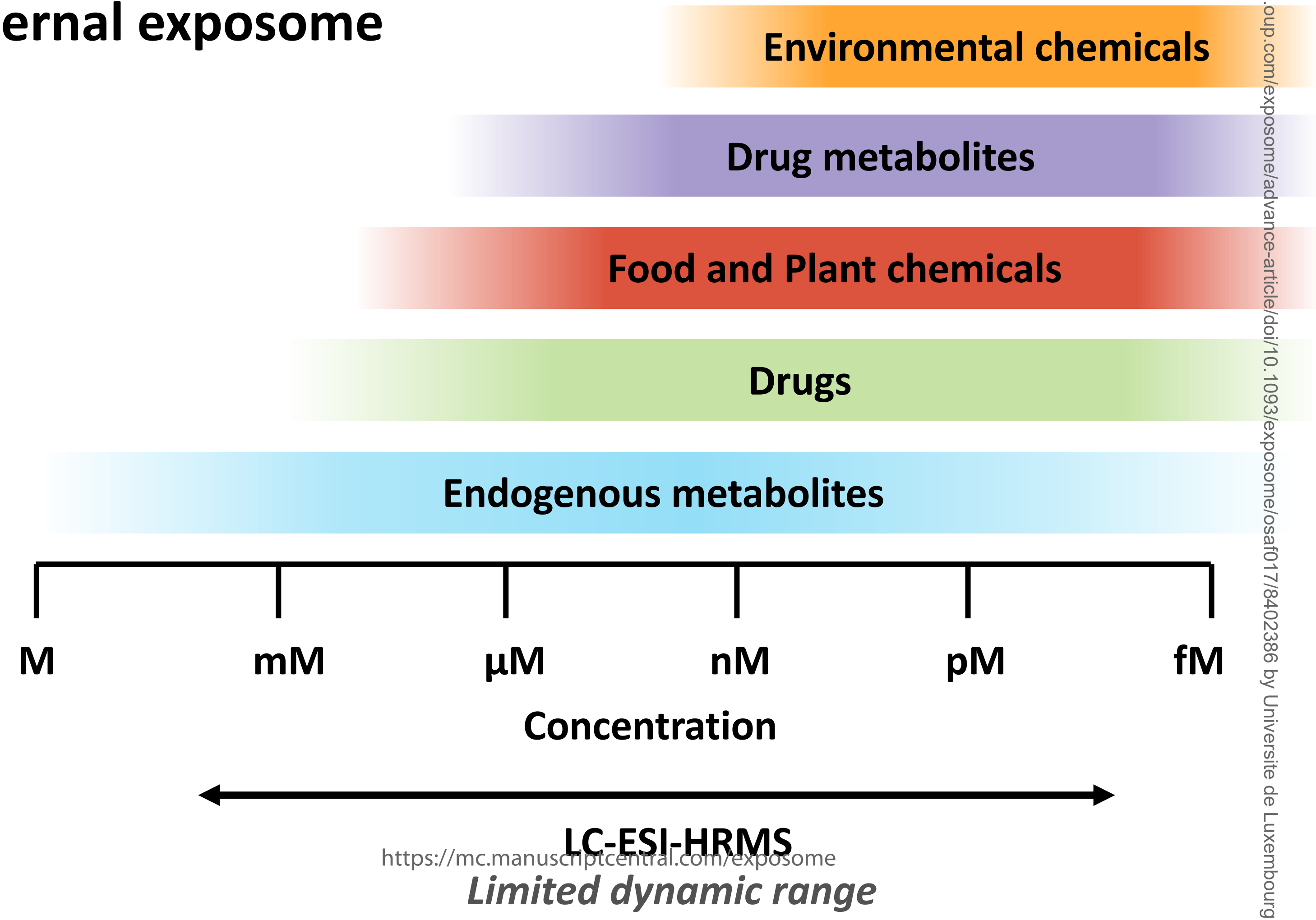
ESI

APCI

APPI

EI

Human metabolome and internal exposome



Environmental chemicals

Drug metabolites

