Electronic Supplementary Information (SI)

Connecting Environmental Exposure and Neurodegeneration using Cheminformatics and High Resolution Mass Spectrometry: Potential and Challenges

Emma L. Schymanski[1,*], Nancy C. Baker[2], Antony J. Williams[3], Randolph R. Singh[1,4], Jean-Pierre Trezzi[5,6], Paul Wilmes[6], Pierre L. Kolber[7,8], Rejko Kruger[7,8], Nicole Paczia[9], Carole L. Linster[9], Rudi Balling[10].

[1] Environmental Cheminformatics Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 avenue du Swing, L-4367 Belvaux, Luxembourg.
[2] Leidos, Research Triangle Park, North Carolina, USA.
[3] National Centre for Computational Toxicity (NCCT), United States Environmental Protection Agency, Research Triangle Park, North Carolina, USA.
[4] Oak Ridge Institute for Science and Education Research Fellow, United States Environmental Protection Agency, Research Triangle Park, North Carolina, USA.
[5] Integrated Biobank of Luxembourg, Luxembourg Institute of Health, 1 rue Louis Rech, L-3555 Dudelange, Luxembourg.
[6] Eco-Systems Biology Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 avenue du Swing, L-4367 Belvaux, Luxembourg.
[7] Clinical and Experimental Neuroscience Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 avenue du Swing, L-4367 Belvaux, Luxembourg.
[8] Neurology, Centre Hospitalier de Luxembourg, Luxembourg City, Luxembourg.
[9] Enzymology and Metabolism Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 avenue du Swing, L-4367 Belvaux, Luxembourg.
[10] Systems Biomedicine, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 avenue du Swing, L-4367 Belvaux, Luxembourg.

*Corresponding author: emma.schymanski@uni.lu

This supplementary information document contains passages of text to compliment the perspectives article, primarily describing methods and resources mentioned in the main article in greater detail. We note that we have used SI to abbreviate the Electronic Supplementary Material instead of the journal term “ESI” to avoid confusion with the ionisation method. The contents are as follows:

The CompTox Chemicals Dashboard

The US EPA's CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard) is one example of a web-based application developed with the intention of providing access to various types of data associated with environmental science.[1] The application provides access to data associated with ~875,000 chemical substances (May, 2019) including experimental and predicted physicochemical properties, environmental fate and transport data, *in vivo* hazard data, *in vitro* bioactivity data, and various types of exposure data. The application also provides access to a set of chemical lists (https://comptox.epa.gov/dashboard/chemical_lists) associated with, and generally linking to, peer-reviewed publications or data assembled from other publicly accessible online databases (see examples above and further explanations below). A number of search capabilities exist, including the ability to search for chemicals based on chemical identifier strings or substrings based on systematic names, synonyms, Chemical Abstract Services (CAS) registry numbers (CASRN) and InChI Keys. Searching by product use and category is available so that chemicals associated with a particular product use can be identified (*e.g.* chemicals in cigarettes). Searching by gene and assay associated with the ToxCast and Tox21 high throughput screening programs is also available.[2] The identification of potential toxicants via mass spectrometry is supported via advanced and batch search functions, including mass and formula-based searches that also take advantage of "MS-Ready" preparations of the data.[3,4] While the details of the underlying databases and cheminformatics developments have been fully described elsewhere,[1] updates to both the data and the functionality of the dashboard are on a biannual release cycle and new functionality has been introduced since the initial publication. The metadata and MS-ready functionality of the CompTox Chemicals Dashboard has been integrated into the *in silico* fragmentation identification approach MetFrag,[3,5] to allow users access to this rich source of environmentally-relevant metadata within computational non-target identification efforts. This is available via the MetFrag web interface (https://msbi.ipb-halle.de/MetFrag/) under the CompTox "Local Database" options. The metadata options implemented to date include overall literature counts, suspect list tagging and selected predicted properties.

