UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

InSpectra - A platform for identifying emerging chemical threats

Feraud, M.; O'Brien, J.W.; Samanipour, S.; Dewapriya, P.; van Herwerden, D.; Kaserzon, S.; Wood, I.; Rauert, C.; Thomas, K.V.

DOI
10.1016/j.jhazmat.2023.131486

Publication date
2023

Document Version
Final published version

Published in
Journal of Hazardous Materials

License
CC BY

Link to publication

Citation for published version (APA):
Feraud, M., O'Brien, J. W., Samanipour, S., Dewapriya, P., van Herwerden, D., Kaserzon, S., Wood, I., Rauert, C., & Thomas, K. V. (2023). *InSpectra* - A platform for identifying emerging chemical threats. *Journal of Hazardous Materials*, *455*, Article 131486. https://doi.org/10.1016/j.jhazmat.2023.131486

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:27 May 2024

Journal of Hazardous Materials 455 (2023) 131486

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

InSpectra – A platform for identifying emerging chemical threats

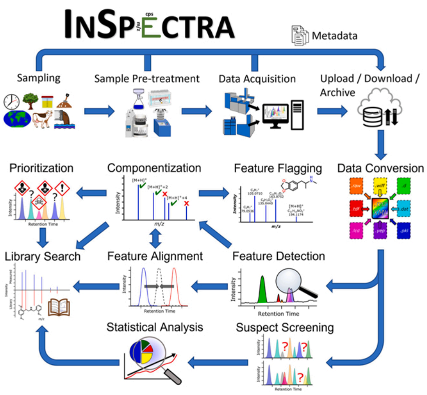
Mathieu Feraud [a,1], Jake W. O'Brien [a,b,*,1], Saer Samanipour [a,b,c,*,1], Pradeep Dewapriya [a], Denice van Herwerden [b], Sarit Kaserzon [a], Ian Wood [d], Cassandra Rauert [a], Kevin V. Thomas [a]

[a] *Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, Australia*
[b] *Van 't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, Netherlands*
[c] *UvA Data Science Center, University of Amsterdam, Netherlands*
[d] *School of Mathematics and Physics, The University of Queensland, Australia*

HIGHLIGHTS

- Open-source/access high-resolution mass spectrometry data processing platform was built.
- Automated vendor independent non-target analysis and suspect screening workflows.
- Archival of HRMS data and metadata in a relational database for retrospective processing.
- Access to largest community curated high-resolution mass spectrometry library.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Lingxin Chen

Keywords:
Non-target analysis
Suspect screening
Archiving
Chemicals of emerging concern
High-resolution mass spectrometry

ABSTRACT

Non-target analysis (NTA) employing high-resolution mass spectrometry (HRMS) coupled with liquid chromatography is increasingly being used to identify chemicals of biological relevance. HRMS datasets are large and complex making the identification of potentially relevant chemicals extremely challenging. As they are recorded in vendor-specific formats, interpreting them is often reliant on vendor-specific software that may not accommodate advancements in data processing. Here we present *InSpectra,* a vendor independent automated platform for the systematic detection of newly identified emerging chemical threats. *InSpectra* is web-based, open-source/access and modular providing highly flexible and extensible NTA and suspect screening workflows. As a cloud-based platform, *InSpectra* exploits parallel computing and big data archiving capabilities with a focus for sharing and community curation of HRMS data. *InSpectra* offers a reproducible and transparent approach for the identification, tracking and prioritisation of emerging chemical threats.

[*] Corresponding authors at: Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, Australia.
E-mail addresses: j.w.obrien@uva.nl (J.W. O'Brien), s.samanipour@uva.nl (S. Samanipour).

[1] These authors contributed equally.

https://doi.org/10.1016/j.jhazmat.2023.131486
Received 10 November 2022; Received in revised form 20 April 2023; Accepted 23 April 2023
Available online 24 April 2023
0304-3894/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

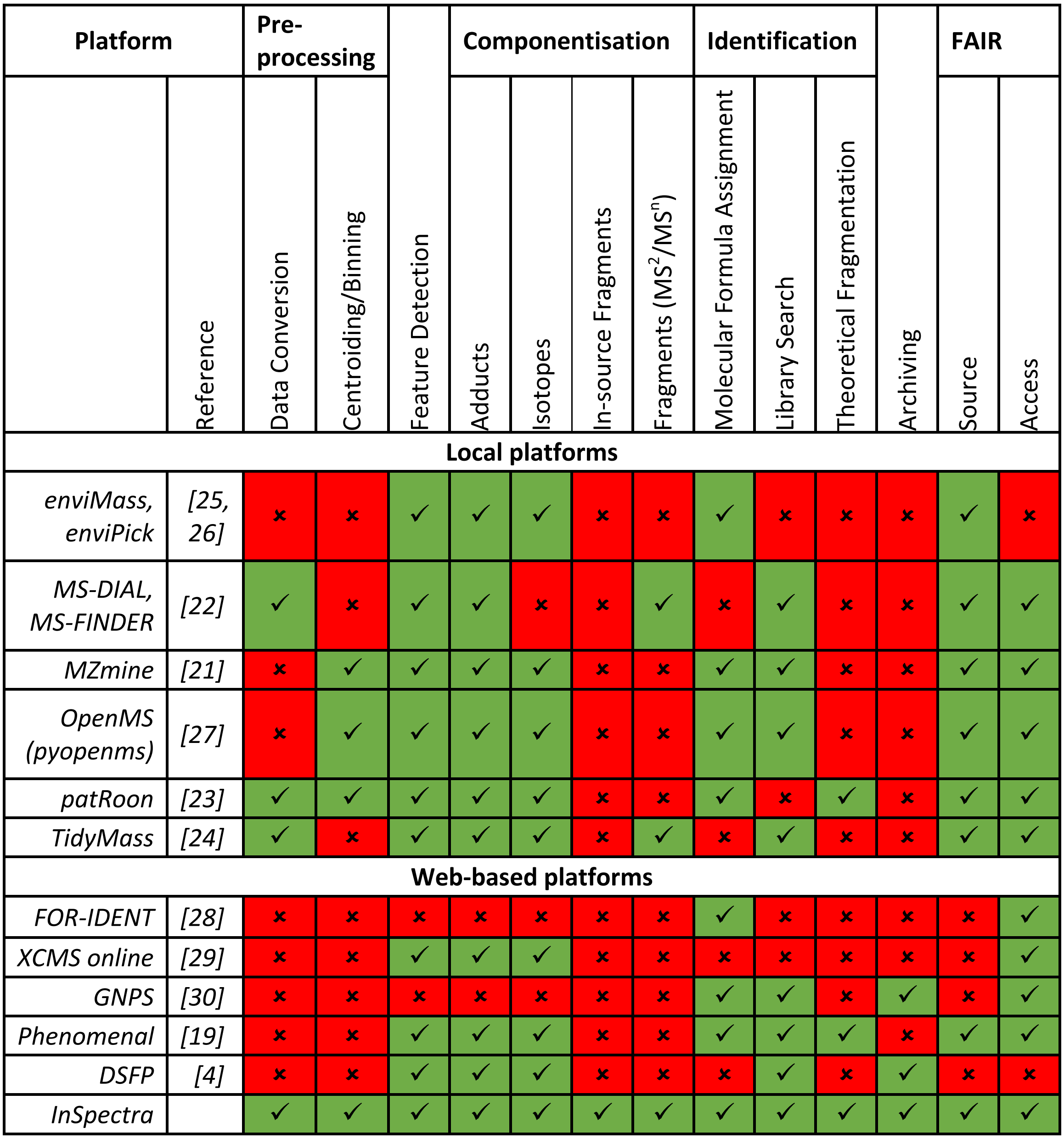
1. Introduction

In 2016, the World Health Organization reported 1.6 million deaths and 45 million disability-adjusted life-years lost because of known chemical exposures. Which chemicals are responsible is poorly understood.[1] A key challenge for regulators is the lack of and difficulty in collecting sufficient experimental evidence between chemical exposure and effects on humans and the environment [2–4]. Part of this challenge is that our “chemosphere” is overly complex, dynamic, and ever-expanding with most of the chemicals indexed in the Chemical Abstract Service (currently 193 million) not characterised with respect to their potential effects on human safety and environmental health.[5]

Non-target analysis (NTA) employing high-resolution mass spectrometry (HRMS) is becoming one of the most comprehensive approaches analytical chemists can use to answer questions related to the fate and exposure of chemicals [6,7]. NTA uses full-scan HRMS data without a priori assumptions about chemical composition of the samples, independently from their levels of complexity. A sub-type of NTA, suspect screening, also uses full scan HRMS data but limits data analysis to a predefined (suspect) list of chemicals [7–12]. Sharing of the full-scan HRMS data and applying retrospective analysis has been proposed as an early warning system for rapidly identifying emerging chemical threats across the globe.[13] This has yet to be realised as challenges remain particularly regarding the archiving of HRMS data, metadata, and processing capabilities.

Advancements in HRMS technology, such as time-of-flight and Orbitrap instruments, particularly when coupled with liquid or gas chromatography (LC-MS and GC-MS), have facilitated rapid NTA and suspect screening assays. These experiments generate thousands of MS/MS spectra per sample in a matter of minutes [14] and whole projects potentially generating millions of spectra. Such large volumes of complex data implies that manual analysis is unfeasible (i.e., combination of

Table 1
Overview of commonly used open-access and/or open-source data processing tools/platforms for the analysis of non-target analysis and suspect screening data [4,19,21–30].



Platform		Pre-processing			Componentisation				Identification				FAIR	
	Reference	Data Conversion	Centroiding/Binning	Feature Detection	Adducts	Isotopes	In-source Fragments	Fragments (MS^2/MS^n)	Molecular Formula Assignment	Library Search	Theoretical Fragmentation	Archiving	Source	Access
Local platforms														
enviMass, enviPick	*[25, 26]*	×	×	✓	✓	✓	×	×	✓	×	×	×	✓	×
MS-DIAL, MS-FINDER	*[22]*	✓	×	✓	✓	×	×	✓	×	✓	×	×	✓	✓
MZmine	*[21]*	×	✓	✓	✓	✓	×	×	✓	✓	×	×	✓	✓
OpenMS (pyopenms)	*[27]*	×	✓	✓	✓	✓	×	×	✓	✓	×	×	✓	✓
patRoon	*[23]*	✓	✓	✓	✓	✓	×	×	✓	×	✓	×	✓	✓
TidyMass	*[24]*	✓	×	✓	✓	✓	×	✓	×	✓	×	×	✓	✓
Web-based platforms														
FOR-IDENT	*[28]*	×	×	×	×	×	×	×	✓	×	×	×	×	✓
XCMS online	*[29]*	×	×	✓	✓	✓	×	×	×	×	×	×	×	✓
GNPS	*[30]*	×	×	×	×	×	×	×	✓	✓	×	✓	×	✓
Phenomenal	*[19]*	×	×	✓	✓	✓	×	×	✓	✓	✓	×	✓	✓
DSFP	*[4]*	×	×	✓	✓	✓	×	×	×	✓	×	✓	×	×
InSpectra		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

multiple steps).[15]

The data processing workflow for NTA and suspect screening assays include several steps from data conversion to structural elucidation.[16] There are several open-access and/or open-source data processing tools that tackle parts of such workflows (see Table 1 for examples). As of now, all such tools and workflows do not take into consideration information associated with the sample itself and its preparation. Previous studies have shown that sample matrix, sample collection, pre-treatment, separation, and data acquisition all impact the explored chemical space (i.e., the coverage of all organic chemicals within a sample). For example, each sample matrix has its own chemical space, hence matrix selection limits the types of chemicals that are present [7] and tools are becoming more available to better explore and identify the chemical space of a samples matrix both pre- and post- acquisition for better representation of NTA results.[17] As such, for a global early warning system for rapidly identifying chemicals of emerging concern to work, the metadata associated with the entire process (from sample collection through to chemical identification) must be archived and retrievable. Currently available platforms (both local and web-based) for processing HRMS data are particularly limited in terms of sharing, archiving and retrospective analysis of NTA data – let alone for capturing metadata.