Food and Plant chemicals

Drugs

Endogenous metabolites

M

mM

µM

nM

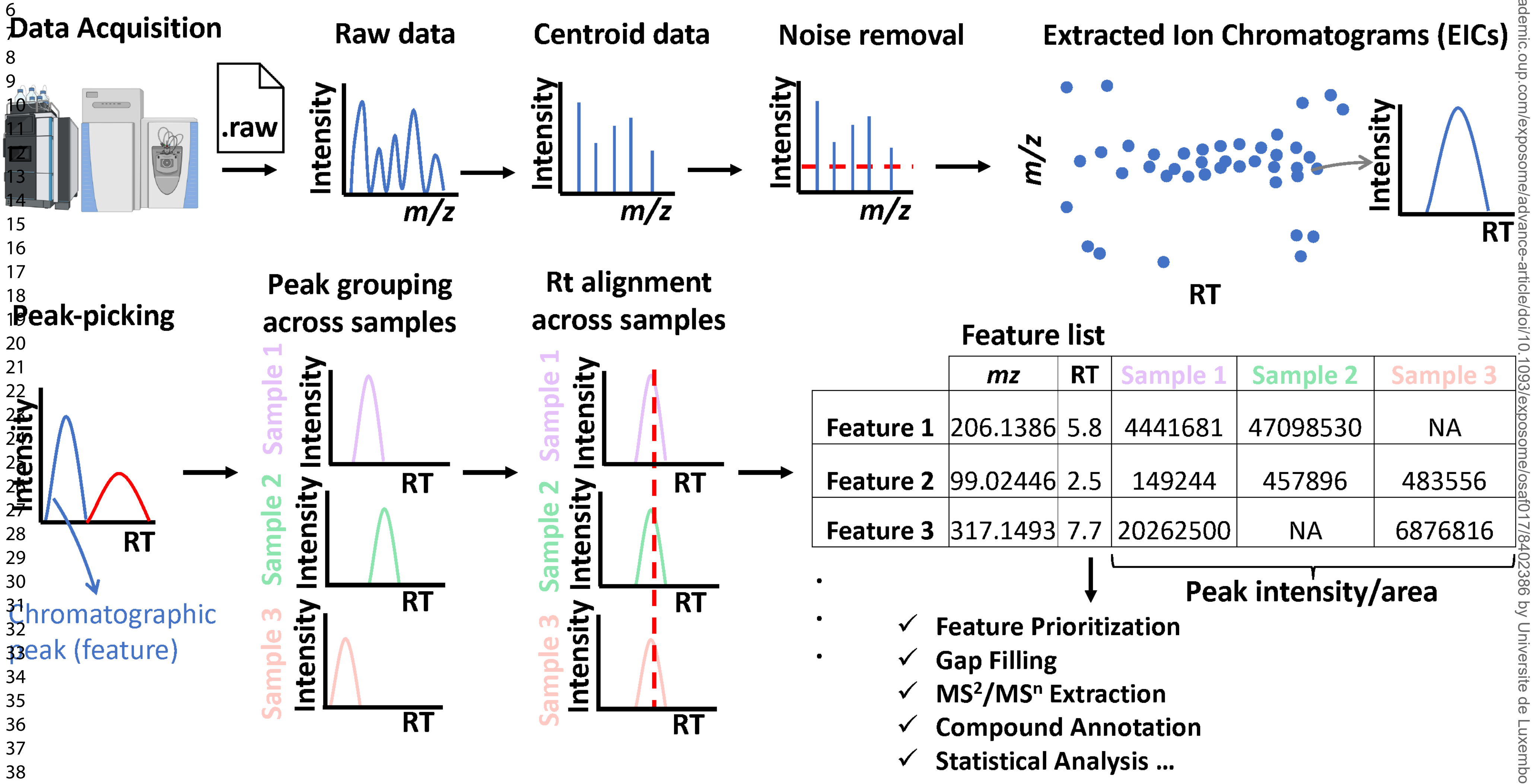
pM

fM

Concentration

LC-ESI-HRMS