Compiling Neurotoxicity-relevant Suspect Lists

A number of lists of neurotoxicants were compiled for the purposes of this perspective, summarized in Table 1 in the main article. These include small, carefully validated lists explained extensively in the source publications (DNTEFFECTS,[6] DNTINVIVO[7] and HUMANNEUROTOX[8]), where few suspect hits will ever be found in environmental samples. The DNTPOTNEG[7] list contains potential negative controls for neurotoxicity which, if found, should not be associated with neurotoxic effects. Two larger lists (NEUROTOXINS[9] and LITMINEDNEURO[10]) are described further below. Note that here the term "(neuro)toxicant" is being used in the broad sense, as toxin refers to toxicants of natural origin only; here "neurotoxin" is only used where this was the actual word used to perform the data mining. For the smaller lists mentioned above, the referenced table from each publication was mapped to the Dashboard identifier DTXSIDs using Name and, where available, CASRN and any additional information (*e.g.* additional synonyms) to verify the match. These DTXSIDs were then used to compile the lists for the Chemical Lists page (https://comptox.epa.gov/dashboard/chemical_lists).

The NEUROTOXINS list contains chemicals reported as neurotoxins that was compiled from public resources including ChEBI (https://www.ebi.ac.uk/chebi/), Wikipedia (from source pages https://en.wikipedia.org/wiki/Neurotoxin and https://en.wikipedia.org/wiki/Category:Neurotoxins), T3DB (http://t3db.ca/ - entries tagged as neurotoxin) and various literature resources, detailed further in a file deposited to Zenodo[11] along with all lists mentioned in this article.

Text Mining of PubMed to Create the LITMINEDNEURO List

One of the largest sources of information describing the effect of chemicals on biological systems, including the nervous system, is the biomedical literature in PubMed (https://www.nlm.nih.gov/bsd/pubmed.html). Most PubMed articles have been annotated with terms from the Medical Subject Headings (MeSH) vocabulary (https://meshb.nlm.nih.gov/search). In earlier work, Baker and Hemminger[12] reported a method to extract the MeSH terms for chemicals and for diseases from each PubMed MEDLINE record and store these records in a database. To identify chemicals annotated with, and thus potentially contributing to, neurotoxicity, this literature database of MeSH terms was queried to retrieve all articles in which a nervous system disease was annotated as being caused by a chemical by looking for the disease subheading "chemically induced". Nervous system diseases were identified through a look-up in the MeSH tree (https://meshb.nlm.nih.gov/treeView) for terms in the family of diseases belonging to the MeSH node C10 (https://id.nlm.nih.gov/mesh/describe?uri=http://id.nlm.nih.gov/mesh/C10). Then, articles were searched for chemical MeSH terms that were the main topic of the article and annotated with a subheading that was either "toxicity", "poisoning", or "adverse effects". To focus the list on substances that could be identified through small molecule mass spectrometry, several categories of substances were subsequently omitted: proteins, mixtures, terms identifying families of chemicals rather than specific chemicals, and biological factors. These steps eliminated entries such as "Particulate Matter", "Contrast Media", and "Pertussis Vaccine". To ensure that the relationships were robust and more likely significant, chemicals with fewer than five articles connecting them to nervous system diseases were also omitted. All chemicals not yet in the Dashboard that satisfied these criteria were registered, to allow a complete listing.