PhenoMeNal is one of the most transparent (i.e., following the FAIR principles [18]), modular, and scalable web-based platforms currently available.[19] This platform is a Galaxy based [20] system where the user can choose the type of workflow and the individual tools used for each step. Currently the *PhenoMeNal* platform does not have a user interface and requires users to tackle the compatibility of the inputs and outputs of different tools within the workflow (e.g., a feature list generated by one tool may not be compatible with a specific identification tool). *PhenoMeNal* also does not have any archiving or retrospective analysis capabilities. Consequently, this platform has not been widely utilised within the exposomics community. Another web-based platform with limited open-access and closed source is *XCMS O*nline (17). This platform enables feature detection, alignment, and simple statistical analysis. However, feature identification is not part of the workflow and must be performed independently. Additionally, this platform is not able to deal with Data Independent Acquisition (DIA) or profile mode data types. Furthermore, the platform is not modular, and the source code is closed with no archiving capabilities.

Two web-based platforms that have archiving capacity are the *Norman Digital Sample Freezing Platform* (*DSFP*) [4] and *Global Natural Products Social Molecular Networking* (*GNPS*).[15] Both platforms are closed source and *DSFP* is limited open-access while *GNPS* is completely open-access. The *GNPS* workflow focuses on molecular networking with limited other capabilities. On the other hand, *DSFP* has a relatively complete suspect screening workflow from the feature detection to library search. However, it does not perform any MS^2 clean-up, therefore being more suitable to process Data Dependent Acquisition (DDA) chromatograms.

Of the local platforms available, *MZmine 2* [21] and *MS-DIAL/MS--FINDER* [22] are the most used due to their user-friendly graphical user interfaces and complete workflows. While *MZmine 2* is very modular with different options for each step of the workflow, *MS-DIAL* follows a fixed workflow. *MS-DIAL/MS-FINDER* can handle both DDA and DIA files as well as feature identification. The current version of *MZmine 2* does not have the capability to handle DIA chromatograms and requires either a local spectral library or external *R* plug-ins to be able to identify features. Recent developments in available local platforms are *patRoon* [23] and *TidyMass* [24] which are both *R* based platforms integrating several tools for different steps of the workflow [23,24]. While *patRoon* has more extensive identification/annotation capabilities compared to *MZmine 2* and *MS-DIAL/MS-FINDER*, its graphical user interface is much simpler – needing minimal *R* programming knowledge. *TidyMass* however doesn't have a graphical user interface.[24] Like *MZmine 2*, *patRoon* is not able to handle DIA data. All the mentioned platforms have limited scalability (i.e., parallel computing without user intervention) nor archiving capacity incorporated in them. See Table 1 for further detail on each platform's current capabilities.

Considering the above, we present a web-based open-source and open-access software platform called *InSpectra* that provides vendor independent complete NTA and suspect screening workflows (i.e., from data conversion through to identification). Currently, *InSpectra* can only process LC/HRMS and not GC/HRMS data. As a cloud-based platform, to optimise the way emerging chemical threats are identified, *InSpectra* takes advantage of parallel computing and the ability to archive all data and associated metadata with a view for sharing and community curation of HRMS data with rapid retrospective analysis capabilities. Additionally, *InSpectra* is completely modular with a future vision to incorporate state-of-the-art algorithms and tools. Furthermore, with this paper we invite collaboration with research teams across all disciplines to trial *InSpectra* and assist with the global curation of HRMS datasets.

2. Methods

2.1. Sample preparation

To demonstrate *InSpectra's* capabilities and current workflows, the Library Search Workflow was applied to multiple LC-HRMS datafiles acquired on both QToF and Orbitrap mass spectrometers coupled to liquid chromatography systems and covering multiple complex sample matrices (cow blood extracts, cow serum extracts, stormwater, and wastewater). All data were collected using electrospray ionisation (ESI) positive ionisation mode to allow comparison. The following sample types used different instrumentation, columns and mobile phases to illustrate that *InSpectra* can process regardless of these parameters and the outputs can still be compared.

Stormwater samples were collected from, a tributary of the Brisbane River, during major storm events in June and October 2020 (Fig. S1 [31]). The samples were solid phase extracted (SPE) as described previously.[31] Cow blood and serum samples were collected from cattle exposed to contaminated groundwater. All blood and serum samples were collected by a qualified person under the guidelines described by UQ ethics approval (#ANRFA/ENTOX/153/16). The samples were kept frozen (−20 °C) until extraction. An equal amount (300 µl) of blood and serum from randomly selected individuals (n = 4) was pooled together for extraction. Here, we did not consider pooling blood and serum necessarily from the same individuals. The pooled cow blood and serum samples (1 mL) were extracted as described previously.[32] Wastewater samples were collected on August 9, 2016.[33] Aliquots of the wastewater sample (1 mL) were spiked with a mixture of isotope-labelled standards (50 ng of each compound) and filtered through a 0.45 µm PTFE syringe filter directly into a glass LC-vial. The prepared samples were kept frozen until analysis.

2.2. Instrumental analysis

The stormwater, blood and serum samples were analysed with high-performance liquid chromatography (ExionLC AD, AB Sciex, Ontario, Canada) coupled to a SCIEX X500R Quadrupole Time-of-Flight (QTOF) mass spectrometer (AB Sciex, Ontario, Canada) equipped with electrospray ionisation (ESI). The stormwater samples were eluted using a Kinetex C18 100 Å analytical column (2.6 µm, 100 mm × 2.1 mm; Phenomenex, Lane Cove, Australia) fitted with a guard cartridge (SecurityGuard™, Phenomenex, Lane Cove, Australia). Chromatographic separation was achieved with mobile phases consisting of Milli-Q water (A) and methanol (B) both acidified with 0.1 % formic acid. The cow blood and serum samples were eluted using ACQUITY UPLC HSS T3 Column (1.8 µm, 2.1 mm × 100 mm; Waters Corporation, Milford, MA) equipped with an ACQUITY guard cartridge. Mobile phased used were Milli-Q water (A) and methanol (B) both containing 2 mM ammonium acetate. The injection volume was set at 10 µl, and the column temperature was maintained at 40 °C for all the samples. Full scan high-resolution mass spectrometric data were collected across 100–1100 *m/z* (MS^1) and 50–1100 *m/z* (MS/MS) in SWATH operation mode. The parameters of the SWATH analysis were as follows: ion source temperature 550 °C; ion spray voltage 5000 V; curtain gas 30 L/min; ion source gas 1 and 2, 60 psi; declustering potential 80 V (DP); and collision energy 35 V (CE). The SWATH window parameters are given in Table S4.

The wastewater samples were analysed using an ultrahigh performance liquid chromatography coupled to a Q Exactive™ HF Hybrid Quadrupole-Orbitrap™ mass spectrometer (UHPLC-OrbitrapMS/MS, Thermo Fisher Scientific, San Jose, USA) with ESI. Separation was achieved with a reverse-phase Hypersil GOLD™ aQ C18 polar-endcapped column (1.9 µm, 2.1 mm × 100 mm; Thermo Fisher Scientific, San Jose, USA) using a binary mobile phase gradient consisting of Milli-Q water (A) and acetonitrile (B), both containing 0.1 % formic acid. Detailed information on the gradients used is given in Table S3. The mass spectrometry parameters used for the analysis have been described previously.[34]

2.3. Quality control and quality assurance (QA/QC)

All samples were spiked with internal standards to monitor instrument conditions throughout the analysis (from sample-to-sample variations) and assess potential drift. A mixture of reference standards was injected in regular intervals to monitor chromatographic and MS performance. Procedural blanks (Milli-Q water spiked with IS and extracted), solvent blanks (methanol) and instrument blanks (Milli-Q water) were analysed alongside samples. All the samples were injected in triplicate. Instrument calibration and resolution adjustments were performed regularly throughout the analysis. System calibration error was maintained at less than 2 ppm.

2.4. Data processing for *InSpectra* workflow demonstration

To demonstrate the potential of the platform, two workflows were discussed and compared. First, all instrument raw datafiles were uploaded to *InSpectra* and processed as per the selected workflow. The raw files were automatically converted to mzXML format, using the parameters listed in table S5. The two workflows are the library search and *Suspect Screening* workflow, which are described in more detail in section Library search identification workflow and *Suspect Screening* workflow, respectively. Briefly, for the library search workflow, feature detection, componentization, and library searching are performed subsequently, extracting as much information from the datafile and matching this to the spectral library. Whereas, the *Suspect Screening* workflow, performs suspect screening directly on the mzXML files, obtaining potential candidates for selected compounds from the provided suspect list. The parameters corresponding to each of the steps in the workflows can be found in table SI2.

The compounds for the suspect lists were selected based on frequently found library search identifications for this dataset. For this, only the components matching with experimental library entries were considered (as opposed to theoretical). The chemicals were selected based on confidence of results, expert knowledge, and relevance. Using the InChIKeys of those compounds, their mass spectra were obtained from the *InSpectra* mass spectral library. Only the results which had a final match factor of 0.4 were accepted as potential candidate. In the case multiple potential candidate signals were found for a compound, the one with the highest match factor, followed by intensity was selected. This obtained an overview of which selected compounds were potentially present in each of the mzXML files, using the *Suspect Screening* workflow.

Additionally, this was also done for the library search workflow for the same selected features. A feature was considered identified if it was matched with a component with a match score of at least 4 or higher. If a feature has multiple identified candidates, only the top 5 (those with the highest match factor, followed by those with the highest intensity if match factor is equal) are kept and used for the comparison of identified features found with the library search and *Suspect Screening* workflow. To compare relative chemical intensities between the samples, the areas of all features in each sample were normalised to the sum of all areas within that sample, and for simplicity only results for library matches which had a final score greater than 5 out of 7 (maximum possible value) were kept and the top 14 by highest relative area plotted as a dendrogram (Fig. 3). Finally, where multiple matches occur for the same feature at different retention times, this potentially indicates isomers which are analytically challenging even with HRMS. However, recent advances in retention indices may be able to assist with correct identification of isomers [35] and it is our intention to incorporate retention indices modelling into future versions of *InSpectra*.

3. Results

3.1. *InSpectra* – the platform

InSpectra is hosted online on a cloud platform that provides many advantages over offline solutions, including independence of end user computer; scalability; ability to archive all data and metadata, and traceability of all processing. The web platform integrates a suite of open source and open- access tools enabling the generation of multiple workflows (Fig. 1). Examples of such workflows and case studies using different workflows are discussed in detail below (see section “Example Workflows”).

