McEachran *et al. J Cheminform (2018) 10:45*
https://doi.org/10.1186/s13321-018-0299-2

Journal of Cheminformatics

METHODOLOGY

Open Access

"MS-Ready" structures for non-targeted high-resolution mass spectrometry screening studies

Andrew D. McEachran[1,2*], Kamel Mansouri[1,2,3], Chris Grulke[2], Emma L. Schymanski[4], Christoph Ruttkies[5] and Antony J. Williams[2*]

Abstract

Chemical database searching has become a fixture in many non-targeted identification workflows based on high-resolution mass spectrometry (HRMS). However, the form of a chemical structure observed in HRMS does not always match the form stored in a database (e.g., the neutral form versus a salt; one component of a mixture rather than the mixture form used in a consumer product). Linking the form of a structure observed via HRMS to its related form(s) within a database will enable the return of all relevant variants of a structure, as well as the related metadata, in a single query. A Konstanz Information Miner (KNIME) workflow has been developed to produce structural representations observed using HRMS ("MS-Ready structures") and links them to those stored in a database. These MS-Ready structures, and associated mappings to the full chemical representations, are surfaced via the US EPA's Chemistry Dashboard (https://comptox.epa.gov/dashboard/). This article describes the workflow for the generation and linking of ~ 700,000 MS-Ready structures (derived from ~ 760,000 original structures) as well as download, search and export capabilities to serve structure identification using HRMS. The importance of this form of structural representation for HRMS is demonstrated with several examples, including integration with the in silico fragmentation software application MetFrag. The structures, search, download and export functionality are all available through the CompTox Chemistry Dashboard, while the MetFrag implementation can be viewed at https://msbi.ipb-halle.de/MetFragBeta/.

Keywords: High-resolution mass spectrometry (HRMS), Structure identification, Structure curation, Database searching

Background

In recent years the use of high-resolution mass spectrometry (HRMS) instrumentation coupled to gas and liquid chromatography has become increasingly common in environmental, exposure and health sciences for the detection of small molecules such as metabolites, natural products and chemicals of concern [1–5]. Advances in instrumentation have led to faster acquisition times, lower limits of detection, and higher resolution, improving the rapid identification of chemicals of interest. However, the bottleneck of data processing has evolved to become the foremost challenge for non-targeted and suspect screening analyses (NTA and SSA, respectively) [1, 2, 6]. Workflows to address data processing can vary substantially between laboratories and depend on access to various software and programming capabilities. Common data processing workflows in NTA and SSA often utilize a combination of vendor-specific software, open source platforms, and in-house resources [1, 3, 7].

In NTA the analyst generally uses peak-picking software to identify molecular features to find the (pseudo) molecular ion (*m/z*) along with associated isotopic peaks and calculate the neutral monoisotopic mass (Fig. 1a, b). Monoisotopic masses can be searched in structure databases to retrieve tentative candidates or can be used in combination with isotopic distributions and/or

*Correspondence: mceachran.andrew@epa.gov; williams.antony@epa.gov
[2] National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Mail Drop D143-02, 109 T.W. Alexander Dr., Research Triangle Park, NC 27711, USA
Full list of author information is available at the end of the article

