Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard

Andrew D. McEachran[1*], Jon R. Sobus[2], Antony J. Williams[3*]

1 Oak Ridge Institute for Science and Education (ORISE) Research Participation Program, U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park, North Carolina, USA 27711

2 National Exposure Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park, North Carolina, USA 27711

3 National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park, North Carolina, USA 27711

*Corresponding authors:
Andrew D. McEachran
mceachran.andrew@epa.gov
Phone: 1-919-541-3001
Fax: 1-919-541-1194

Antony J. Williams
williams.antony@epa.gov
Phone: 1-919-541-1033
Fax: 1-919-541-1194

Acknowledgements: The authors would like to thank Jim Little for graciously providing the dataset used in Little *et al* (2012). This work was supported in part by an appointment to the ORISE participant research program supported by an interagency agreement between the US EPA and DOE. This work has been internally reviewed at the US EPA and has been approved for publication. The views expressed in this paper are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

Compliance with Ethical Standards: The authors declare no potential conflicts of interest.

Abstract: Chemical features observed using high-resolution mass spectrometry can be tentatively identified using online chemical reference databases by searching molecular formulae and monoisotopic masses and then rank-ordering of the hits using appropriate relevance criteria. The most likely candidate “known unknowns,” which are those chemicals unknown to an investigator but contained within a reference database or literature source, rise to the top of a chemical list when rank-ordered by the number of associated data sources. The U.S. EPA’s CompTox Chemistry Dashboard is a curated and freely available resource for chemistry and computational toxicology research, containing more than 720,000 chemicals of relevance to environmental health science. In this research, the performance of the Dashboard for identifying known unknowns was evaluated against that of the online ChemSpider database, one of the primary resources used by mass spectrometrists, using multiple previously studied datasets reported in the peer-reviewed literature totaling 162 chemicals. These chemicals were examined using both applications via molecular formula and monoisotopic mass searches followed by rank-ordering of candidate compounds by associated references or data sources. A greater percentage of chemicals ranked in the top position when using the Dashboard, indicating an advantage of this application over ChemSpider for identifying known unknowns using data source ranking. Additional approaches are being developed for inclusion into a non-targeted analysis workflow as part of the CompTox Chemistry Dashboard. This work shows the potential for use of the Dashboard in exposure assessment and risk decision-making through significant improvements in non-targeted chemical identification.

Keywords: Non-targeted Analysis, Suspect Screening, DSSTox, High Resolution Mass Spectrometry

Introduction

Data processing workflows in non-targeted analysis (NTA) and suspect screening analysis (SSA) routinely identify a small percentage (often <5%) of likely chemical compounds in environmental samples [1, 2]. Improvements in compound identification can enhance exposure assessment, especially when the use of confirmation standards is not practical or possible (at the 'tentative' or 'probable' degrees of certainty [3-5]). Online reference databases can be useful for identifying "known unknowns" by searching intrinsic properties, specifically molecular formula and monoisotopic mass, and rank-ordering by the number of associated references or data sources [6, 7]. In this process, the most likely candidate "known unknowns," which are those compounds unknown to a researcher but known in a reference dataset or resource, are elevated to the top of a search results list. Researchers have previously reported that the freely available chemical database ChemSpider (http://www.chemspider.com/) [8, 9] proved more useful than the Chemical Abstract Service (CAS) Registry^{SM} when identifying known unknowns, with a key distinction of ChemSpider being the ability to search by monoisotopic mass [7]. Since this initial work, additional studies have reported using ChemSpider (amongst other databases) to support structure identification [10-13]. However, to enhance compound identification strategies, calls have also increased for improvements to open reference databases and analysis workflows (including "one-pass analysis"), and for public sharing of mass spectral data [2, 10, 11, 14].