3.1.1. Online processing and scalability

All processes are performed on an online opensource platform, thus there are no requirements on the user’s computer to install any software, have a specific operating system, meet specific/minimum system requirements, pay licensing fees, etc. The only requirement is that the user has a computer with a web browser and an internet connection capable of uploading the files they wish to process. Currently all data is stored and processed within the cloud hosted by Amazon Web Services (AWS), which can potentially be moved to other cloud providers. *InSpectra* processes have been configured so that regardless of the number of the files in a job, an adequate number of optimised processing computers will be started to perform the needed tasks. The number of computers open is linearly correlated to the number of files for processing. The scalability of the infrastructure allows it to process hundreds of requests as efficiently as one without wait-time. Currently *InSpectra* is configured so that all samples in a batch are processed at the same time on individual and independent computers, thus the size of a batch and the number of users has negligible effect on wait-time for a sample to be processed. The processing time is dependent on the size of the data file and the computer used averaging 3–5 h to process a job. It is expected as the different algorithms of *InSpectra* are updated the processing time will be reduced.

3.1.2. Archiving of data, metadata, and traceability of processing

InSpectra has an in-built archiving system to store all data from raw instrument datafiles, metadata, experimental conditions to the outputs of the individual tools within the workflow. This includes the processing parameters, tools used and their versions. These files are stored in a repository and, depending on access requirements, can be stored on low access requirement infrastructure to minimise storage costs. The metadata recorded automatically includes parameters used for processing the data, the metadata of the HRMS files themselves (e.g., instrument used, brand, ionisation mode, etc., which are read directly from the raw HRMS datafiles themselves), and versions and inputs of algorithm used while processing the files. The metadata recorded manually are of the samples themselves (e.g., sample matrix, location, time, sample preparation, etc.). This data is stored in a relational database, enabling rapid and easy analysis and further processing (See Fig. 2). In fact, this relational

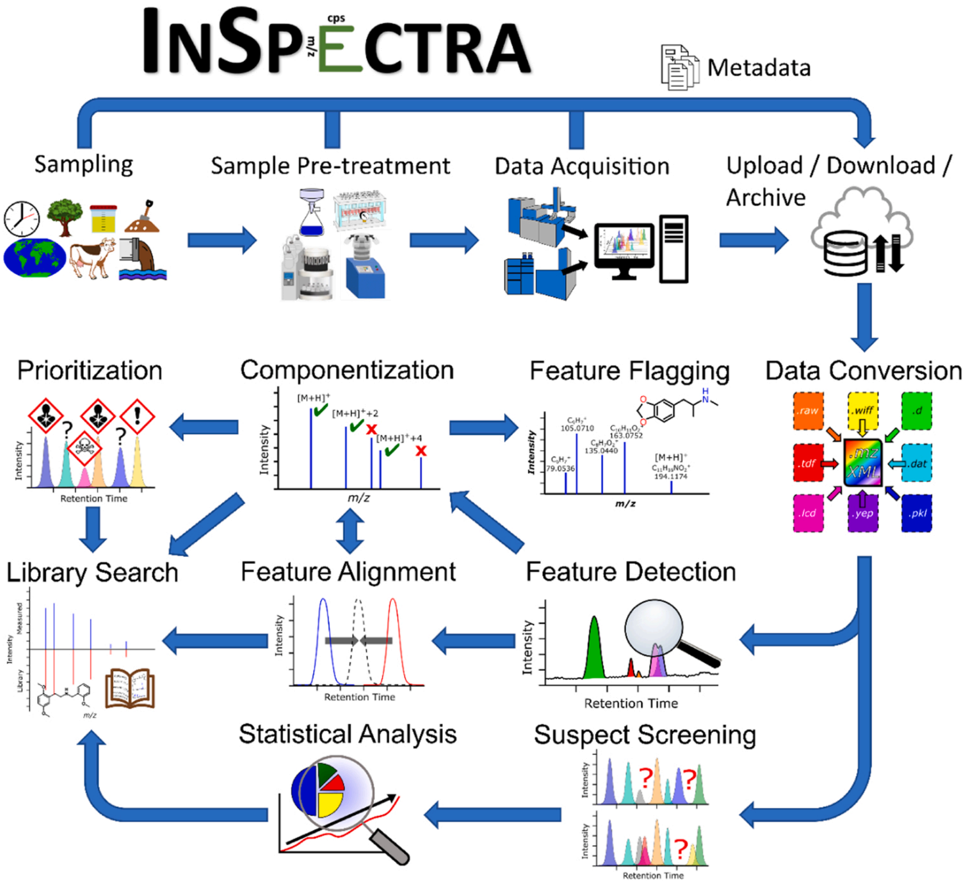


Fig. 1. Overview of the InSpectra platform including current and future tools and workflows.

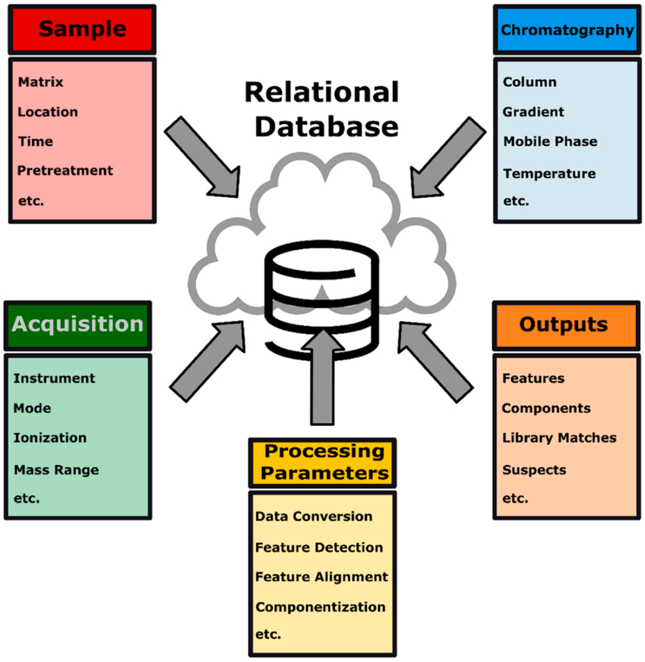


Fig. 2. Overview of InSpectra's relational database.

database is key to allowing the platform to be used for sharing and retrospective analysis of full scan HRMS data as an early warning system for rapid detection of chemicals of emerging concerns across the globe. The collection of such metadata (e.g., sample type and sample preparation steps) are currently hindered by the lack of a user-friendly graphical web interface, which will be addressed soon.

3.2. Tools and workflows

The algorithm description and the validation procedure are provided under the section Code Availability.

3.2.1. Use of open-source tools

InSpectra was built using open-source algorithms which are available via the Git repositories (i.e., Bitbucket), resulting in reproducible and transparent outputs and workflows. Such a level of transparency is often difficult to achieve, given that HRMS instrument vendor software is proprietary, closed source, and closed access. This black box strategy hinders the objective and fair evaluation of the existing algorithms as well as the direct comparison of their outputs. The algorithms used in *InSpectra* have been tested, validated, peer-reviewed, and published [36–39]. The use of algorithms maintained on Bitbucket also means that updated (and often improved) versions of such algorithms are automatically integrated into *InSpectra*, providing users with access to state-of-the-art processing tools while providing the means for open collaboration. It also allows for complete transparency for all parties to understand all parts of the data processing if they wish to do so.

The platform makes use of multiple languages and tools to facilitate a seamless data processing workflow from conversion of raw HRMS data files to identification/annotation and statistical analysis. Python is used for the connection of the algorithms on the backend, as it is a very well-supported language, is quick to code in, and supports a multitude of tools to communicate with other languages and computers. MySQL is used as the database, where all data and metadata, are stored. The different modules are mainly written in Julia, which is a dynamic language that is quite efficient at process-heavy tasks, However, it is important to note that *InSpectra* as a modular platform is not dependent on any language for processing, if an algorithm can run on UNIX or windows, it can be integrated into *InSpectra*. *ProteoWizard* (21), used for HRMS data conversion, is written in C+ + which is an extremely efficient language. Because Python has extraordinarily strong application programming interfaces (APIs), the modules can be written in any language, such as R, matlab, C#, PHP, etc. All scripts for the platform management and database structures used can be made available upon request. For cases where a different combination of tools is needed for custom workflows, a local version of *InSpectra* can be deployed on both local workstations and/or high-performance computing servers. This also enables the cases where due to data sensitivity the data cannot be uploaded to the commercial clouds (e.g., forensic laboratories).

3.2.2. Modularity

The steps included in the workflows constitute different modules of the platform, providing maximum flexibility on the potential workflows and tools to be used. The core algorithms of *InSpectra* are sourced directly from their respective GIT repositories which allows updates and fixes from collaborators to be automatically incorporated in *InSpectra*. Because the software versions are stored with the metadata, if a new update has an impact on the quality of results, the files can be easily reprocessed. As new tools are developed for *InSpectra*, they can be added as separate modules to improve existing workflows or as additional workflows.

The main/core workflow in *InSpectra* includes data conversion, feature detection, componentisation, and identification steps. A brief description of these tools is provided below.

3.2.3. Conversion of raw HRMS datafiles

Once the HRMS datafiles are uploaded into the platform, they are converted into the *mzXML* [38] format. This was chosen as it is an open-source format and creates coherency between the many different vendor formats and *InSpectra's* algorithms. Future versions of *InSpectra* will include the mzML [40] format to facilitate the use of data with ion mobility information. Currently *ProteoWizard's* format conversion utility *msConvert* is used for HRMS data conversion.[41] The parameters used for this conversion are stored in the database. InSpectra can handle all the data formats that *ProteoWizard* can, because the *self-adjusting feature detection* (*SAFD*) [36] is conducted on MS^1 only and *CompCreate* can handle DDA, DIA as well as SWATH.

3.2.4. Feature detection

Feature detection is used to obtain the MS^1 information on the parent, adduct, isotope, and in-source fragment ions, for which the *SAFD* algorithm was used. This algorithm performs feature detection by fitting a three-dimensional gaussian on profile data, requiring no prior binning or centroiding. The current version of *SAFD* is capable of handling both profile and centroided data.[39] As three-dimensional feature detection (i.e., profile mode) is more resource intensive compared to two-dimensional (i.e., centroided data) feature detection, *SAFD* benefits from *InSpectra*'s cluster computing processing capabilities. The *SAFD* algorithm takes an mzXML file and a set of parameters as inputs, comprising of the maximum number of iterations, maximum and minimum peak width in the time domain, mass resolution of the instrument, minimum peak width in the mass domain, correlation threshold, minimum intensity, signal to noise ratio, and signal increment threshold. During the process of fitting a three-dimensional gaussian, the user defined parameters (e.g., widths in the mass and time domain) are only utilised as the first guess and subsequently adapted according to the experimental data. The *SAFD* algorithm outputs a CSV file with the detected features, containing information on the retention time, mass, area, intensity, peak purity, and mass resolution. *SAFD* has been shown to produce more reliable results compared to XCMS,[42] a state-of-the-art algorithm.