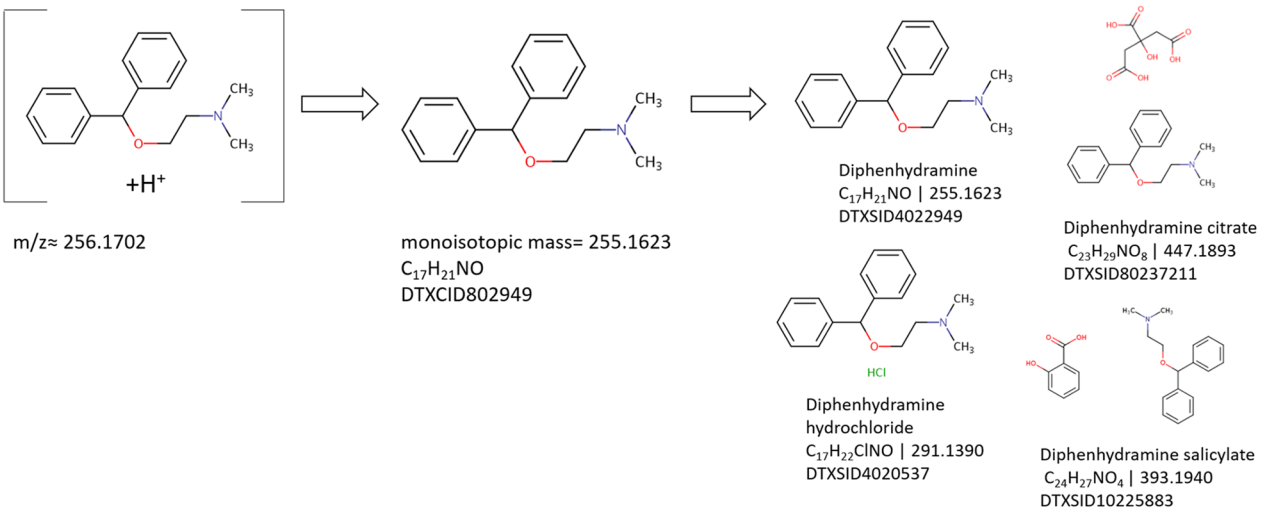

Fig. 1 Using the example of the structure of diphenhydramine (DTXSID4022949 [17]): in HRMS, molecular features and associated ions are used to identify the pseudomolecular ion at a specific *m/z* (**a**). This information is then used to calculate the neutral monoisotopic mass and/or molecular formula (**b**). Both a neutral mass and formula can be searched in structure databases to retrieve matching candidate results (**c**). The MS-Ready form of a structure (**b** DTXCID802949 [18]) and the substance form(s) of a chemical (**c** DTXSID4022949 [17]; DTXSID80237211 [19]; DTXSID4020537 [20]; DTXSID10225883 [21]) are linked such that all can be retrieved in a single query with the EPA's DSSTox database. DTXCID indicates the unique *chemical* identifier and DTXSID indicates the unique *substance* identifier, linked to metadata

fragmentation data to arrive at a molecular formula(e) before candidate searching (Fig. 1c). Candidate selection often combines concepts such as database searching and data source ranking [7–9], spectral matching [10, 11] and retention time feasibility [7, 12–14] to identify the most probable structures, with database presence and metadata proving critical to success [7, 15]. When fragmentation information was combined with metadata and retention time information in MetFrag2.2, the number of correct identifications improved from 22% (105 of 473 correct) to 89% (420 of 473) on candidates retrieved from ChemSpider [16] using molecular formulae [7]. However, mixtures and salts (and thus their associated metadata) were excluded from candidate lists as these would not be observed at the calculated exact mass or formula used for searching. Yet, multi-component forms of a chemical (e.g., mixtures and salts, Fig. 1c) may contain the component observed via HRMS. Excluding these from database searches limits which substances can be identified by excluding variants of a structure and associated metadata.

Despite the prevalence of structure databases and online chemistry resources in NTA workflows, relatively little work has been done within the community to curate and standardize chemical structures in databases to optimize searching and identification with HRMS data [22, 23]. To maximize the search capabilities of structure databases, both the substance form, commonly represented by a structure (Fig. 1c), and the "MS-Ready" form (Fig. 1b) of the structure should be contained within databases and linked. When properly linked, both the observed form and variants of the structure observed via HRMS can be presented, thereby allowing the analyst to subsequently access metadata that may provide increased evidence in structure identification [5, 9, 15, 22, 24].

To link particular forms of a substance to their structure components (i.e., salts and mixtures) and their related MS-Ready forms, structure standardization is required. Various curation and standardization approaches are already defined in cheminformatics [25–28] and in use within the quantitative structure–activity relationship (QSAR) modeling community [27, 29]. QSAR modelers generally need desalted, neutralized, non-stereospecific structures, typically excluding inorganics and mixtures, to facilitate calculating molecular descriptors used in subsequent modeling approaches. Workflows describing the generation of QSAR-Ready structures have previously been published [27, 28, 30]. The requirements to produce MS-Ready structures are similar (vide infra), thus the processing rule set to produce QSAR-Ready files could be altered to provide an MS-Ready form of the data with a number of appropriate extensions. Hence, a previous QSAR-Ready structure preparation workflow [28, 30] was adapted to produce MS-Ready chemical structure forms that are amenable to structure identification

using database searching. The resulting Konstanz Information Miner (KNIME) workflow, associated rule set and software processing module for the generation of MS-Ready structures are provided as an outcome of this work and available for download from a Github repository [31]. In addition, this workflow was used to generate MS-Ready forms (~700,000) for the ~760,000 chemicals substances in DSSTox [32] for access via the US EPA's CompTox Chemistry Dashboard (hereafter "Dashboard") [33]. The functionality in the Dashboard includes the ability to search, export and download MS-Ready structures. Several examples are provided to demonstrate the value of MS-Ready structures, including integration and demonstration of identification in NTA through the in silico fragmenter MetFrag [7]. Through accessibility to MS-Ready structures and the integration between the Dashboard and MetFrag, valuable resources to support structural identification of chemicals, now including mixtures and salts, are available to the community.