The United States Environmental Protection Agency (US EPA) is developing a public resource for computational chemistry, toxicology, and exposure research efforts. This freely-available resource, known as the CompTox Chemistry Dashboard (https://comptox.epa.gov/dashboard; hereafter referred to as the Dashboard), is part of a suite of databases and applications developed by the National Center for Computational Toxicology (https://www.epa.gov/aboutepa/about-national-center-computational-toxicology-ncct), and integrates data from the Distributed Structure-Searchable Toxicity (DSSTox) database (DSSTox_v2) [15]. The underlying database has been expanded, with an emphasis on curation and characterizing data quality, to include hundreds of thousands of chemicals. Recent efforts have involved incorporating specific search tools into the Dashboard to benefit NTA. The Dashboard's current utilities include the ability to search a reference database of ~720,000 chemicals by monoisotopic mass and molecular formula. In this research, we evaluated the effectiveness of the Dashboard in the identification of known unknowns, comparing results against those from the *de facto* freely available online database for mass spectrometry based structure identification, ChemSpider, using the same method of rank-ordering of associated references or data sources reported by Little *et al* [7]. Determining the utility of the Dashboard relative to the current standard of freely available chemistry databases will benefit future research applications both within the US EPA and the scientific community as a whole by highlighting the effectiveness of tools designed for NTA users with a new, highly curated chemical reference database.

Methods

A total of 162 chemicals were selected for the assessment of the Dashboard using search and data source rank-ordering techniques (see Electronic Supplementary Material Table 1). The selected chemicals (n=162) were compiled from the Little *et al* [7] article that initiated this approach for NTA and from recent environmental and NTA literature. Selected chemicals include pharmaceuticals, dyes, surfactants, chemicals used in manufacturing, and personal care products that have been previously reported in environmental media (water [2, 13, 16], wastewater [16], dust [1], etc.). Monoisotopic masses, formulae, and structural identifiers for all chemicals are reported in the Electronic Supplementary Material (see Electronic Supplementary Material Table 1).

The workflow of known unknown identification by data source ranking has been previously described [6, 7]. The same workflow was followed here with minor amendments. Using the Advanced Search option in the Dashboard, a user can enter either a defined mass range (i.e. 263.87 to 263.89 amu) or a single mass with an associated error range (i.e. 263.881 amu ± 0.005 amu), see Electronic Supplementary Material Figures 1-6 for more details. Currently the Dashboard allows for mass search ranges and error to be entered in atomic mass units (amu) only. Therefore, monoisotopic masses of selected chemicals were searched using the Advanced Search tools in both ChemSpider and the Dashboard with an error of 0.005 amu. Most accurate mass measurement instruments can achieve a standard deviation of 5 ppm or better mass error; in order to be applicable for users with a range of accurate mass capabilities, the error window used in this work (0.005 amu) encompasses at least two standard deviations for all but the highest molecular weight chemicals. Advanced Search results were sorted in descending order by the number of associated references (in ChemSpider per Little *et al* [7]) or data sources (in the Dashboard). References in ChemSpider are the number of external IDs for a given chemical and data sources in the Dashboard represent the number of times that a dataset in the DSSTox database contains a particular chemical. Prevalence across many data sources and/or references is indicative, in this context, of a chemical's relative likelihood of occurrence [7]. The rank of each chemical of interest within the search results after sorting was recorded (Figure 1). The method was repeated in each application using molecular formulae for every chemical of interest to compare results of formula-based searching to those of mass-based searching.

For a complete comparison, ranking results in both applications of the 89 chemicals from Little *et al* [7] were also evaluated independently to explicitly assess the Dashboard relative to the dataset that initiated this approach. Little *et al* [7] also evaluated their workflow on a set of large molecular weight unique commercial polymers not included in the set of 89. For continuity of comparison, these 12 compounds were searched and rank-ordered following the above methods separately from the 162 chemicals.

No modifications to the search parameters or software were made during this study. All methods are demonstrated in the Electronic Supplemental Material (see Electronic Supplementary Material Figures 1-6) and can be repeated in the publicly available Dashboard. Searches were executed in both applications in July 2016. Statistical analyses were conducted in the R Statistical Computing Environment [17].