3.2.5. Feature alignment

For feature alignment, an in-house algorithm was developed, where the feature lists (i.e., SAFD outputs) run using the same experimental conditions are combined to generate aligned feature list. This algorithm employs a user defined percentage of the measured peak widths in time and *m/z* domains to group the features, having a default value of 50 % (i.e., 0.5). This algorithm follows the same principal as the conventional approaches using a retention window and *m/z* window for aligning the features. The main difference is that these windows are dynamically adjusted based on the peak widths in time and mass domains. Previous tests via internal standards and different matrices have indicated the applicability of this tool. Currently, this algorithm is implemented as a part of the SAFD package [43] and is also able to align feature lists generated by other feature detection algorithms.

3.2.6. Componentisation

Componentisation is used for grouping information belonging to unique chemical constituents, including adducts, isotopologues, and fragments (including in-source fragments). For this, the componentisation algorithm *CompCreate* [37] was used, since it can obtain both MS^1 (i.e., parent, isotopes, adduct, and in-source fragments) and MS^2 (i.e., fragments) information. The *CompCreate* algorithm can process data coming from both DDA and DIA approaches. Additionally, it has built in processes for the Sciex's SWATH and multi-collision data types. The algorithm uses the MS^1 features obtained during feature detection as potential precursor ions. For all these potential precursor ions, both the MS^1 features and MS^2 peaks were grouped based on the time difference between the retention time at the apex, Pearson's correlation of the extracted ion chromatograms (i.e., peak shape check), and information specific to the ion type (Fig. 3). For the latter, adducts are identified based on a database of frequently detected single charged adducts in LC-HRMS experiments (e.g., M+Na).[44] Isotopes are detected based on the elemental mass defect between the parent and potential isotope mass, assuming that elemental mass defect is similar for these ions since they have the same molecular structure.[45] Whereas (in-source) fragments are further filtered based on the probability of the neutral loss (i.e., mass difference between fragment and parent ion) database, which is further elaborated in section Neutral loss database. The *CompCreate* algorithm outputs a CSV file that contains both the generated components and un-grouped features as well as the spectral information at MS^1 and MS^2 levels. The results coming from *CompCreate* have not yet effectively been compared with existing algorithms, since, to our knowledge, there is no open-access/source algorithm providing both the MS^1 and MS^2 information (Table 1).

3.2.7. Neutral loss database

To calculate the neutral losses (NLs) the measured parent ion mass of each high-resolution (i.e., resolutions $\geq$ 10,000) entry in MassBank EU were subtracted from each measured fragment associated with that parent ion.[46] The absolute value of the resulting NLs was binned using a mass tolerance of 0.003 Da. Consequently, we were able to calculate the frequency of occurrence for each NL value (Fig. 4).

To calculate the probability of false detection (FD) associated with each NL occurrence probability, bootstrapping with 20,000 iterations was used. During each iteration, a randomly selected parent ion was matched with a randomly selected set of fragments (i.e., true negatives). These true negative NLs were used for calculating the false detection rates, binning them with the same mass window (i.e., 0.003 Da), and calculating the corresponding occurrence probabilities. The probability of FD appeared to be $\leq$ 2 % for the NLs with an occurrence probability larger than 3 %. During the componentization process, only NLs with an occurrence probability $\geq$ 5 % were considered for fragment detection, corresponding to an FD rate of $\leq$ 0.1 %.

3.2.8. Library search

Library search is used for the identification of components or features

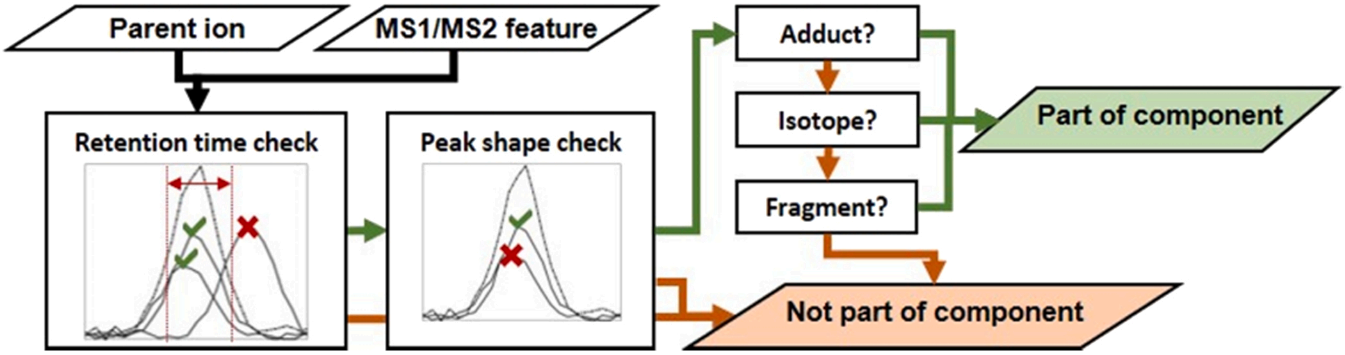


Fig. 3. Workflow for the grouping of information belonging to the parent ion. The signals from the MS^1 feature list are all evaluated as potential parent ion. For each parent ion, the corresponding MS^1 and MS^2 signals are obtained and evaluated based on their apex retention time, peak shape correlation. When these two requirements are passed, the MS^1 or MS^2 signal is further evaluated based on the type (i.e., adduct, isotope, or fragment).

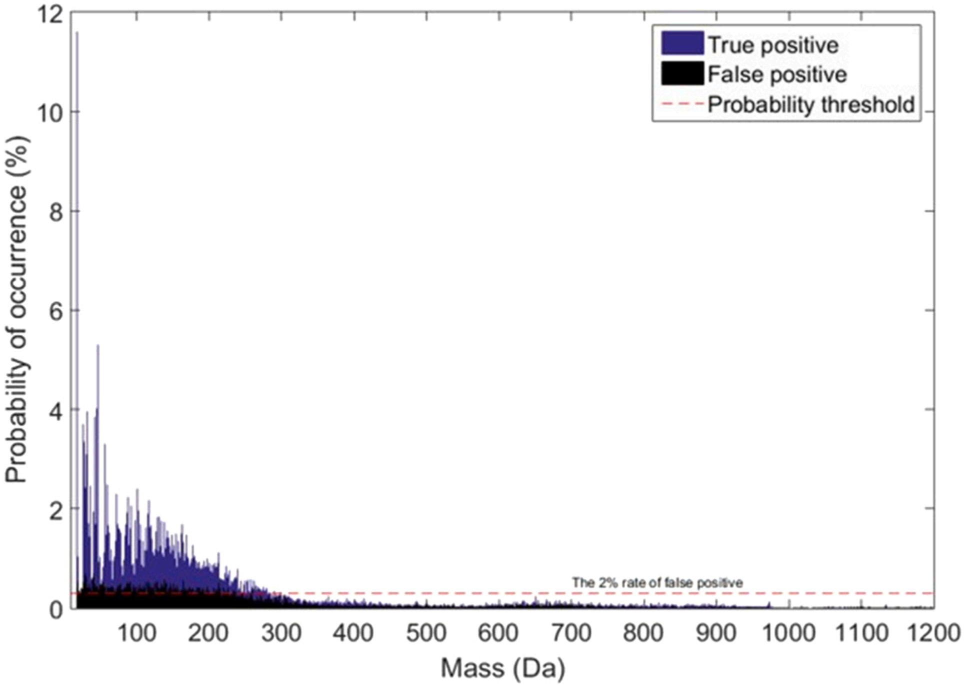


Fig. 4. The TP and FP Occurrence probability for each NL mass. The red line represents the occurrence probability, for which the overall NL false detection rate is ≤ 2 %.

based on similarity with database spectra. For the library search InSpectra uses the Universal Library Search Algorithm (ULSA). This algorithm uses the combination of seven parameters, these are, the number of matched fragments in the measured spectrum, number of matched fragments in the library spectrum, mass error of precursor ion, average mass error of matched fragments, standard deviation of the mass error of the matched fragments, reverse match factor, and forward match factor to rank the potential candidate for each component. This combination has shown to be effective in reducing the impact of the instrumental conditions on the quality of identifications.[37] Additionally, for the components a complete library search is performed while for the remaining features a molecular formula assignment is performed based on the compounds present in the databases. For spectral matching of components, an initial search in the *InSpectra* database is performed based on the precursor ion mass using the mass window associated with each component. On average this mass window ranges between ± 10 mDa ± 30 mDa. For each of those spectra, a Final Score (quality of spectral match) is calculated based on seven different parameters(see above). This combination has shown to be effective in reducing the impact of the instrumental conditions on the quality of identifications.[37] The influence (i.e., weight) of each parameter can be specified by the user via a weight vector of seven values ranging between zero and one. In future versions of *InSpectra* machine learning based tools to assess the probability of true positive identification will be implemented as an additional source of information for the analyst. As for the features that are found in the input list, molecular formula assignment is performed with the *InSpectra*'s database (detailed below). Potential molecular formula assignment is mainly based on the mass error between the measured *m/z* values and the theoretical mass. This deviation is normalized by the mass tolerance provided by the user and then converted to a score between zero and one, where one is related to a case with no mass error. *ULSA* outputs a list containing all potential candidate identifications or molecular formula assignments with their corresponding Final Score as well as the list of matched fragments for the components and features, respectively. The output is stored on the platform and its metadata is stored and referenced in the database.

3.2.9. *InSpectra's library*

The database of the library search is obtained from two sources, the *MassBank Project* [48] and CFM-ID.[49] The *MassBank Project* [48] contains experimental spectra of both endogenous and exogenous compounds of which there are 89,826 distinct spectra, 15,059 unique compounds, and 16,840 unique isomers. 25,935 of these entries have a recorded resolution of 7500 or above. The library database also includes 700,000 theoretical (predicted via CFM-ID [49]) spectra from the EPA's National Centre for Computational Toxicology CompTox Chemical Dashboard [47] with spectra predicted for EI-MS and ESI-MS/MS in both positive and negative ionisation modes.[50] To our knowledge the *InSpectra* platform is the only platform capable of searching against such a large spectral library, which provides the researchers access to such resources. While *InSpectra* in its current form does not have capability for users to add new spectra to its library, as the experimental libraries are currently sourced from the *MassBank* project, new submissions to *MassBank* will be incorporated when the library image is updated. This was chosen for design, efficiency and reliability considerations. As of now, the database of *InSpectra* does not automatically update its database from the *MassBank* database. Future revisions are planned to automate this.

3.2.10. *Exploratory data analysis*

To be able to perform statistical analysis, first the information (e.g., features or components) need to be linked across these samples. Connecting identified components to each other across different samples would be the easiest case, enabling direct spatial and temporal trend analysis of chemicals across different matrices. The identity-based alignment functionality currently implemented in InSpectra is essential for performing the detection of emerging chemical threats.

As for unidentified features, feature alignment can be performed for samples analysed with the same method. *InSpectra*, through its relational database, can group the datasets measured via the same methods and align their feature lists and/or components, enabling similar types of trend analysis as for the identified features and increased understanding of the covered chemical space of a sample set.