The output of this processing was exported to Microsoft® Excel and is included in the SI and on FigShare[13]. The CASRN and the CompTox Chemicals Dashboard substance identifiers (DTXSID) were included in the spreadsheet for chemicals for which this relationship was captured and associated with MeSH identifiers. The overview tab of this workbook contains 1,250 chemicals (1,243 unique DTXSIDs) and the co-annotations with 554 nervous system diseases in over 53,000 chemical – disease pairs ("Detail" tab). These relationships were described in 38,192 articles. A batch search of the Dashboard by DTXSID will return all related chemical information needed for generating suspect lists with subject-specific reference scores for disease or effect subsets of this list. Note that MeSH terms for chemicals are often more general than CASRNs. For instance, one MeSH term may refer to multiple CASRNs and this is a key challenge to mapping literature and exact chemical data as discussed above. For example, the MeSH term "Propylene Glycol" maps to six CASRNs in the MeSH dictionary (4254-14-2, 4254-15-3, 4254-16-4, 57-55-6, 58858-91-6, 63180-48-3). The Dashboard has one record for Propylene Glycol (DTXSID0021206) with the CASRN 57-55-6 and lists the old deleted CASRN 4254-16-4 as a synonym (all synonyms listed here). The remaining CASRN belong to the R- and S- isomers, the sodium salt and the monohydrate, all of which match to (and would be observed as) the "MS-ready" form DTXSID0021206 in NT-HR-MS.

Taking advantage of the volume of data (number of occurrences) can help minimise the effects of the noise. Ranking the chemical list by article count is a useful method to bring the chemicals most commonly associated with neurological disorders to the top of the list and relegating the chemicals with the weakest or erroneous signals to the bottom. As mentioned above, all chemicals mentioned in 5 articles or more were selected for the LITMINEDNEURO list.

Related Structures: Mapping "Chemicals" and "Substances"

Reviewing some of the lists in Table 1 via the Dashboard reveals several interesting cases where new cheminformatics approaches are needed and already under development to capture less well-defined substance information in the form of discrete chemical structures suitable for NT-HR-MS screening studies. Firstly, not all chemical substances have distinct chemical structure representations. Chemicals of unknown or variable composition, complex reaction products and biological materials (UVCBs) may be displayed simply as a chemical record with a name and, where available, a CASRN or also as Markush structure representations[14]. Examples relevant to potential neurotoxicants (all from the HUMANNEUROTOX list) include the Markush form of xylenes and the records for the poorly defined Metaldehyde and the generic category "Arsenic and arsenic compounds".

The following text includes examples taken from the lists provided in Table 1 in the main text, with selected structures given in Figure 1 of the main text. The "MS-ready" approach applied on the Dashboard, for instance, associates mixtures with their constituent chemicals, removes salts, and crosslinks this association in a way to retain both substance-specific and individual component information.[3] The mixtures, components and neutralised forms of nicotine are, for example, available here: https://comptox.epa.gov/dashboard/dsstoxdb/mixture_search?cid=930. The individual components and the metadata associated with the substances can then be used in candidate ranking during NT-HR-MS, for instance using MetFrag[3,5] shown in the top row of Figure 1. In this figure, nicotine is to the left in the red box; the blue boxes contain three alternative forms, with the number of active hit calls/number of assays in which the chemical was analyzed as part of the TOXCAST[2] project, and the number of associated data sources. A further strategy applied is linking and displaying chemicals via the "related structures" tab. Examples include the display of transformation products associated with a chemical (*e.g.* Diazinon, middle row of Figure 1 and full listing here), or classes of chemicals for instance polychlorinated biphenyls (PCBs), see last row of Figure 1 (main text).