Limited dynamic range

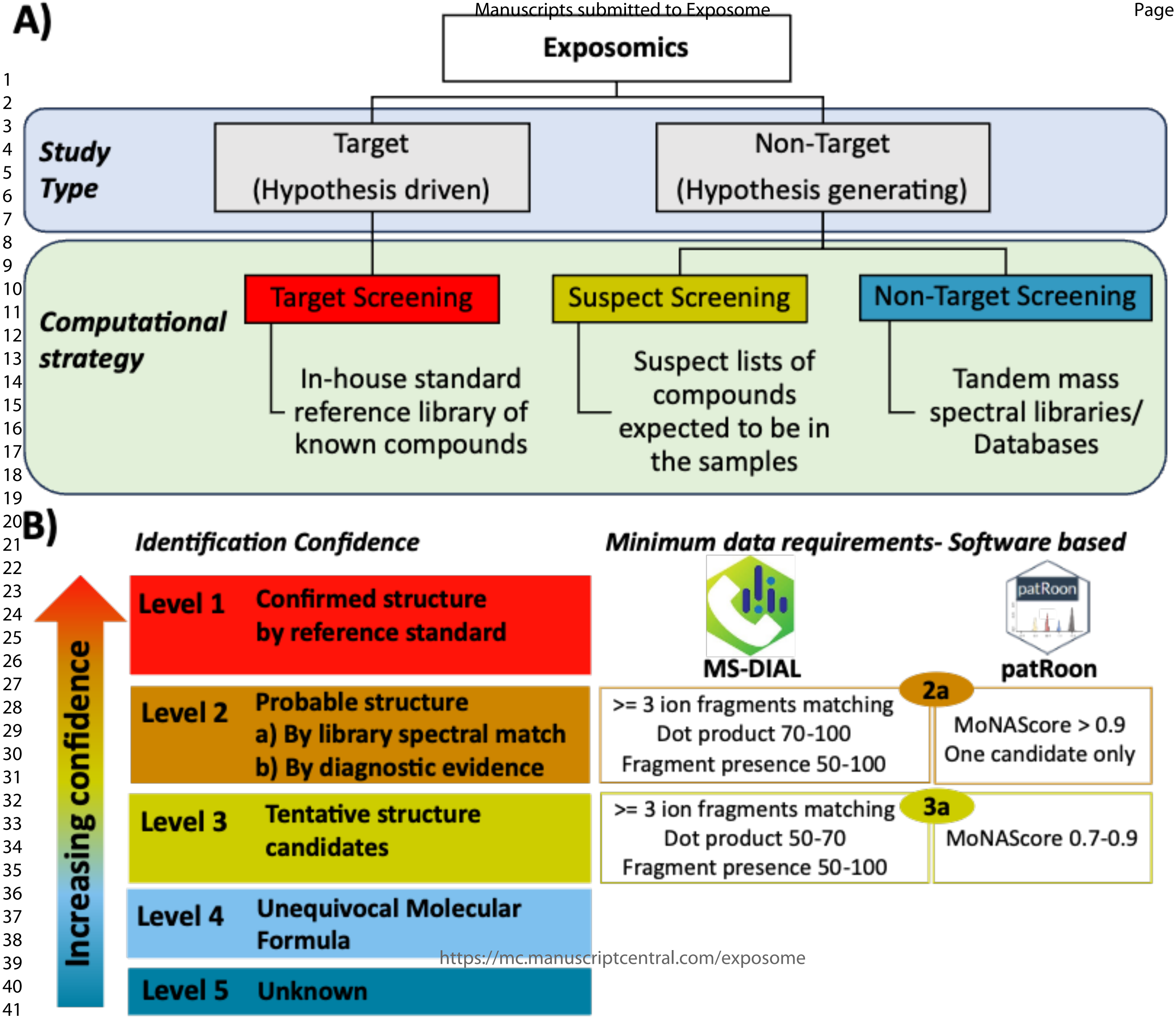


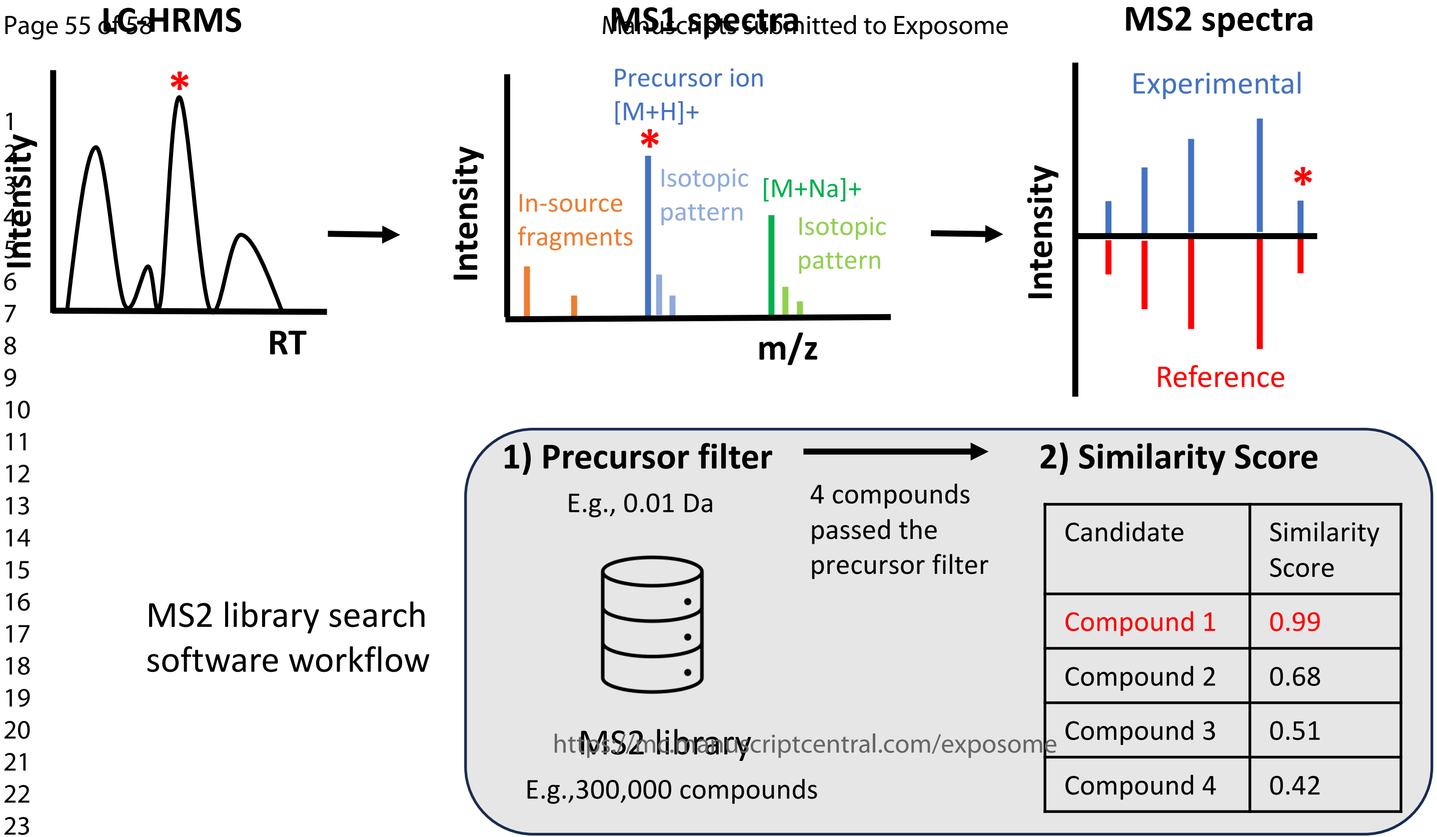
Feature list

	mz	RT	Sample 1	Sample 2	Sample 3
Feature 1	206.1386	5.8	4441681	47098530	NA
Feature 2	99.02446	2.5	149244	457896	483556
Feature 3	317.1493	7.7	20262500	NA	6876816