Results and Discussion

Overall Rank-Ordering

The goal of rank-ordering unidentified chemicals using their monoisotopic mass or molecular formula is to bring the most likely candidate chemicals to the top of the list for either tentative identification or further investigation. Entering monoisotopic masses with an error range of 0.005 amu and ranking by data sources, the average position rank of all 162 chemicals in the Dashboard was 1.31 with the number 1 rank occurring 85% of the time (Table 1). Using ChemSpider the average position rank across all chemicals was 2.20 with the number 1 rank occurring in 70% of the 162 searches (this average includes the removal of an outlier where the rank of one particular chemical was 201); average position rank in the Dashboard was significantly lower than in ChemSpider (Mann-Whitney U test, p= 0.0005). Formula-based searching yielded improved ranking statistics, consistent with what has been previously reported in the literature [7]. Mean rank position and percentage of chemicals occurring in the number one position improved when searching molecular formulae in both applications and independently, the Dashboard significantly outperformed ChemSpider (p=0.0083, Table 1, Electronic Supplementary Material Tables 2-3). Interestingly, mass-based searching in the Dashboard resulted in similar mean rank position and a higher percentage of chemicals in the number one rank position than formula-based searching using ChemSpider. Chemical formula assignment can vary in certainty with varying mass accuracy. As mass accuracy declines, more potential formulae can be generated from the same monoisotopic mass, introducing more error to formula assignment. Therefore, skipping the step of formula generation and assignment before chemical identification would represent an ideal situation leading towards a one-pass analysis [11]. These data indicate that for the chemicals included in this study, it is just as reliable to directly search the Dashboard using a monoisotopic mass than it would be to attempt to first generate a formula and search ChemSpider using the formula.

Table 1. Summary statistics of rank-ordering all 162 chemicals using data sources or associated references in both the CompTox Chemistry Dashboard and in ChemSpider.

	Mass-based Searching		Formula-based Searching	
	Dashboard	ChemSpider	Dashboard	ChemSpider
Average Rank Position	1.3	2.2[a]	1.2	1.4
Percent in #1 Position	85%	70%	88%	80%

[a] Average rank in ChemSpider shown here does not include an outlier where the rank was 201, when added the average rank position is 3.5.

Rank-Ordering of Chemical Class

The two largest classes of compounds compiled for this study were pharmaceutical drugs and industrial chemicals. When searching monoisotopic masses, 82% and 76% of pharmaceutical drugs ranked number one using the Dashboard and ChemSpider, respectively (Tables 2, 3). Pharmaceutical drugs are increasingly important in environmental NTA and risk assessment due to their ubiquitous presence in water and other environmental media [18, 19], and correctly identifying these compounds is important to document for researchers in environmental and human health risk assessment. Greater than 80% of the chemicals in several other compound classes ranked number one using mass-based searches in the Dashboard, including industrial chemicals, steroid hormones, pesticides, and veterinary drugs (Table 2). For those classes containing more than five chemicals, personal care products resulted in the worst average rank position of searched masses in both ChemSpider and the Dashboard. Two chemicals in particular, paraxanthine, a caffeine metabolite, and hexyl dodecanoate, a skin conditioning emollient, fell outside of the top five rank-ordered results when searched by both mass and formula. In the case of paraxanthine, two other more prevalent metabolites of caffeine precede it in the data source ranking. Hexyl dodecanoate has several constitutional isomers, many of which are also emollients, which rank ahead of it in terms of number of sources. This identifies a potential drawback of this rank-ordering workflow in that metabolites and isomers may not be distinguishable by data source ranking alone.

Table 2. Results of searching by monoisotopic mass and rank-ordering by number of data sources in the CompTox Chemistry Dashboard, listed by compound class

Compound class	Number in class	Average Rank	Number of compounds in each position rank-ordered				
			#1	#2	#3	#4	#5+
Pharmaceutical Drug	72	1.3	59	8	3	2	
Industrial Chemicals	42	1.2	38	1	1	2	
Personal Care Products	8	2.6	6				2
Steroid Hormones	7	1.0	7				
Perfluorochemicals	6	1.3	5	1			
Pesticides	12	1.3	10	1	1		
Veterinary Drugs	3	1.0	3				
Dyes	2	1.0	2				
Food product/natural compounds	4	1.5	3		1		
Illicit Drugs	2	1.5	1	1			
Misc. Molecules	3[a]	1.0	3				

[a]One organic molecule (tephrosin) not present in the Dashboard

Table 3. Results of searching by monoisotopic mass and rank-ordering by number of associated references in ChemSpider, listed by compound class

Compound class	Number in class	Average Rank	Number of compounds in each position rank-ordered				
			#1	#2	#3	#4	#5+
Pharmaceutical Drug	72	1.4	55	9	6	2	
Industrial Chemicals	42	5.5	28	6	3		5
Personal Care Products	8	6.1	3	1			4
Steroid Hormones	7	1.0	7				
Perfluorochemicals	6	1.2	5	1			
Pesticides	12	2.3	6	2	3		1
Veterinary Drugs	3	1.3	2	1			
Dyes	2	1.0	2				
Food product/natural compounds	4	3.8	2			1	1
Illicit Drugs	2	2.0	1		1		
Misc. Molecules	3 [a]	1.3	2	1			