At this stage the majority of related substance mappings displayed under the "related substances" tab are compiled manually, but the ability to enumerate Markush structure representations into their individual chemical structures has also been integrated into the registration database underpinning the Dashboard for future expansions. Automation and future-proofing these developments will be critical to upscaling the current approaches to high throughput NT-HR-MS data analysis. Currently, users can download all related structures and thus screen all (available) chemicals for that class, to get an overview of currently-linked information. The download file contains basic chemical information, the relationship definition (*e.g.* predecessor, Markush child) as well as selected metadata (currently including data sources, number of associated PubMed articles, number of associated PubChem data sources, % active assay results in ToxCast[2] and the consumer product database (CPDAT) count[15]). An alternative approach to communicate relationships between chemicals on the Dashboard is to create a list containing members of a discrete class of chemicals, *e.g.* a list containing all PCBs (https://comptox.epa.gov/dashboard/chemical_lists/pcbchemicals). Identifier substring searches *via* the home page of the Dashboard may also help capture some "classes" of chemicals that are not strictly related in a registered chemical mixture, but may co-occur – *e.g.* "phthalates" or "statin" compounds. However, any substring search of this nature is also prone to returning non-related compounds for a class also. For example, searching on the substring "statin" also returns three isotopes of the element A*statin*e as well as Titanium Dioxide since one of the associated synonyms is "Ho*statin*t White R 30". In all cases, therefore, manual inspection of the search results is required. A challenge for typical small molecule analysis in NT-HR-MS is the presence of entries such as "Arsenic and arsenic containing compounds" on many lists. Specific search functionality is being developed to allow users to find all entries in the Dashboard that, for instance, contain specific elements and substructures. For now, a substring search of "arsenic" will capture 116 compounds (12 June 2019), while a search for Arsenic in the molecular formula gives over a thousand hits as not all arsenic containing compounds

contain arsenic in the name (*e.g.* Arsinous cyanide). As some organometallic compounds can be detected with LC coupled with HR-MS, as evidenced in the recent EPA Non-Targeted Analysis Collaborative Trial (ENTACT),[16] this will become increasingly relevant to clarify neurotoxicity and other effects that may be related to the presence of metals. Computational methods for identification will need to be updated to be able to handle metal-containing molecules, as recently done for RMassBank[17] for use during the ENTACT, and the ability to deal with covalently bonded metals (as opposed to ionic salts) was considered in the preparation of MS-Ready structures.[3]

One of the most interesting challenges associated with capturing chemical information for NT-HR-MS related to neurotoxicity (or any potential health impact) is relating what we as humans consume relative to discrete chemical components. For instance, while toxicity testing is most commonly performed on a chemical such as caffeine, patients will instead report coffee (or tea or energy drink) consumption. Coffee, for instance, is documented to contain over 1000 chemical constituents[18], which require a variety of analytical techniques for detection[19]. Nicotine is another common example where the chemical tested is often nicotine (or an associated salt or mixture), but patients will instead report *e.g.* smoking habits. It is well known that nicotine is not the only chemical in cigarettes that may cause detrimental health effects. Several thousand chemicals have been identified in cigarettes[20,21], with 599 chemicals listed as additives[22]. Capturing such knowledge (*e.g. via* cross-mapping and adding as lists or related substances in databases) will be increasingly important to help reconcile NT-HR-MS results in the future, yet expand suspect lists even further. As mentioned above, however, delving into the literature associated with these chemical entries will also be critical, as some documented neurotoxicants can indeed be associated with a neuroprotective action in the context of certain diseases (see discussion above). Some of the literature studies already take into account the question of active ingredients.[23]

Once these chemicals and their relationships are captured, a range of metadata can be included to assist in candidate prioritisation and selection for NT-HR-MS. This is demonstrated in the example shown in Figure 2 (main text). Candidates were exported as a MetFrag Input File (CSV) from the MS-ready formula search for "$C_{10}H_{14}N_2$" in the batch search of the CompTox Dashboard database. Many scoring terms can be selected; for this example the terms selected were "presence in the HUMANNEUROTOX list", PubChem Data Sources and TOXCAST % Active bioassays. The resulting CSV was uploaded as a local CSV file and searched by formula in MetFrag (https://msbi.ipb-halle.de/MetFragBeta, v2.0.19) using a spectrum of nicotine taken from MassBank (EQ300804). To demonstrate how metadata can be used for quick prioritisation, 5 scoring terms were used in total. Two of these are based on experimental information, the MetFrag score, *i.e.* the predicted fragmentation assignment and the Exact Spectral Similarity score, which compares the experimental spectrum to all entries for exactly that substance in open libraries. The three additional terms described above ("presence in HUMANNEUROTOX list", PubChem Data Sources and TOXCAST % Active bioassays) fall under what we refer to as "metadata" in this article. These terms indicate whether the substance may be of interest in the (here hypothetical) experimental context. All scores are scaled between 0 and 1 and added to form the final score (blue dots in Figure 2, main text). The presence of one candidate with a total score well above 2 indicates that for this query, a potential neurotoxicant of interest (in this case nicotine) matches both experimental and metadata and may be a promising hit for further validation. Scaled up to the hundreds of masses typically present in non-target identification efforts, the quick selection of top candidates with high overall scores can enable rapid prioritisation, even in the absence of fragmentation information, as metadata scores will still indicate possible candidates of interest. This can, for instance, enable the quick selection of interesting masses for creation of inclusion lists for re-acquisition of fragmentation data.[24] Instead of taking the bulk PUBCHEM_DATA_SOURCES scores, as demonstrated here, it would be possible to use the chemical information and specific literature counts in the excel macro included in the supplemental information