Peak intensity/area

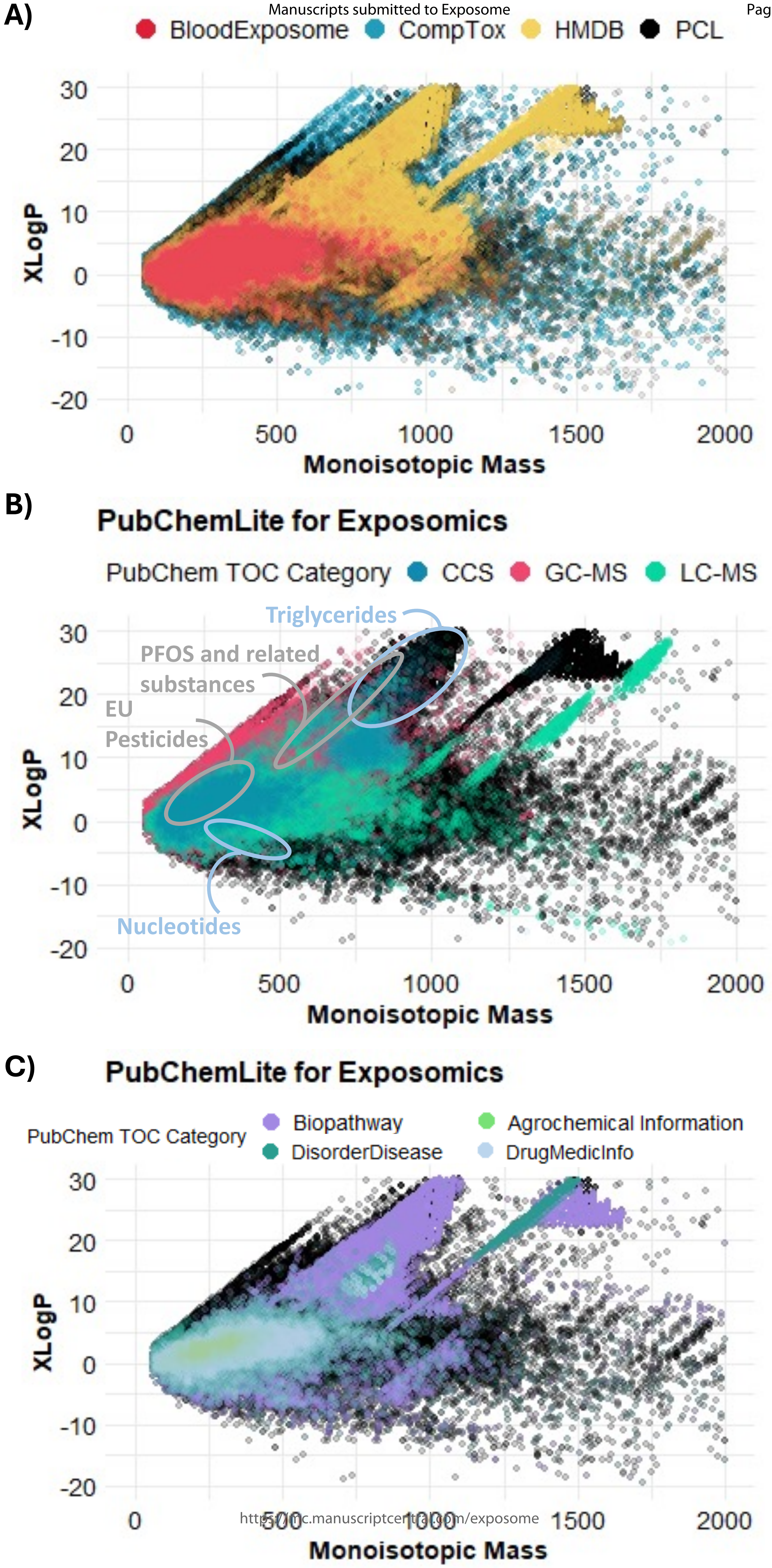
✓ Feature Prioritization
✓ Gap Filling
✓ MS^2/MS^n Extraction
✓ Compound Annotation
✓ Statistical Analysis …