Methods

MS-Ready processing workflow

The MS-Ready processing workflow is an extension of the workflows described in detail by Mansouri et al. to curate and prepare QSAR-Ready structures for use in the development of prediction models [28, 30]. The related QSAR-Ready workflow is openly available on GitHub [34]. The free and open-source environment KNIME (Konstanz Information Miner) was used to design and implement the workflow [35]. Only free and open source KNIME nodes were used in the workflow. Cheminformatic steps were mainly performed using INDIGO nodes [36]. The nodes for each step were grouped into metanodes to ease readability and increase flexibility and future updates.

The MS-Ready workflow and transformation files are available on GitHub [31] and consisted of the following steps:

1. Consistency checking: file format, valence, and structural integrity.
2. Removal of inorganics and separation of mixtures into individual components.
3. Removal of salts and counterions (the salts list is available in Additional file 1).
4. Conversion of tautomers and mesomers to consistent representations. Examples include: nitro and azide mesomers, keto–enol tautomers, enamine–imine tautomers, enol-ketenes, etc. [37–39].
5. Neutralization of charged structures and removal of stereochemistry information.
6. Addition of explicit hydrogen atoms and aromatization of structures.
7. Removal of duplicates using InChIKey [40].

Differences between the QSAR-Ready and MS-Ready workflows exist primarily in the handling of salts and counterions, chemical mixtures, metals, and organometallics (Fig. 2). For the generation of both QSAR and MS-Ready structures, salts and solvents are separated and removed from mixtures via an exclusion list (Fig. 2a). The exclusion list used during QSAR-Ready structure preparation (189 structures, SDF file provided as Additional file 2) was substantially reduced for MS-Ready structures (32 structures, SDF file provided as Additional file 1), allowing a greater number of secondary components that are observable in MS to be retained and linked to the original substances via MS-Ready forms (e.g., benzoate, fumarate, citrate). For MS-Ready structures, all records still containing multiple components were separated out, deduplicated if necessary, and retained, with all components linked to the original substance (Fig. 2b, c). For the QSAR-Ready workflow, in contrast, chemical mixtures are excluded due to the complexity merging activity estimates for components of the mixture (Fig. 2b, c). The MS-Ready workflow retains organometallics containing covalent metal–carbon bonds within the chemical structure while the QSAR-Ready workflow does not (Fig. 2d), primarily because most descriptor packages used for QSAR modeling cannot handle organometallic compounds. However, users of MS-Ready structures for environmental and exposure NTA applications need to include substances such as organomercury and organotin compounds, due to their toxicity and use as, for example, fungicides and antifouling agents.

Mapping MS-Ready structures to substances

For the purpose of structure identification using the Dashboard, MS-Ready structures must be mapped to the associated chemical substances in the underlying DSSTox database [32]. Chemical substances within DSSTox are identified by unique DTXSIDs (DSSTox Substance Identifiers) and can denote a mixture, polymer or single chemical while DTXCIDs (DSSTox Chemical Identifier) are unique chemical structure identifiers. A structure-data file (SDF) of all chemical structures (DTXCIDs) associated with substances (DTXSIDs) was exported and passed through the MS-Ready preparation workflow. The resulting MS-Ready structures were then loaded back into the DSSTox structure table, omitting duplicate structures as identified by standard InChIKey [40] generated using the JChem Java API [41]. Mappings between the original DSSTox structure and its MS-Ready form was stored in a structure relationship mapping table.