to create candidate files and scores for neurological endpoints of interest. This example is also presented online.[25]

Chemical Properties and Analytical Methods

The chemical diversity of potential neurotoxicants poses a challenge in terms of the analysis of their presence in different matrices. The mapping of potential neurotoxicants from public resources (https://comptox.epa.gov/dashboard/chemical_lists/neurotoxins) resulted in a list of over 500 chemicals with varying physicochemical properties. Boxplots (see Figure 4) were generated using the predicted properties available in the Dashboard to demonstrate the diversity in these chemical properties. These properties included the octanol-water partition constant (log K_{ow}), water solubility, vapor pressure, and boiling point, as well as properties that assist in understanding how these chemicals may behave in the environment (moving from source to sink) such as bioconcentration factor and biodegradation.

Understanding environmental fate specifically will help guide the experimenter in deciding which matrices may be of importance to study, in addition to providing knowledge regarding how these chemicals may be absorbed, distributed, metabolised, and excreted (ADME). It is apparent from Figure 4 that there is a huge variability in the properties of the suspected neurotoxicants and this makes experimental design even more challenging. The bioconcentration factor, log K_{ow}, and water solubility plots show that most of the potential neurotoxicants have a higher tendency to accumulate in the human body in addition to a lower susceptibility to be eliminated through urine. Unless metabolised to form a more polar conjugate, these compounds can be expected to be present in the body and accumulate over time. The log K_{ow} and water solubility can also be used to decide which analytical approach is most appropriate. Generally speaking, moderately polar to nonpolar molecules are amenable to chromatography, however the molecules of most interest will determine choices such as liquid or gas chromatography, which solvent(s), separation technique, ionisation mode and polarity to employ. The following sections give a brief overview of different scenarios to identify documented ("known") potential neurotoxicants and discover new ones, more comprehensive overviews (beyond the focus on neurotoxicants) are available elsewhere.[26,27]

Target versus Non-target Mass Spectrometry

Analysis by mass spectrometry is generally the method of choice for identification and, in the case of target compounds, quantification, in complex sample matrices. Targeted analytical methods are ideal for rapid quantification of known (target) compounds; however, this is not the focus of the current article. The following material will concentrate on how to find potential neurotoxicants with NT-HR-MS. This generally entails either liquid or gas chromatography coupled with high resolution tandem mass spectrometry (although other methods are also possible). Tandem mass spectrometry requires a combination of precursor ion, fragment ion(s), and retention time information to verify the presence and concentration of different molecules. Tandem mass spectrometry generally achieves high selectivity and sensitivities and a linear dynamic range spanning a few orders of magnitude. Low resolution gas chromatography mass spectrometry approaches, while less sensitive, still offer advantages in certain cases due to its well-established nature, such as the comprehensive nature of spectral libraries (over 1 million compounds compared with tens of thousands of compounds in tandem mass spectrometry libraries)[28] but is not the focus here. The ability to quickly identify and quantify known compounds helps scientists to form dose-response relationships and provide data to support risk assessments and establish toxicity levels. For exploratory studies, however, NT-HR-MS is ideal as an initial step. Without analytical standards, it can be challenging to discriminate between isobaric compounds. Furthermore, low abundance features may be lost in the noise region. Factored