[a]Tephrosin was removed from average rank calculations as it was not present in a Dashboard search

Comparison to Little et al Datasets

For continuity and comparison, the 89 chemicals used to document ChemSpider's utility in known unknown identification were analyzed further (Table 4). On this smaller subset, the Dashboard again significantly outperformed ChemSpider (p=0.009) when searching monoisotopic mass and the average rank of molecular formula searches were similar (Table 4). A greater number of chemicals ranked number one when rank ordering after a mass search in the Dashboard than after a formula search in ChemSpider, mirroring what was observed on the entire set of 162 chemicals. However, one chemical within the Little *et al* [7] list was present in ChemSpider but not in the Dashboard. Tephrosin, a natural toxin, is not contained within the DSSTox database, and therefore not searchable in the Dashboard. Additionally, ChemSpider's performance based on this analysis did not match that which was previously reported [7]. Specifically, the number of times each chemical ranked number one when searched by molecular formula declined.

A set of 12 large molecular weight chemical compounds (all MW>600 Da) were evaluated separately from the list of 89 in the initial research by Little *et al* [7] to determine identification efficacy of unique commercial polymer additives. For a complete assessment, these 12 compounds were separately evaluated following the same methods. Two of the 12 compounds were absent from the Dashboard while all 12 were contained within ChemSpider (see Electronic Supplementary Material Table 4). By rank-ordering, all of the compounds in this list that were contained in the Dashboard ranked number 1 by both mass and formula searching. However, this does highlight that chemicals outside the domain of the database are not captured in this method, indicating that for true unknowns other identification processes need to be incorporated.

The number of entries in ChemSpider has doubled since 2012, from 26 million to 57 million today. More entries can be beneficial (as reflected in the omissions in the Dashboard), but it can also interfere with the identification of likely candidate chemicals as reported in this research (Table 4). This is also true for other resources such as PubChem (presently containing more than 90 million chemicals [20]) as well as the Chemical Abstracts Service (CAS) Registry^{SM} (containing more than 100 million chemicals). A comparison of the number of possible results returned from formula searches in each platform illustrates this complication (see Electronic Supplementary Material Table 5). For the formula of piperine (C17H19NO3), PubChem returns 20,000 possible results, ChemSpider returns 9000, and the Dashboard returns 100. Based on data source ranking piperine was the top result in the Dashboard and the 4th highest in ChemSpider. The Dashboard is being developed with a focus on high-quality data of particular value to the environmental sciences and toxicology communities. Large scale collections of chemicals extracted from patents and chemical vendor collections are not included in the database as support for these efforts is already provided by PubChem and ChemSpider. This approach leads to a cleaner database allowing for more precise known unknown identification.

Table 4. Summary statistics and rank-ordered position in the CompTox Chemistry Dashboard and ChemSpider of the 89 compound subset from the Little *et al* [7] study

		Average Rank	Number in each position rank-ordered				
		(± SD)	#1	#2	#3	#4	#5+
Mass-based	Dashboard	1.2 ± 0.7	77[a]	5	3	3	
	ChemSpider	2.2 ± 6.1[b]	68	8	7	1	5
Formula-based	Dashboard	1.1 ± 0.4	78[a]	8	2		
	ChemSpider	1.3 ± 1.0	77	8	2	1	2

[a] One chemical (tephrosin) not present in the Dashboard

[b] Average rank in ChemSpider shown here does not include an outlier where the rank was 201, when added the average rank position is 4.4.

Ongoing Work

Rank-Ordering Methods

Additional search and rank-order criteria are presently undergoing testing within the CompTox Chemistry Dashboard for further improvements in known unknown chemical identification. Under the premise of this work and the work of others (e.g. [7, 6]), chemicals of interest in environmental media are likely those with the most sources, or are the most 'popular' chemicals. Preliminary results indicate that searching the unique InChIKey identifier of chemicals of interest in Google, and rank-ordering the results by the number of result hits, provides an even more accurate identification than using the Dashboard and data sources. These data could be used to enhance or replace data sources within the Dashboard for known unknown investigations. Additionally, rank-order statistics improve when tightening the search window around a monoisotopic mass. Further research developing a sliding mass search scale based on relative monoisotopic mass (i.e. a smaller search window around a smaller mass) could result in more accurate identification of known unknowns.