The feature alignment algorithm that *InSpectra* uses is part of the *SAFD* package.[36] This algorithm aligns MS^1 features across multiple samples based on their retention time and measured mass. To evaluate if features should be matched, the algorithm adapts the mass and retention tolerance based on the peak width in the mass and retention domain, respectively. Alternatively, the mass and retention tolerances could also be provided by the user if a different algorithm has been used for obtaining the MS^1 feature lists and peak widths are unavailable. However, since this approach uses the retention time to match features between samples, it is not yet possible for *InSpectra* to perform statistical analysis for unidentified features analysed via different methods. The next version of *InSpectra* will include a validated retention mapping algorithm to seamlessly connect the unidentified features generated via multiple acquisition methods.[51]

Because all the data and metadata are stored in an SQL database, and the data itself is standardized and can be easily queried, complex queries and retrieval can be performed at all levels of NTA for further analysis and projects. Once *InSpectra* has enough data and metadata entered, the gathering of specific data, which depending on the parameters needed can take hours or days to filter through, could be done in minutes with a simple query. It would make the comparisons of distinct spectra easy to perform and likely more relevant to investigate. For example, finding all the peaks from spectra with a distinct parameter (column, date, instrument) where all other relevant parameters are equal in order to investigate the initial distinct feature on the results of the spectra. For example, if instruments used to investigate influence the correlation between retention time, *m/z* and intensity for the same peaks.

3.3. *Example workflows*

InSpectra enables the user to combine multiple tools that each have their own functions and goals. A complete overview of paths or tools' combinations can be seen in Fig. 1. However, to give a better idea of the overall process and possibilities, two frequently used workflows are described below.

3.3.1. *Library search identification workflow*

One of the most used NTA workflows is for feature identification to identify known unknowns starting from raw data through to a list of identified features (i.e., spectra matched against a library; see Fig. 5). In this workflow the HRMS files are converted to mzXML (a common open-source format) that is then processed for feature detection using *SAFD* to obtain the MS^1 information on the parent, adduct, isotope, and in source fragment ions. *Feature Alignment* then groups the features across multiple feature lists or component lists based on their retention time and *m/z*. The file then undergoes componentisation using *CompCreate* which groups information belonging to unique chemical constituents. The componentised file is then searched against the *InSpectra* database using *ULSA*. Lastly, is *Statistical Analysis*, which offers multiple tools to analyse the results either as a standalone or in context of other stored processed files and its metadata, such as heatmaps, temporal and spatial trends via identity-based alignment approach.

3.3.2. *Suspect screening workflow*

Another commonly used NTA workflow is *Suspect Screening*, which is a top-down approach where only a targeted list of chemicals (that can be sourced from outside of *InSpectra*) is searched for within the samples (see Fig. 6 for an overview) as opposed to the complete NTA workflow, which is a bottom-up approach and uses the complete database of 89,826 experimental and 700,000 theoretical spectra as a feature list. The *Suspect Screening* algorithm uses the suspect list to extract the MS^1 and MS^2 information from the raw data and generates a match factor between the user provided spectra and the experimentally measured one. *Suspect Screening* checks for precursor ion, isotopes, isotopic depth, and the presence of fragments in the MS^2 data. This is a faster process than complete NTA, given that it focuses on specific mass channels (i.e., monoisotopic mass of the suspect analytes), it does not incorporate retention time in the search, and the *Suspect Screening* workflow does not perform the feature detection and componentisation. Additionally, the *Suspect Screening* workflow is considered more sensitive in terms of the ability of the workflow to identify the experimentally detected features than the complete NTA workflow (raw data file to library match) due to its more targeted nature. This workflow generates a list of features with their potential structure, isotopic matching, number of matched fragments, and match factors. This information will facilitate the confidence assessment of the identifications by the analysts. Additionally, to facilitate *Suspect Screening* within *InSpectra*, the *InSpectra* database also includes the InChIKeys associated with each entry. The user can use these InChIKeys to run a query against the database and collect all the spectra associated with those chemicals and thus generate a suspect list.

3.3.3. *Demonstration of InSpectra workflows*

To demonstrate the *InSpectra* platform, two workflows were executed on a variety of samples and discussed for a selected group of suspect and non-target analytes. The first workflow, *Library Search*, performed identification on the components obtained from the mzXML files, while the second workflow, *Suspect Screening*, evaluated if these suspect analytes could be present in the mzXML files based on the known spectra of the chemicals.

Applying the *Library Search* workflow to samples representing four different matrices (wastewater, stormwater, cow blood and serum extracts) resulted in a mixture of chemicals being tentatively identified covering multiple classes of chemicals including pharmaceuticals (e.g., pentoxifylline) and agrochemicals (e.g., oxadixyl) (see Fig. 7). This

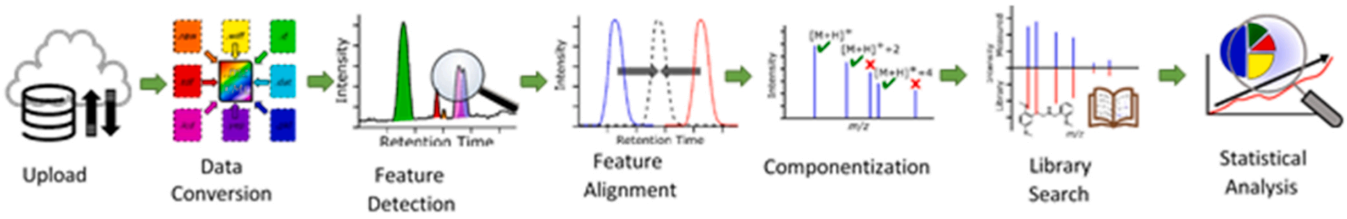


Fig. 5. Example InSpectra workflow: Library Search.

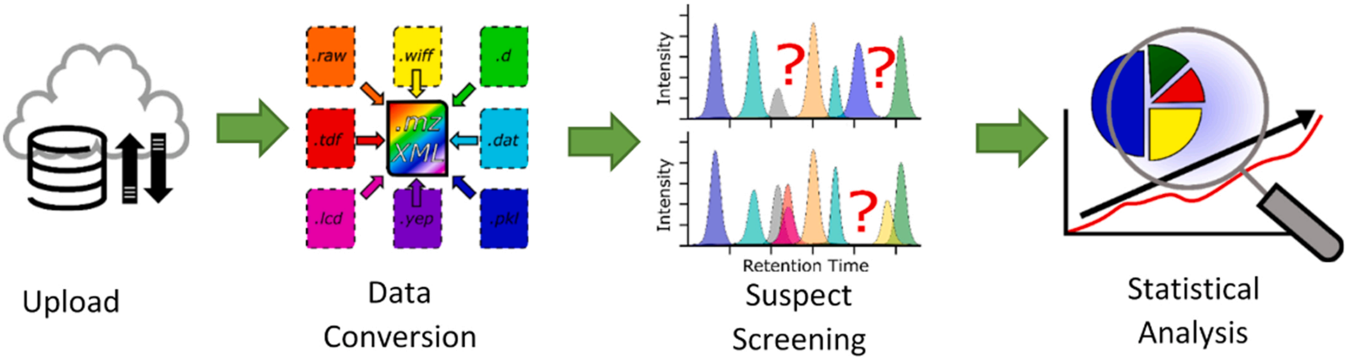


Fig. 6. Example InSpectra workflow: Suspect Screening.

demonstrates the capability of using a full NTA workflow (i.e., Library Search) for biological and environmental analysis, considering the data were acquired on two different vendor instruments using completely independent experimental conditions. The full NTA workflow (*Library Search*) outputs feature lists, extracted component lists and candidate lists that could be used for structural elucidation via other tools/platforms as well as future retrospective analysis for further evaluation. When the same *Library Search* workflow was also applied to stormwater samples collected over a series of time points during multiple storm events but analysed within the same batch, multiple detections of the same tentative chemicals were detected (Fig. 8). In the stormwater samples, the most dominant family of detected chemicals were the agrochemicals (e.g., simeton) and their transformation products (e.g., 2-hydroxyatrazine). The presence and frequency of detection of these chemicals in stormwater may indicate domestic sources contaminating stormwater such as agricultural runoff. These results demonstrate the applicability of *InSpectra* as an early warning system for chemicals of emerging concern.

To demonstrate the *Suspect Screening* workflow, we used chemicals tentatively identified using the *Library Search* workflow to create a suspect list and processed the same files via the *Suspect Screening* workflow. When we performed a direct comparison of the results of the two workflows for temporal data (in Fig. 9 and Fig. 10), in ~58 % of the cases we found complete agreement between the two workflows (see Fig. 11), ~29 % of the cases were only detected by the Library Search workflow, and ~13 % only by *Suspect Screening*. The acceptance criteria for a detection using the *Library Search* workflow was a final score value of ≥ 4 (out of 7) with a minimum of 3 matched fragments, and for *Suspect Screening* a match factor of ≥ 0.2 again with a minimum of 3 matched fragments. To make this comparison consistent, as the samples were from the same batch and were of the same matrix, additional criteria were applied. This considered a consistent retention time between the samples and was based this on the *Library Search* detection. The retention time for each compound was selected by selecting the most frequently recurring retention time in all samples for a distinct accession in the *Library Search* output. The 20 most recurring unique accessions from distinct compounds from the *Library Search* output were then plotted as a categorical heatmap (Fig. 9). Given that the list of chemicals of interest was skewed towards the *Library Search* workflow, a higher detection frequency for this workflow was expected. Additionally, upon further investigation of the discrepancy cases (e.g., 2,6-xylidine Fig. 11 A) the *Suspect Screening* workflow includes peak detection during the fragment matching which is not included in the *Library Search* workflow (Fig. S2 in the SI). This implies that the *Library Search* workflow is less sensitive towards noisy MS2 signals, and thus higher number of tentatively identified features.

The detailed investigation of the extracted ion chromatograms, (XICs) for the tentatively identified chemicals for both MS levels and workflows, further indicated the meaningfulness of the extracted information, and thus the confidence associated with them (Fig. 11 B).

4. Discussion

4.1. Potentials and limitations

In this study we have reported the development and release of an open-source/access platform for the analysis of novel and conventional chemicals of emerging concern in complex samples, from water to biological matrices. This platform is fully modular allowing the inclusion of future tools within the existing workflows. Additionally, the outputs of each step are fully tracible and analysable with the data processing tools outside of *InSpectra* (e.g., marker discovery [52] and/or molecular networking [53]). Additionally, the platform enables community level collaboration, which should ultimately result in a true early warning system of chemicals of emerging concern.