together, these limitations may end up preventing the discovery of toxicologically significant compounds in a complex matrix. However, recent advances in mass spectrometry have made it possible to overcome these challenges to a certain extent, as the sensitivity and mass resolving power of the instruments have improved over the last two decades. Moreover, new separation techniques and ionisation sources have made it possible to expand the breadth of chemical space that can be covered if the different techniques are used side by side. The following text discusses this in the light of neurotoxicant analysis and discovery.

Sample "Pre-Processing"

Sample preparation and knowledge of the matrix plays a vital role in the discovery of new neurotoxicants as it dictates recovery or loss of the analytes of interest. As no single method is sufficient to capture all potential neurotoxicants, the best way to explore the presence of neurotoxicants would be to perform a categorical extraction with the intent of initial (non-target) identification using HR-MS and subsequent quantification using targeted methods. Categorical extraction builds on the knowledge that known/suspected neurotoxicants have very different properties, as demonstrated above. Categorical extraction is accomplished by creating theoretical compartments based on physicochemical properties: volatiles, nonpolar compounds, polar compounds, and metal containing compounds. Analysis of volatile components can be accomplished by using gas chromatography (GC). GC analysis with electron ionisation (EI) is optimal for compounds that are nonpolar but do not easily acquire or lose a hydrogen in the gas phase. Although EI is a harder ionisation than atmospheric pressure ionisation (API) techniques, the additional fragmentation provides a "fingerprint" that can be searched against spectral libraries such as NIST.[29] Accurate mass GC-EI methods are slowly becoming available. One lesson that can be learned from the list of potential neurotoxicants is that the lower mass range must be used to accommodate low mass components (*e.g.* acrylamide and acrylonitrile); almost 60 compounds in the potential neurotoxicant list of 511 have masses below 100 (see Figure 4). While small components are often GC-amenable, care must be taken to prevent analyte loss by volatilisation (hence rapid analysis is desirable). Employing solid phase microextraction or headspace analysis followed by GC-HR-MS analysis would be two approaches to overcome this challenge.[30,31] Extraction using non-polar solvents such as hexane, isooctane or ethyl acetate would also enable the recovery of nonpolar components from the matrix that have similar properties to the following potential neurotoxicant classes: pyrethroid pesticides, polyhalogenated biphenyls, and diphenyl ethers. Due to the efficiency of separation brought by GC, improved separation of analytes is achieved especially in cases where multiple congeners of an analyte class may be present. Further chromatographic selectivity may be achieved by employing multidimensional GC[32].

References

1 A. J. Williams, C. M. Grulke, J. Edwards, A. D. McEachran, K. Mansouri, N. C. Baker, G. Patlewicz, I. Shah, J. F. Wambaugh, R. S. Judson and A. M. Richard, The CompTox Chemistry Dashboard: A community data resource for environmental chemistry, *J. Cheminformatics*, 2017, **9**, 61.

2 A. M. Richard, R. S. Judson, K. A. Houck, C. M. Grulke, P. Volarath, I. Thillainadarajah, C. Yang, J. Rathman, M. T. Martin, J. F. Wambaugh, T. B. Knudsen, J. Kancherla, K. Mansouri, G. Patlewicz, A. J. Williams, S. B. Little, K. M. Crofton and R. S. Thomas, ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology, *Chem. Res. Toxicol.*, 2016, **29**, 1225–1251.

3 A. D. McEachran, K. Mansouri, C. Grulke, E. L. Schymanski, C. Ruttkies and A. J. Williams, "MS-Ready" structures for non-targeted high-resolution mass spectrometry screening studies, *J. Cheminformatics*, 2018, **10**, 45.