To further identify chemicals in environmental media, functional use and product occurrence data, as contained in the US EPA's CPCat database [21], can be incorporated into searching and rank-ordering. Chemical use and function category data, organized with descriptors such as detergent, food_additive, etc., are currently available in the Dashboard. These data may further inform tentative chemical identification through filtering by use category relative to sample medium or through compiled use ranking metrics; testing in the Dashboard is ongoing. Further research to create a weighting-based or tiered ranking approach for identification using all aforementioned criteria as inputs is underway.

MS-Ready Structures

Charged and salted forms of chemicals contained within chemical reference databases complicate the search and identification process as these forms are not consistent with the form an analyst would detect via high resolution mass spectrometry in NTA. As an example, the colorant FD&C Blue No. 1 (or Brilliant Blue FCF) is present in both ChemSpider and the Dashboard as a charged molecule with two sodium ions. Therefore, when searching a neutral unidentified monoisotopic mass on both applications, neither resource would return the chemical identified via NTA. Chemical structure curation and standardization can remove duplicates and inconsistencies in structures to allow for cleaner tentative identification. Mansouri *et al* (2016) developed chemical structure standardization approaches to create QSAR (Quantitative Structure Activity Relationship)-ready structures for use in estrogenic receptor activity screening [22]. This workflow has since been applied to all chemical structures contained in the DSSTox databased and exposed in the Dashboard. QSAR-ready structures are neutral, de-salted, and contain no stereochemistry information, and are consistent with the chemical forms detected in mass spectrometry (when corrected for charge-state). In other words, structures standardized into QSAR-ready form happen to offer us MS-ready structures as a benefit. These will be incorporated into the Dashboard, allowing users to be able to easily identify the associated substances, whether they be salts, associated with solvents of hydration, etc. The ability to search MS-ready structures has already been delivered via an iOS mobile app by making our data freely available from the NCCT website (ftp://newftp.epa.gov/COMPTOX/Sustainable_Chemistry_Data/Chemistry_Dashboard). The m/z EPA CompTox app (https://itunes.apple.com/app/m-z-comptox/id1148436331) is already freely available, thereby providing accessibility for NTA users.