The *Library Search* workflow is reliant on the quality and diversity of the publicly available spectra. Consequently, low quality spectra (e.g., noisy) will result in lower confidence tentative identifications. For example, an entry in MassBank for *erucamide* contains 32 unique fragments (MassBank Accession id: *FIO00884*), but the relative intensities for these range from 89 to 999. For *usnone a*, a compound with a molecular weight of only 344 Da, the MassBank entries are only for a low-resolution instrument and in some cases have more than 200 fragments recorded (MassBank Accession id: *NGA01929*). Simply applying a resolution threshold may not be sufficient due to most entries not having recorded the resolution at which they were acquired. When generating

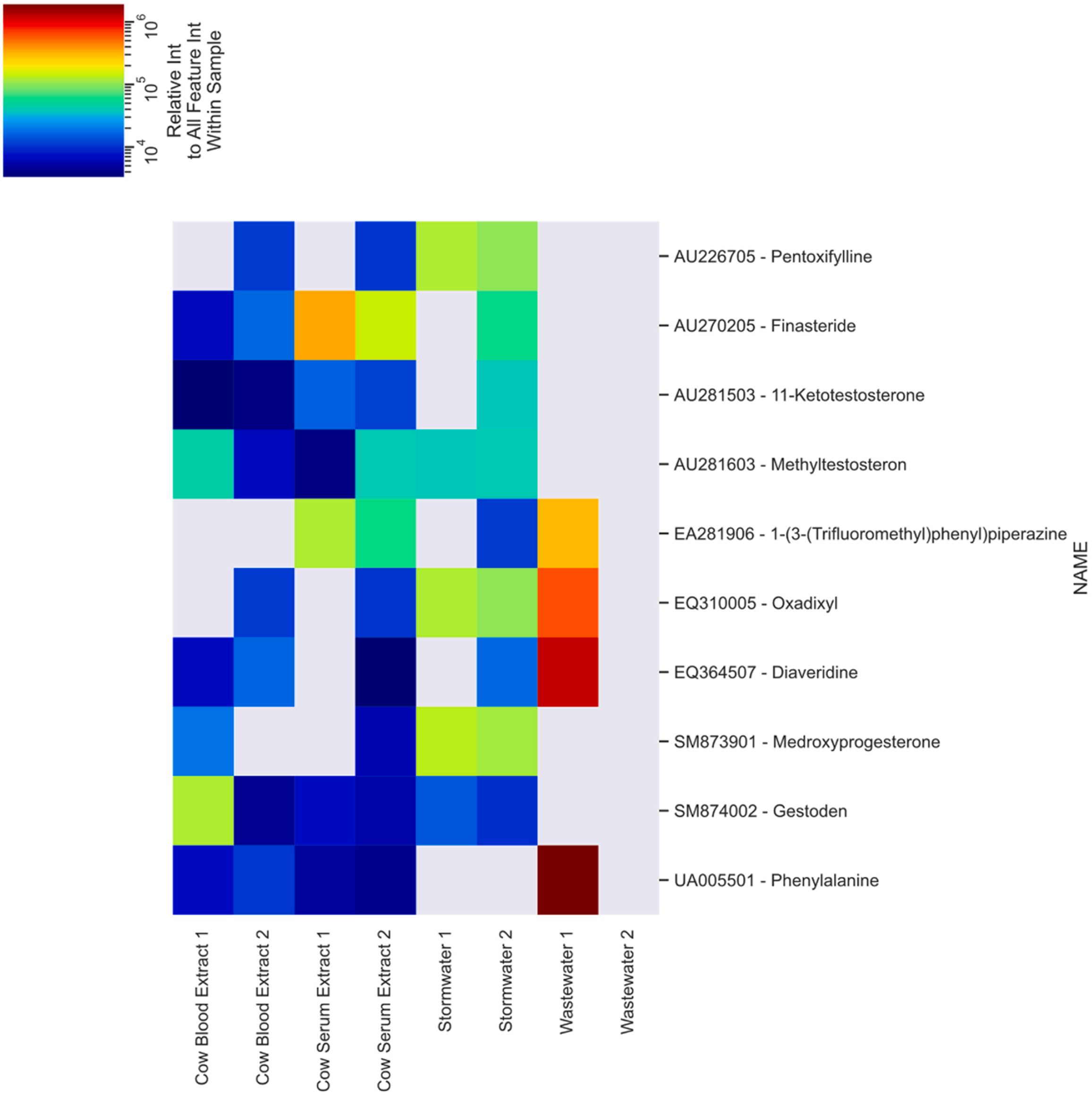


Fig. 7. Heatmap of relative intensity (relative to all intensities in the sample set) of the most prevalent tentative chemical identifications in different matrices (two different samples each of wastewater, stormwater, cow blood and cow serum extracts) using the Library search workflow.

the suspect lists, expert knowledge plays an important role as there should be a balance between the number of fragments included in such a list and the threshold set for the Match Score. A suspect list entry with too many non-diagnostic/low-probability fragments will result in low Match Scores. Finally, like any NTA workflow, the outputs of *InSpectra* may require further expert evaluation to assess their levels of confidence and accuracy and refine the output. To facilitate that, the reports generated from *InSpectra* include all the information (e.g., number of matched fragments and the associated mass error) used for the detection and identification of chemical signals, which can be used by the analyst during the post processing.

4.2. Outlook

The outlook for the *InSpectra* platform includes expanding its user-base through improving the usability of the platform and incorporating additional tools and workflows. Having focused on developing the backend of the platform, it is currently using a command-line interface (CLI) which requires the user to be comfortable using three pre-built functions with arguments to communicate with *InSpectra*. Replacing the CLI with a graphical user interface (GUI) would increase accessibility and usability of the program. As a platform designed for sharing, processing and archival of HRMS data, this will need to be simple and informative to guide users through the various workflows. Priority will be given to ensure users can easily and securely upload their HRMS datasets (including associated metadata) and put them through the existing workflows such a NTA and *Suspect Screening* described above. Already work has begun on new tools such as peak alignment to facilitate easy comparison between samples and novel prioritisation and structural elucidation tools.

In its current form the *InSpectra* platform is limited to processing LC-HRMS data as expansion to include GC-HRMS capability requires a sufficiently large public GC-HRMS library to build a neutral loss model. At present the online crowd-sourced libraries (e.g., MassBank EU) contain insufficient data to build such a model.

As a platform for the early warning of emerging chemical threats, key developments will focus on statistical analysis of the data to allow trends of interest to be explored such as the emergence of new chemicals/features in multiple geographic locations. As *InSpectra* was built using a relational database, it is expected that such statistical packages will

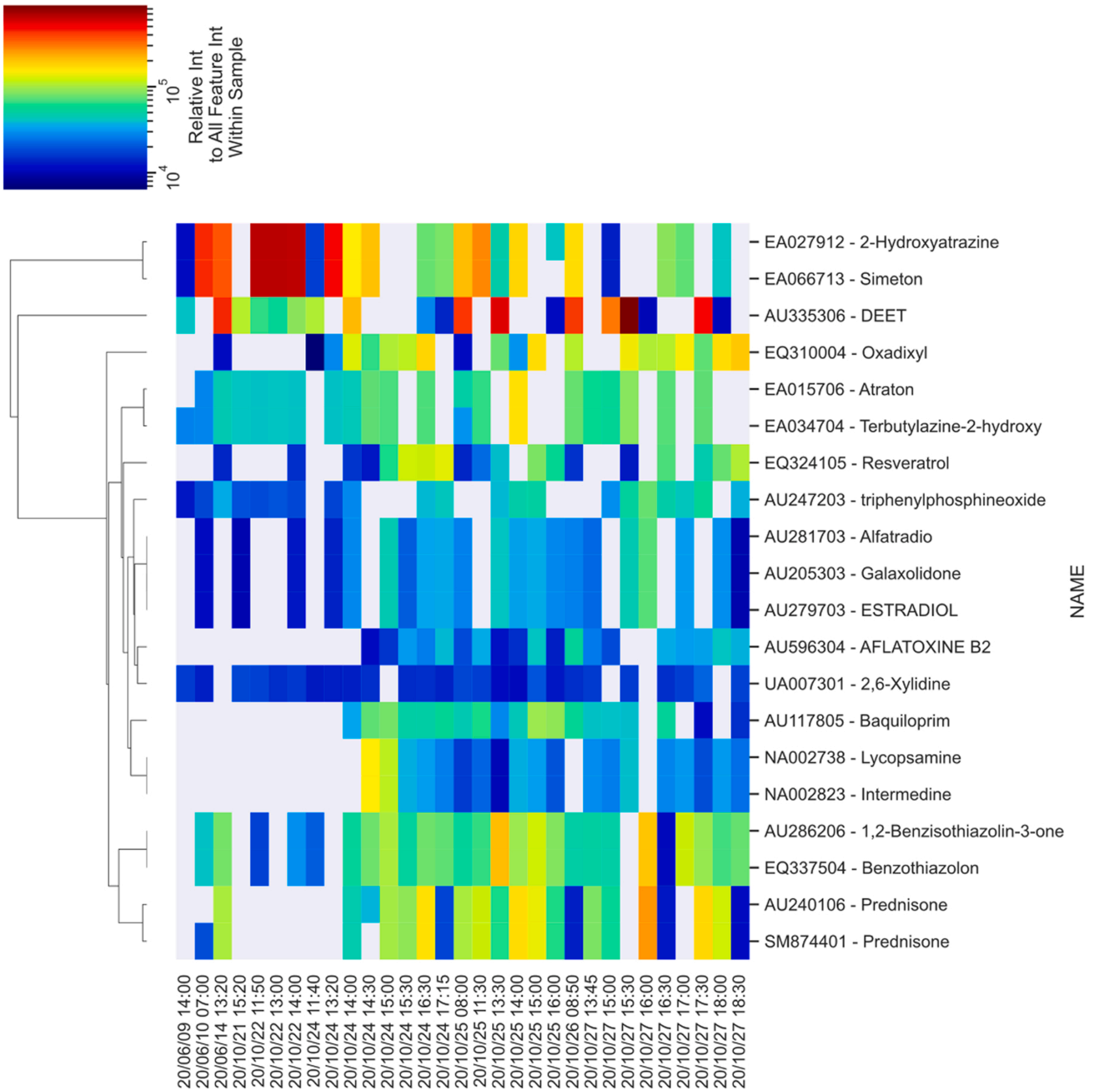


Fig. 8. Temporal Analysis of relative intensities (relative to all intensities in the sample set) of the most prevalent most prevalent tentative chemical identifications within stormwater using the Library Search workflow. The chemicals are ordered based on temporal trend similarity using hierarchical clustering.

easily be implemented into the platform.

The capabilities of *InSpectra* could be expanded to metabolomic and compound discoveries. Feature detection, componentisation and library search outputs could be effortlessly combined to perform a molecular networking analysis to identify and visualise molecules with similar spectral data. This can be achieved by pair wise comparison of componentised MS^2 spectral data using a spectral alignment algorithm and create a network of spectral relations. The metadata (sample information) such as type of sample, spatial and temporal information can be incorporated into the molecular network to facilitate the data analysis and interpretation. Currently, only the Global Natural Products Social Molecular Networking (GNPS) platform offers such analysis as a complete workflow. *InSpectra* could provide a couple of advantages in the identification of environmental compounds compared to GNPS due to its ability to search against more up to date reference libraries (MassBank and *theoretical* spectra from EPA's DSSTox database) and its capacity to archive all the data, including metadata.

With this paper we invite collaboration with research teams across all disciplines to trial *InSpectra*. Researchers who want their HRMS data to be analysed may contact us to do so. The plan of *InSpectra* is to have a website dedicated for all interested parties to have access to upload and process their files independently, however, because of our limited resources, only the backend workflows and maintenance of them are currently working. By showing the current capabilities and potential of *InSpectra*, we hope to have further evidence for the necessity of *InSpectra* and help us secure the resources to have a fully supported platform.

Statement of Environmental Implication

While chemicals continue to improve quality of life, chemical pollution can cause detrimental effects to the environment. Determining which chemicals are responsible however is challenging and the difficulty for regulators is the lack of sufficient experimental evidence between chemical exposures and effects. Non-target analysis employing high-resolution mass spectrometry (HRMS) is increasingly being used to identify chemicals of biological relevance; but these datasets are large and complex. Here we present an open-source/access platform for processing and archiving HRMS data to transparently identify chemicals with a future focus of sharing and community curation as an early warning system for emerging chemical threats.