4 E. L. Schymanski and A. J. Williams, Open Science for Identifying "Known Unknown" Chemicals, *Environ. Sci. Technol.*, 2017, **51**, 5357–5359.

5 C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender and S. Neumann, MetFrag Relaunched: Incorporating strategies beyond in silico fragmentation, *J. Cheminformatics*, 2016, **8**, 3.
6 W. R. Mundy, S. Padilla, J. M. Breier, K. M. Crofton, M. E. Gilbert, D. W. Herr, K. F. Jensen, N. M. Radio, K. C. Raffaele, K. Schumacher, T. J. Shafer and J. Cowden, Expanding the test set: Chemicals with potential to disrupt mammalian brain development, *Neurotoxicol. Teratol.*, 2015, **52**, 25–35.
7 M. Aschner, S. Ceccatelli, M. Daneshian, E. Fritsche, N. Hasiwa, T. Hartung, H. Hogberg, M. Leist, A. Li, W. Mundy, S. Padilla, A. Piersma, A. Bal-Price, A. Seiler, R. Westerink, B. Zimmer and P. Lein, Reference compounds for alternative test methods to indicate developmental neurotoxicity (DNT) potential of chemicals: Example lists and criteria for their selection and use, *ALTEX*, 2017, **34**, 49–74.
8 P. Grandjean and P. Landrigan, Developmental neurotoxicity of industrial chemicals, *The Lancet*, 2006, **368**, 2167–2178.
9 N. C. Baker, E. L. Schymanski and A. J. Williams, S43 | NEUROTOXINS | Neurotoxicants Collection from Public Resources, *Zenodo*, 2019, DOI: 10.5281/zenodo.2656729.
10 N. C. Baker, E. L. Schymanski and A. J. Williams, S37 | LITMINEDNEURO | Neurotoxicants from literature mining PubMed, *Zenodo*, 2019, DOI: 10.5281/zenodo.3242298.
11 E. L. Schymanski, N. C. Baker and A. J. Williams, Suspect Lists for Neurotoxicant Screening, *Zenodo*, 2019, DOI: 10.5281/zenodo.3243472.
12 N. C. Baker and B. M. Hemminger, Mining connections between chemicals, proteins, and diseases extracted from Medline annotations, *J. Biomed. Inform.*, 2010, **43**, 510–519.
13 N. C. Baker, E. L. Schymanski and A. J. Williams, Literature Neurotoxicants: Excel Macro File, *FigShare*, DOI:10.23645/epacomptox.7334603.
14 J. M. Barnard, A comparison of different approaches to Markush structure handling, *J. Chem. Inf. Model.*, 1991, **31**, 64–68.
15 K. L. Dionisio, K. Phillips, P. S. Price, C. M. Grulke, A. J. Williams, D. Biryol, T. Hong and K. K. Isaacs, The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products, *Scientific Data*, 2018, **5**, 180125.
16 E. M. Ulrich, J. R. Sobus, C. M. Grulke, A. M. Richard, S. R. Newton, M. J. Strynar, K. Mansouri and A. J. Williams, EPA's non-targeted analysis collaborative trial (ENTACT): genesis, design, and initial findings, *Anal. Bioanal. Chem.*, **411**, 853–866.
17 M. A. Stravs, E. L. Schymanski, H. P. Singer and J. Hollender, Automatic recalibration and processing of tandem mass spectra using formula annotation: Recalibration and processing of MS/MS spectra, *J. Mass Spectrom.*, 2013, **48**, 89–99.
18 Clarke, R.J., *Coffee Volume 1 Chemistry*, Springer, New York, 2013.
19 M. Jeszka-Skowron, A. Zgoła-Grześkowiak and T. Grześkowiak, Analytical methods applied for the characterization and the determination of bioactive compounds in coffee, *Eur. Food Res. Technol.*, 2015, **240**, 19–31.
20 R. Talhout, T. Schulz, E. Florek, J. van Benthem, P. Wester and A. Opperhuizen, Hazardous compounds in tobacco smoke, *Int. J. Environ. Res. Public Health*, 2011, **8**, 613–628.
21 C. Baumung, J. Rehm, H. Franke and D. W. Lachenmeier, Comparative risk assessment of tobacco smoke constituents using the margin of exposure approach: the neglected contribution of nicotine, *Sci. Rep.-UK*, 2016, **6**, 35577.
22 Tri-County Cessation Centre, Cigarette Ingredients, https://web.archive.org/web/20160121165220/http:/www.tricountycessation.org/tobaccofacts/Cigarette-Ingredients.html#list, (accessed 9 June 2019).
23 A. Ascherio, S. M. Zhang, M. A. Hernán, I. Kawachi, G. A. Colditz, F. E. Speizer and W. C. Willett, Prospective study of caffeine consumption and risk of Parkinson's disease in men and women, *Ann. Neurol.*, 2001, **50**, 56–63.
24 C. D. Broeckling, E. Hoyes, K. Richardson, J. M. Brown and J. E. Prenni, Comprehensive Tandem-Mass-Spectrometry Coverage of Complex Samples Enabled by Data-Set-Dependent Acquisition, *Anal. Chem.*, 2018, **90**, 8020–8027.
25 E. L. Schymanski and A. J. Williams, Community Resources Connecting Chemistry and Toxicity Knowledge to Environmental Observations, *Zenodo,* 2019, DOI: 10.5281/zenodo.3242674.