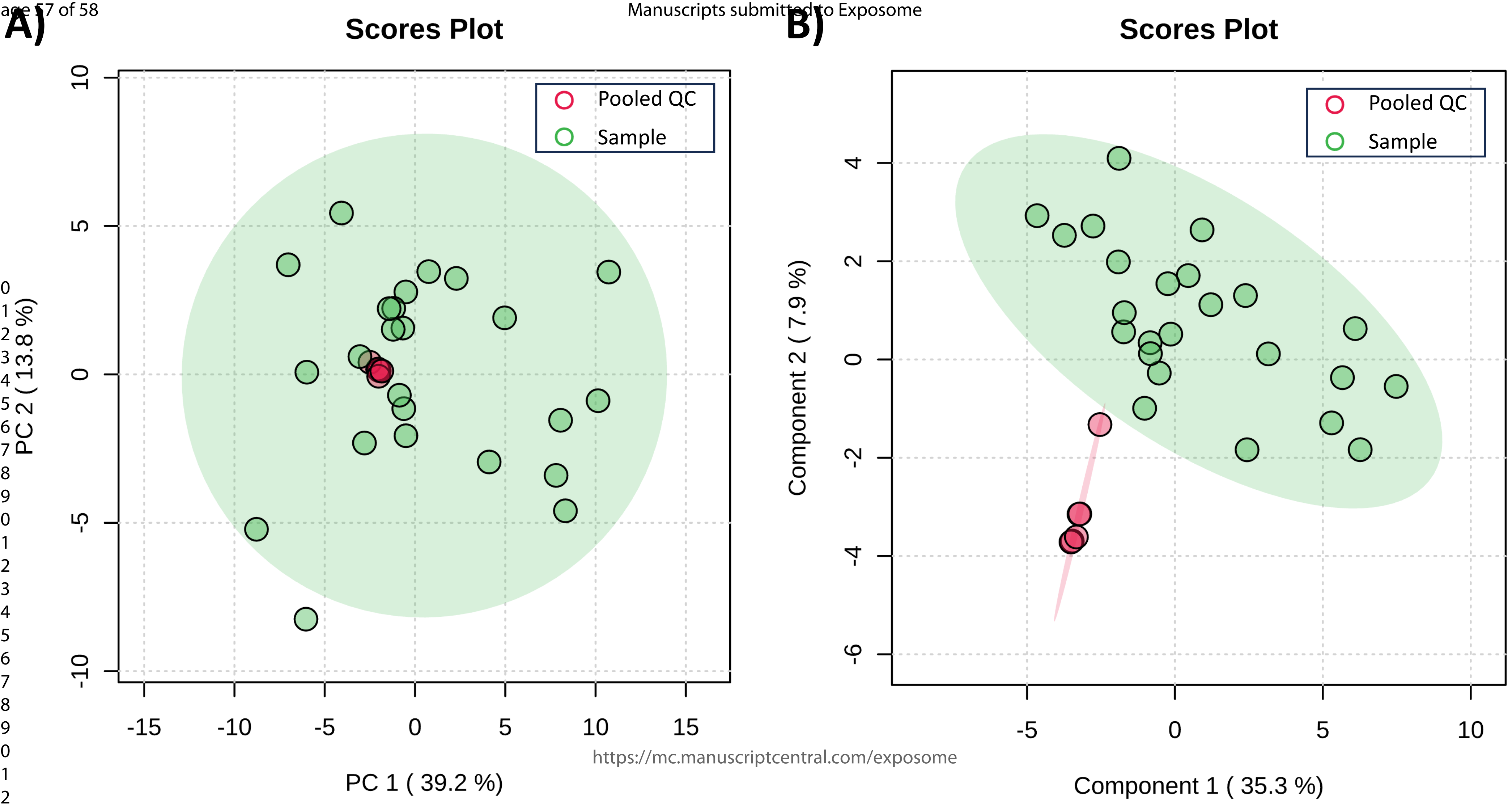

A)
Exposomics
Study Type
Target
(Hypothesis driven)
Non-Target
(Hypothesis generating)
Computational strategy
Target Screening
Suspect Screening
Non-Target Screening
In-house standard reference library of known compounds
Suspect lists of compounds expected to be in the samples
Tandem mass spectral libraries/ Databases
B)
Identification Confidence
Minimum data requirements- Software based
Increasing confidence
Level 1
Confirmed structure by reference standard
Level 2
Probable structure
a) By library spectral match
b) By diagnostic evidence
Level 3
Tentative structure candidates
Level 4
Unequivocal Molecular Formula
Level 5
Unknown
MS-DIAL
patRoon
patRoon
2a
>= 3 ion fragments matching
Dot product 70-100
Fragment presence 50-100
MoNAScore > 0.9
One candidate only
3a
>= 3 ion fragments matching
Dot product 50-70
Fragment presence 50-100
MoNAScore 0.7-0.9