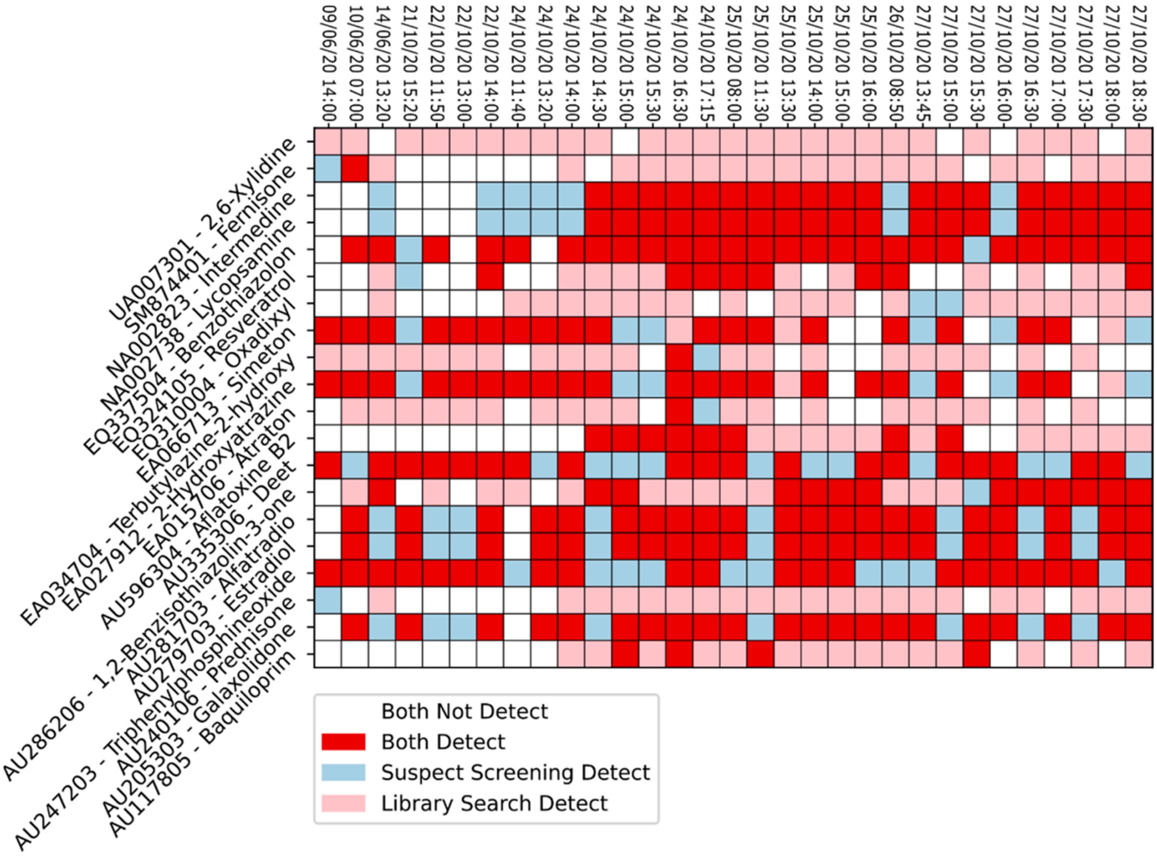


Fig. 9. Comparison of tentative chemical detections for the Library Search and Suspect Screening workflows on select suspect analytes found in stormwater temporal samples. 241 cases were detected by both workflows as indicated by the red squares. 116 cases were detected by neither workflow as indicated by the white squares, 181 cases were only identified using the Library Search workflow as indicated by the light red squares, and 82 cases where only the Suspect Screening workflow detected the suspect chemicals as indicated by the blue squares.

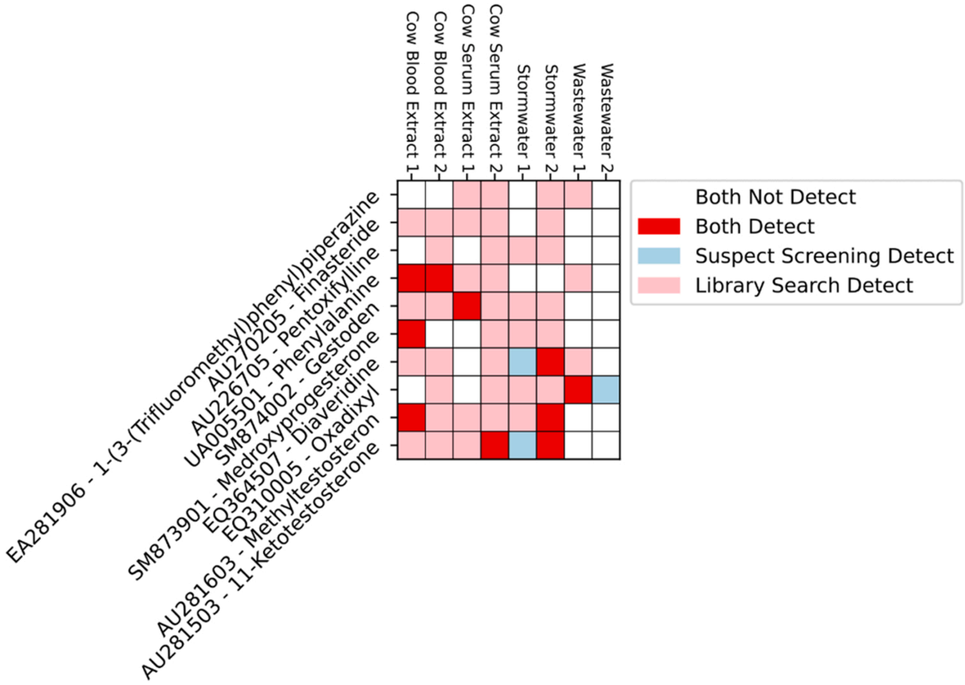


Fig. 10. Comparison of tentative chemical detections for the Library Search and Suspect Screening workflows on select suspect analytes found in different matrices (wastewater, storm water, cow blood and serum extracts). There were 10 cases where both workflows detected the suspects as indicated by the red squares, 39 where only the Library Search workflow resulted in detection as indicated by the light red squares, 2 cases where only the Suspect Screening workflow resulted in detection indicated by the blue squares and 19 cases where neither workflow resulted in a detection as indicated by the white squares.

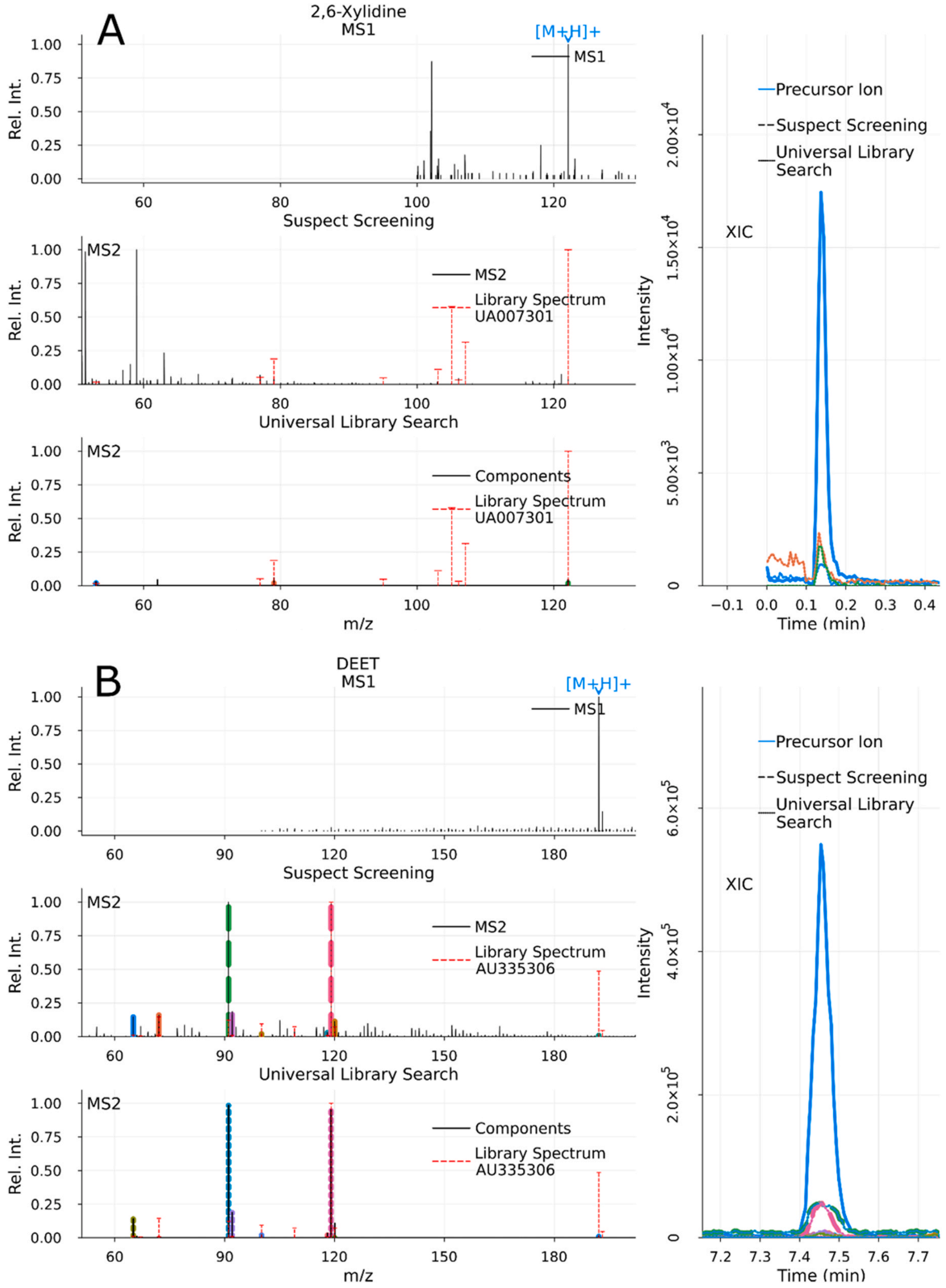


Fig. 11. Example comparisons of tentative identifications of both the Suspect Screening and Library Search workflows. The top panel shows the MS^1 spectrum at the apex of the precursor ion, the middle panel the MS^2 spectrum at the same time and the bottom panel the componentised fragments. The red dashed lines indicate the reference spectra for the matched accession and matched fragments are coloured for easier reference. The right panel shows the extracted ion chromatogram (XIC) for the precursor ion (solid blue line), matched Suspect Screening fragments as dashed lines, and matched Library Search components as dotted lines. Rel. Int. = relative intensity.