26 J. Hollender, E. L. Schymanski, H. P. Singer and P. L. Ferguson, Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go?, *Environ. Sci. Technol.*, 2017, **51**, 11505–11512.
27 W. Brack, S. Ait-Aissa, R. M. Burgess, W. Busch, N. Creusot, C. Di Paolo, B. I. Escher, L. Mark Hewitt, K. Hilscherova, J. Hollender, H. Hollert, W. Jonker, J. Kool, M. Lamoree, M. Muschket, S. Neumann, P. Rostkowski, C. Ruttkies, J. Schollee, E. L. Schymanski, T. Schulze, T.-B. Seiler, A. J. Tindall, G. De Aragão Umbuzeiro, B. Vrana and M. Krauss, Effect-directed analysis supporting monitoring of aquatic environments — An in-depth overview, *Sci. Total Environ.*, 2016, **544**, 1073–1118.
28 B. Y. L. Peisl, E. L. Schymanski and P. Wilmes, Dark matter in host-microbiome metabolomics: Tackling the unknowns–A review, *Anal. Chim. Acta*, 2018, **1037**, 13–27.
29 S. Stein, Mass Spectral Reference Libraries: An Ever-Expanding Resource for Chemical Identification, *Anal. Chem.*, 2012, **84**, 7274–7282.
30 E. Gionfriddo, É. A. Souza-Silva and J. Pawliszyn, Headspace versus Direct Immersion Solid Phase Microextraction in Complex Matrixes: Investigation of Analyte Behavior in Multicomponent Mixtures, *Anal. Chem.*, 2015, **87**, 8448–8456.
31 I. Domínguez, F. J. Arrebola, R. Gavara, J. L. Martínez Vidal and A. G. Frenich, Automated and simultaneous determination of priority substances and polychlorinated biphenyls in wastewater using headspace solid phase microextraction and high resolution mass spectrometry, *Anal. Chim. Acta*, 2018, **1002**, 39–49.
32 C. Veenaas, A. Bignert, P. Liljelind and P. Haglund, Nontarget Screening and Time-Trend Analysis of Sewage Sludge Contaminants via Two-Dimensional Gas Chromatography–High Resolution Mass Spectrometry, *Environ. Sci. Technol.*, 2018, **52**, 7813–7822.