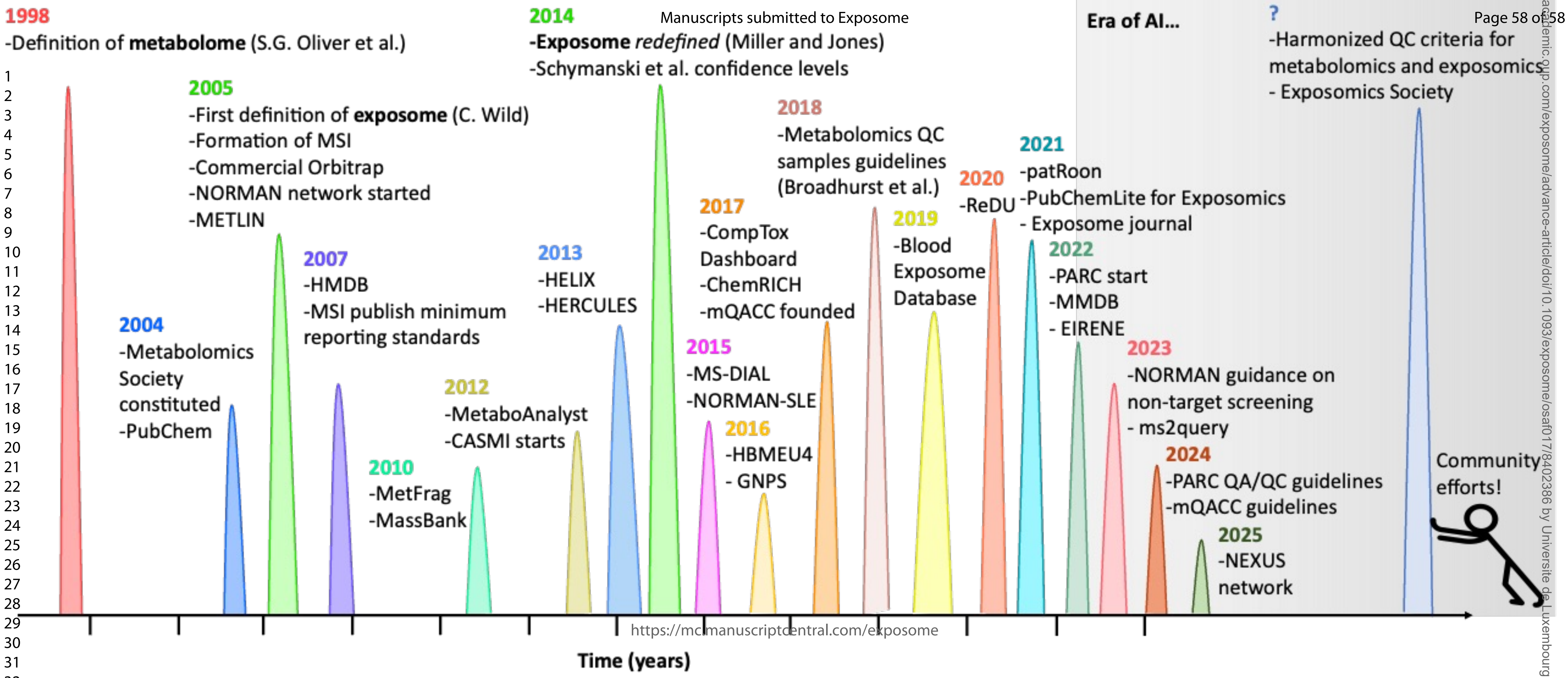


Candidate	Similarity Score
Compound 1	0.99
Compound 2	0.68
Compound 3	0.51
Compound 4	0.42

MS2 library search software workflow


A)
● BloodExposome ● CompTox ● HMDB ● PCL
XLogP
30
20
10
0
-10
-20
0
500
1000
1500
2000
Monoisotopic Mass
B)
PubChemLite for Exposomics
PubChem TOC Category ● CCS ● GC-MS ● LC-MS
Triglycerides
PFOS and related substances
EU Pesticides
Nucleotides
XLogP
30
20
10
0
-10
-20
0
500
1000
1500
2000
Monoisotopic Mass
C)
PubChemLite for Exposomics
PubChem TOC Category
● Biopathway
● DisorderDisease
● Agrochemical Information
● DrugMedicInfo
XLogP
30
20
10
0
-10
-20
0
500
1000
1500
2000
Monoisotopic Mass


A)
Scores Plot
Pooled QC
Sample
PC 2 (13.8 %)
PC 1 (39.2 %)
10
5
0
-5
-10
-15
-10
-5
0
5
10
15
B)
Scores Plot
Pooled QC
Sample
Component 2 (7.9 %)
Component 1 (35.3 %)
4
2
0
-2
-4
-6
-5
0
5
10



1998
-Definition of **metabolome** (S.G. Oliver et al.)

2004
-Metabolomics Society constituted
-PubChem

2005
-First definition of **exposome** (C. Wild)
-Formation of MSI
-Commercial Orbitrap
-NORMAN network started
-METLIN

2007
-HMDB
-MSI publish minimum reporting standards

2010
-MetFrag
-MassBank

2012
-MetaboAnalyst
-CASMI starts

2013
-HELIX
-HERCULES

2014
-Exposome *redefined* (Miller and Jones)
-Schymanski et al. confidence levels

2015
-MS-DIAL
-NORMAN-SLE

2016
-HBMEU4
- GNPS

2017
-CompTox Dashboard
-ChemRICH
-mQACC founded

2018
-Metabolomics QC samples guidelines (Broadhurst et al.)

2019
-Blood Exposome Database

2020
-ReDU

2021
-patRoon
-PubChemLite for Exposomics
- Exposome journal

2022
-PARC start
-MMDB
- EIRENE

2023
-NORMAN guidance on non-target screening
- ms2query

2024
-PARC QA/QC guidelines
-mQACC guidelines

2025
-NEXUS network

Era of AI...

?
-Harmonized QC criteria for metabolomics and exposomics
- Exposomics Society