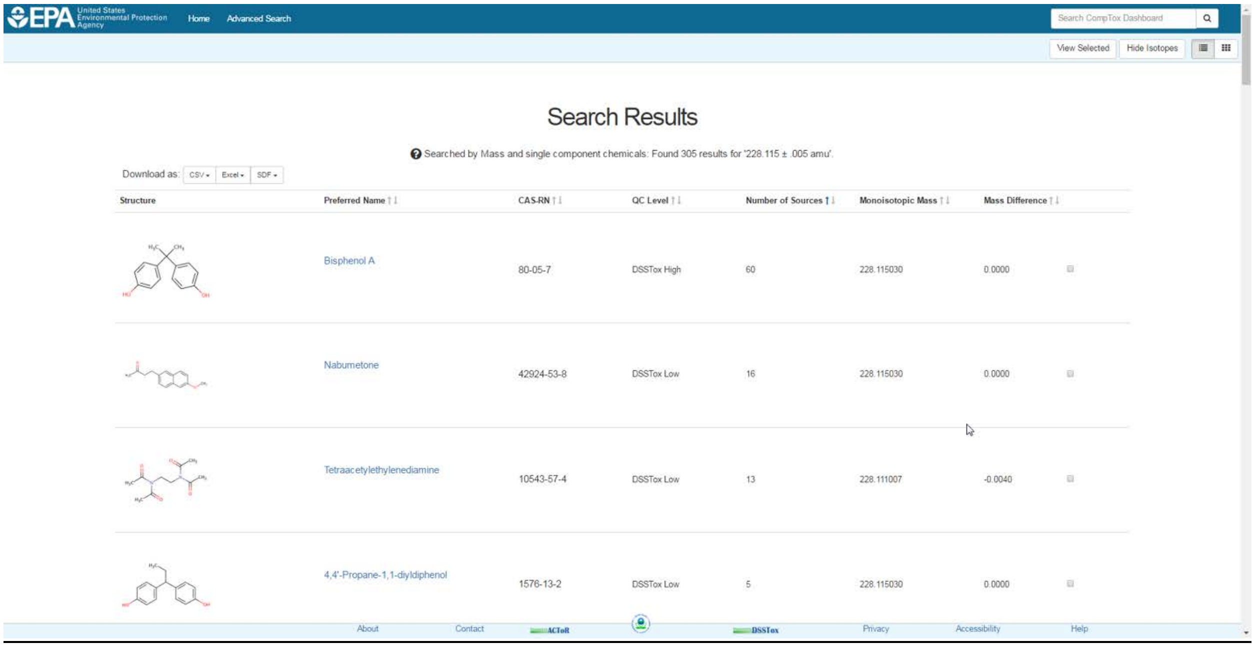
API Development

Planned developments for the Dashboard include an application programming interface (API) and access to a suite of web services. Programmatic access will allow third parties to investigate and interrogate the data within the database for their own known unknowns analyses. Within an investigation of observed chemical features, a user could include ChemSpider for expansive coverage, the Dashboard for focused high-quality data, and even more focused resources like FOR-IDENT (http://for-ident.hswt.de/) [23] for water-specific analyses, among others. Additional capabilities within the API will enable the user to access and incorporate algorithmically generated mass spectral fragmentation resources and metabolite databases for known unknown chemical identification (including spectral library resources like MassBank [24] and mzcloud [25], *in silico* fragmentation resources like MetFrag [26, 12], and metabolite databases such as Metlin [27]). Chemical metabolites and degradants in environmental media present a difficult problem from an identification perspective. Using the Dashboard to identify known unknowns in the workflow presented here does not include an avenue for metabolites or fragments. However, linking the Dashboard via web services to the open resources available for algorithmically generated metabolites and mass spectra can advance chemical identification in NTA through structure elucidation and metabolite identification.

Conclusions

The Dashboard is a highly curated freely available online reference database that is an effective investigative tool for the identification of known unknowns. Comparisons with the ChemSpider database, a primary database for mass spectrometrists to utilize for structure identification purposes, show better performance overall for the test sets reported here. Expanding the data, functionality and access to support projects within the EPA, and in the scientific community as a whole, will further demonstrate its utility for risk analysis and general chemical identification both as part of larger, more developed workflows and as a stand-alone investigative tool. Future research on expanded utility employing further chemical identification mechanisms will advance the field of NTA and chemical identification in a public arena for widespread use.

Figures



Search Results

Searched by Mass and single component chemicals: Found 305 results for '228.115 ± .005 amu'.

Structure	Preferred Name	CAS-RN	QC Level	Number of Sources	Monoisotopic Mass	Mass Difference
	Bisphenol A	80-05-7	DSSTox High	60	228.115030	0.0000
	Nabumetone	42924-53-8	DSSTox Low	16	228.115030	0.0000
	Tetraacetylethylenediamine	10543-57-4	DSSTox Low	13	228.111007	-0.0040
	4,4'-Propane-1,1-diyldiphenol	1576-13-2	DSSTox Low	5	228.115030	0.0000

Figure 1. Advanced search results table in the CompTox Chemistry Dashboard (https://comptox.epa.gov/dashboard) after an advanced search of monoisotopic mass 228.115 ± 0.005 amu. Results are ranked in descending order by the number of data sources.

References

1. Rager JE, Strynar MJ, Liang S, McMahen RL, Richard AM, Grulke CM et al. Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring. Environ Int. 2016;88:269-80.

2. Schymanski EL, Singer HP, Slobodnik J, Ipolyi IM, Oswald P, Krauss M et al. Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. Anal Bioanal Chem. 2015;407(21):6237-55.

3. Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP et al. Identifying small molecules via high resolution mass spectrometry: communicating confidence. Environ Sci Technol. 2014;48(4):2097-8.

4. Letzel T, Bayer A, Schulz W, Heermann A, Lucke T, Greco G et al. LC–MS screening techniques for wastewater analysis and analytical data handling strategies: Sartans and their transformation products as an example. Chemosphere. 2015;137:198-206.