CRediT authorship contribution statement

Mathieu Feraud: Conceptualization, Methodology, Software, Writing - Original Draft, Writing – Review & Editing, Jake W. O'Brien: Conceptualization, Methodology, Software, Writing - Original Draft, Writing – Review & Editing, Saer Samanipour: Conceptualization, Methodology, Software, Writing - Original Draft, Writing – Review & Editing, Funding acquisition, Pradeep Dewapriya: Methodology, Writing - Original Draft, Writing – Review & Editing, Denice van Herwerden: Conceptualization, Methodology, Software, Writing - Original Draft, Writing – Review & Editing, Sarit Kaserzon: Writing - Original Draft, Writing – Review & Editing, Funding acquisition, Ian Wood: Methodology, Writing - Original Draft, Cassandra Rauert: Methodology, Writing - Original Draft, Writing – Review & Editing, Kevin V. Thomas: Conceptualization, Funding acquisition, Writing - Original Draft, Writing – Review & Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors gratefully acknowledge the financial support from the Australian Research Council (ARC) Discovery Project (DP190102476) and National Health and Medical Research Council (NHMRC) (2019/GNT1185347). J.W.O is the recipient of an NHMRC Emerging Leadership Fellowship (EL1 2009209). The Queensland Alliance for Environmental Health Sciences, The University of Queensland, would like to acknowledge the financial support of the Queensland Department of Health. S.S. acknowledges the University of Amsterdam Data Science Centre for financial support (project Edified). The authors wish to acknowledge Li Zhe for providing Orbitrap data and Samual MacDonald for early platform development.

The NTA Study Reporting Tool (SRT) was used during peer review to document and improve the reporting and transparency of this study (10.1021/acs.analchem.1c02621; 10.6084/m9.figshare.19763482 [PDF] or 10.6084/m9.figshare.19763503 [Excel]).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jhazmat.2023.131486.

References

[1] O. World Health, The public health impact of chemicals: knowns and unknowns, in, World Health Organization, Geneva, 2016.

[2] Pleil, J.D., 2012. Categorizing biomarkers of the human exposome and developing metrics for assessing environmental sustainability. J Toxicol Environ Health B Crit Rev 15, 264–280.

[3] Kortenkamp, A., Faust, M., Scholze, M., Backhaus, T., 2007. Low-level exposure to multiple chemicals: reason for human health concerns? Environ Health Perspect 115 (Suppl 1), 106–114.

[4] Alygizakis, N.A., Oswald, P., Thomaidis, N.S., Schymanski, E.L., Aalizadeh, R., Schulze, T., et al., 2019. NORMAN digital sample freezing platform: a European virtual platform to exchange liquid chromatography high resolution-mass spectrometry data and screen suspects in "digitally frozen" environmental samples. TrAC Trends Anal Chem 115, 129–137.

[5] Muir, D.C.G., Howard, P.H., 2006. Are there other persistent organic pollutants? A challenge for environmental chemists. Environ Sci Technol 40, 7157–7166.

[6] S. Samanipour, J.W. Martin, M.H. Lamoree, M.J. Reid, K.V. Thomas, Letter to the Editor: Optimism for nontarget analysis in environmental chemistry, Environ Sci Technol, 53 (2019) 5529–5530.

[7] Hollender, J., Schymanski, E.L., Singer, H.P., Ferguson, P.L., 2017. Nontarget screening with high resolution mass spectrometry in the environment: ready to go? Environ Sci Technol 51, 11505–11512.

[8] Hernandez, F., Bakker, J., Bijlsma, L., de Boer, J., Botero-Coy, A.M., Bruinen de Bruin, Y., et al., 2019. The role of analytical chemistry in exposure science: Focus on the aquatic environment. Chemosphere 222, 564–583.

[9] Albergamo, V., Schollee, J.E., Schymanski, E.L., Helmus, R., Timmer, H., Hollender, J., et al., 2019. Nontarget screening reveals time trends of polar micropollutants in a riverbank filtration system. Environ Sci Technol 53, 7584–7594.

[10] Chiaia-Hernandez, A.C., Gunthardt, B.F., Frey, M.P., Hollender, J., 2017. Unravelling contaminants in the anthropocene using statistical analysis of liquid chromatography-high-resolution mass spectrometry nontarget screening data recorded in lake sediments. Environ Sci Technol 51, 12547–12556.

[11] Sjerps, R.M.A., Vughs, D., van Leerdam, J.A., Ter Laak, T.L., van Wezel, A.P., 2016. Data-driven prioritization of chemicals for various water types using suspect screening LC-HRMS. Water Res 93, 254–264.

[12] Chiaia-Hernandez, A.C., Schymanski, E.L., Kumar, P., Singer, H.P., Hollender, J., 2014. Suspect and nontarget screening approaches to identify organic contaminant records in lake sediments. Anal Bioanal Chem 406, 7323–7335.

[13] Alygizakis, N.A., Samanipour, S., Hollender, J., Ibanez, M., Kaserzon, S., Kokkali, V., et al., 2018. Exploring the potential of a global emerging contaminant early warning network through the use of retrospective suspect screening with high-resolution mass spectrometry. Environ Sci Technol 52, 5135–5144.

[14] Bouslimani, A., Sanchez, L.M., Garg, N., Dorrestein, P.C., 2014. Mass spectrometry of natural products: current, emerging and future technologies. Nat Prod Rep 31, 718–729.

[15] M. Wang , J.J. Carver , V.V. Phelan , L.M. Sanchez , N. Garg , Y. Peng, et al. , Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking, Nat Biotechnol, 34 (2016) 828–837.

[16] Schulze, B., Jeon, Y., Kaserzon, S., Heffernan, A.L., Dewapriya, P., O'Brien, J., et al., 2020. An assessment of quality assurance/quality control efforts in high resolution mass spectrometry non-target workflows for analysis of environmental samples. TrAC Trends Anal Chem 133.

[17] Black, G., Lowe, C., Anumol, T., Bade, J., Favela, K., Feng, Y.-L., et al., 2023. Exploring chemical space in non-targeted analysis: a proposed ChemSpace tool. Anal Bioanal Chem 415, 35–44.

[18] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. In: Sci. Data, 3, 160018.

[19] Peters, K., Bradbury, J., Bergmann, S., Capuccini, M., Cascante, M., de Atauri, P., et al., 2019. PhenoMeNal: processing and analysis of metabolomics data in the cloud. Gigascience 8.

[20] Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., et al., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 46, W537–W544.

[21] Pluskal, T., Castillo, S., Villar-Briones, A., Orešič, M., 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinforma 11, 395.

[22] Tsugawa, H., Ikeda, K., Takahashi, M., Satoh, A., Mori, Y., Uchino, H., et al., 2020. A lipidome atlas in MS-DIAL 4. Nat Biotechnol 38, 1159–1163.

[23] Helmus, R., Ter Laak, T.L., van Wezel, A.P., de Voogt, P., Schymanski, E.L., 2021. patRoon: open source software platform for environmental mass spectrometry based non-target screening. J Chemin 13, 1.

[24] Shen, X., Yan, H., Wang, C., Gao, P., Johnson, C.H., Snyder, M.P., 2022. TidyMass an object-oriented reproducible analysis framework for LC–MS data, Nature. Communications 13, 4365.

[25] M. Loos, enviMass version 3.5 LC-HRMS trend detection workflow—R package, in, 2018.

[26] M. Loos, enviPick: Peak Picking for High Resolution Mass Spectrometry Data, in, 2016.

[27] Rost, H.L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., et al., 2016. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods 13, 741–748.

[28] T. Letzel, Inaugural Presentation of the open-access platform FOR-IDENT, in: NORMAN Annual General Assembly Meeting, 2015.

[29] Tautenhahn, R., Patti, G.J., Rinehart, D., Siuzdak, G., 2012. XCMS Online: a web-based platform to process untargeted metabolomic data. Anal Chem 84, 5035–5039.

[30] Aron, A.T., Gentry, E.C., McPhail, K.L., Nothias, L.-F., Nothias-Esposito, M., Bouslimani, A., et al., 2020. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. Nat Protoc 15, 1954–1991.

[31] Rauert, C., Charlton, N., Okoffo, E.D., Stanton, R.S., Agua, A.R., Pirrung, M.C., et al., 2022. Concentrations of tire additive chemicals and tire road wear particles in an Australian urban tributary. Environ Sci Technol 56, 2421–2431.

[32] Nilsson, S., Mueller, J.F., Rotander, A., Braunig, J., 2021. Analytical uncertainties in a longitudinal study - a case study assessing serum levels of per- and poly-fluoroalkyl substances (PFAS). Int J Hyg Environ Health 238, 113860.

[33] O'Brien, J.W., Grant, S., Banks, A.P.W., Bruno, R., Carter, S., Choi, P.M., et al., 2018. A National Wastewater Monitoring Program for a better understanding of public health: a case study using the Australian census. Environ Int 122, 400–411.

[34] M.S. McLachlan, Z. Li, L. Jonsson, S. Kaserzon, J.W. O'Brien, J.F. Mueller, Removal of 293 organic compounds in 15 WWTPs studied with non-targeted suspect screening, Environmental Science: Water Research & Technology, (2022).

[35] Haddad, P.R., Taraji, M., Szücs, R., 2021. Prediction of analyte retention time in liquid chromatography. Anal Chem 93, 228–256.

[36] Samanipour, S., O'Brien, J.W., Reid, M.J., Thomas, K.V., 2019. Self adjusting algorithm for the nontargeted feature detection of high resolution mass spectrometry coupled with liquid chromatography profile data. Anal Chem 91, 10800–10807.

[37] Samanipour, S., Reid, M., Baek, K., Thomas, K.V., 2018. Combining a deconvolution and a universal library search algorithm for the non-target analysis of data independent LC-HRMS spectra. Environ Sci Technol.

[38] Pedrioli, P.G., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., et al., 2004. A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol 22, 1459–1466.

[39] Samanipour, S., Choi, P., O'Brien, J.W., Pirok, B.W.J., Reid, M.J., Thomas, K.V., 2021. From centroided to profile mode: machine learning for prediction of peak width in HRMS data. Anal Chem 93, 16562–16570.

[40] Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., et al., 2011. mzML–a community standard for mass spectrometry data. Mol Cell Proteom 10, R110.

[41] Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., et al., 2012. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30, 918–920.

[42] Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., Siuzdak, G., 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78, 779–787.

[43] Samanipour, S., O'Brien, J.W., Reid, M., Thomas, K.V., 2019. Self adjusting algorithm for the non-targeted feature detection of high resolution mass spectrometry coupled with liquid chromatography profile data. Anal Chem 91, 10800–10807.

[44] U.D.F. Lab, Mass Spectrometry Adduct Calculator, in, 2022.

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

[46] M. Project, High Quality Mass Spectra Database, in: N.A. MassBank Project, MassBank Consortium (Ed.), http://www.massbank.jp/, 2021.

[47] Williams, A.J., Grulke, C.M., Edwards, J., McEachran, A.D., Mansouri, K., Baker, N. C., et al., 2017. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J Chemin 9, 61.

[48] Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., et al., 2010. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 45, 703–714.

[49] 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, W94–99.

[50] Allen, F., Pon, A., Greiner, R., Wishart, D., 2016. Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification. Anal Chem 88, 7689–7697.

[51] J. Boelrijk, S. Samanipour, D. Van Herwerden, B. Ensing, P. Forré, Predicting RP-LC retention indices of structurally unknown chemicals from mass spectrometry data, in, American Chemical Society (ACS), 2022.

[52] Chen, C.J., Lee, D.Y., Yu, J., Lin, Y.N., Lin, T.M., 2022. Recent advances in LC-MS-based metabolomics for clinical biomarker discovery. Mass Spectrom Rev, e21785.

[53] Nothias, L.F., Petras, D., Schmid, R., Duhrkop, K., Rainer, J., Sarvepalli, A., et al., 2020. Feature-based molecular networking in the GNPS analysis environment. Nat Methods 17, 905–908.