5. Letzel T, Lucke T, Schulz W, Sengl M, Letzel M. OMI (Organic Molecule Identification) in water using LC-MS (/MS): Steps from “unknown” to “identified”: a contribution to the discussion In a class of its own. lab&more. 2014; 4: 24-28. http://www.int.laborundmore.com/archive/921107/OMI-(Organic-Molecule-Identification)-in-water-using-LC-MS(-MS)%3A-Steps-from-%E2%80%9Cunknown%E2%80%9D-to-%E2%80%9Cidentified%E2%80%9D%3A-a-contribution-to-the-discussion.html

6. Little JL, Cleven CD, Brown SD. Identification of “known unknowns” utilizing accurate mass data and chemical abstracts service databases. J Am Soc Mass Spectr. 2011;22(2):348-59.

7. Little JL, Williams AJ, Pshenichnov A, Tkachenko V. Identification of “known unknowns” utilizing accurate mass data and ChemSpider. J Am Soc Mass Spectr. 2012;23(1):179-85.

8. Pence HE, Williams A. ChemSpider: an online chemical information resource. J Chem Educ. 2010;87(11):1123-4.

9. Royal Society of Chemistry. ChemSpider. 2016. http://www.chemspider.com/.

10. Schymanski EL, Singer HP, Longrée P, Loos M, Ruff M, Stravs MA et al. Strategies to Characterize Polar Organic Contamination in Wastewater: Exploring the Capability of High Resolution Mass Spectrometry. Environ Sci Technol. 2014;48(3):1811-8. doi:10.1021/es4044374.

11. Godfrey AR, Brenton AG. Accurate mass measurements and their appropriate use for reliable analyte identification. Anal Bioanal Chem. 2012;404(4):1159-64. doi:10.1007/s00216-012-6136-y.

12. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminform. 2016;8(1):1-16. doi:10.1186/s13321-016-0115-9.

13. Bade R, Causanilles A, Emke E, Bijlsma L, Sancho JV, Hernandez F et al. Facilitating high resolution mass spectrometry data processing for screening of environmental water samples: An evaluation of two deconvolution tools. Sci Total Environ. 2016;569:434-41.

14. Zedda M, Zwiener C. Is nontarget screening of emerging contaminants by LC-HRMS successful? A plea for compound libraries and computer tools. Anal Bioanal Chem. 2012;403(9):2493-502. doi:10.1007/s00216-012-5893-y.

15. Richard AM, Williams CR. Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. Mutat Res-Fund Mol M. 2002;499(1):27-52. doi:http://dx.doi.org/10.1016/S0027-5107(01)00289-5.

16. McEachran AD, Shea D, Bodnar W, Nichols EG. Pharmaceutical occurrence in groundwater and surface waters in forests land-applied with municipal wastewater. Environ Toxicol Chem. 2016;35(4):898-905. doi:10.1002/etc.3216.

17. R Team Core. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.

18. Kolpin DW, Furlong ET, Meyer MT, Thurman EM, Zaugg SD, Barber LB et al. Pharmaceuticals, hormones, and other organic wastewater contaminants in U.S. streams, 1999-2000: A national reconnaissance. Environ Sci Technol. 2002;36(6):1202-11.

19. Klosterhaus SL, Grace R, Hamilton MC, Yee D. Method validation and reconnaissance of pharmaceuticals, personal care products, and alkylphenols in surface waters, sediments, and mussels in an urban estuary. Environ Int. 2013;54:92-9.

20. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A et al. PubChem Substance and Compound databases. Nucleic Acids Res. 2016;44(D1):D1202-13. doi:10.1093/nar/gkv951.

21. Dionisio KL, Frame AM, Goldsmith M-R, Wambaugh JF, Liddell A, Cathey T et al. Exploring consumer exposure pathways and patterns of use for chemicals in the environment. Toxicol Rep. 2015;2:228-37.

22. Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A et al. CERAPP: Collaborative estrogen receptor activity prediction project. Environ Health Persp. 2016. doi:10.1289/ehp.1510267.

23. RISK-IDENT. STOFF-IDENT. 2013. http://risk-ident.hswt.de/pages/de/links.php.

24. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010;45(7):703-14. doi:10.1002/jms.1777.

25. HighChem. mzCloud. 2016. https://www.mzcloud.org/. 16 August 2016.

26. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics. 2010;11(1):1.

27. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR et al. METLIN: a metabolite mass spectral database. Ther Drug Monit. 2005;27(6):747-51.