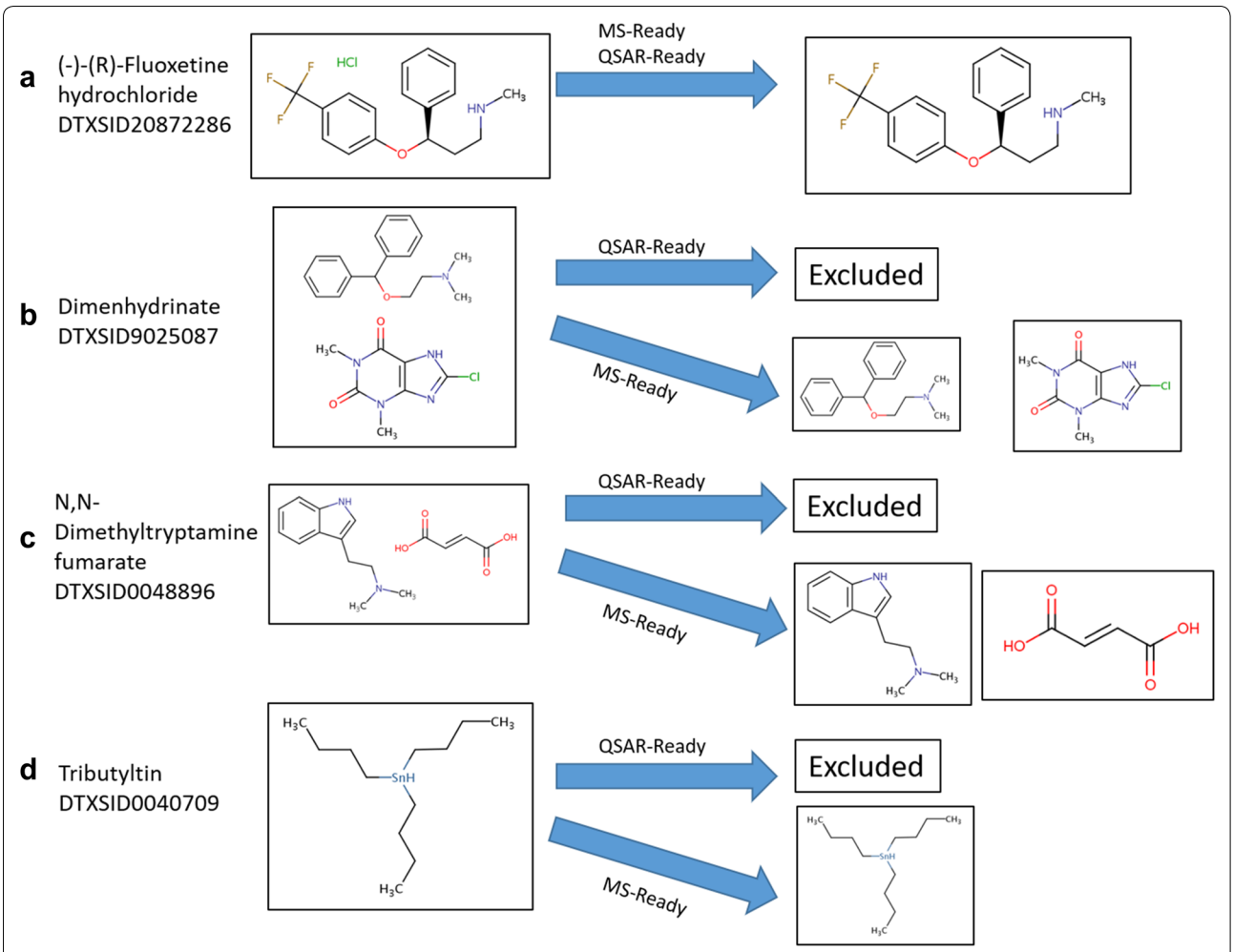

Fig. 2 Original substances (left) and processed, linked chemical structures (right) indicating similarities and differences between the QSAR-Ready and MS-Ready workflows. **a** Salt and stereochemistry removed for both QSAR- and MS-Ready purposes; **b**, **c** mixtures separated and linkages retained for MS-Ready, discarded for QSAR-Ready; **d** organometallics with metal–carbon bonds retained in MS-Ready, discarded in QSAR-Ready. The identities of the associated MS-Ready structures are visible in the "Linked Substances" tab of individual substance records in the Dashboard

Accessibility to MS-Ready results

Once mapped within the database, functionality to support searching based on MS-Ready structures was incorporated into the Dashboard [33] to support mass spectrometry-based NTA and SSA. MS-Ready structures can be searched using the Advanced Search page based on a single molecular formula [42] or can be searched in batch mode (i.e., 1–100 s of masses or formulae at a time) in the Batch Search interface [43]. The Batch Search interface allows for MS-Ready structure searching of both molecular formulae and monoisotopic masses. As the form of a chemical structure observed via HRMS is linked to all substances containing the structure (e.g., the neutral form, all salt forms, mixtures), when a molecular formula or monoisotopic mass is searched using MS-Ready structures, both single component and multi-component substances can be returned. This is distinct from an exact formula search whereby results returned match the input formula exactly (e.g., excluding mixtures where only a component matches that given formula). Figure 3 demonstrates the difference between an exact formula search (returning candidates to the left of the figure) and an MS-Ready search (which returns all candidates shown in the figure). Both exact formula and MS-Ready formula searches can be conducted within the Advanced Search and Batch Search pages of the Dashboard. Screenshots of the search interfaces and resulting file are provided in Additional file 3: Figs. S1–S4. Users can download the results with export options including SMILES and the identifiers that correspond with the substance (CASRN, preferred name, synonyms), chemical and MS-Ready forms. Column

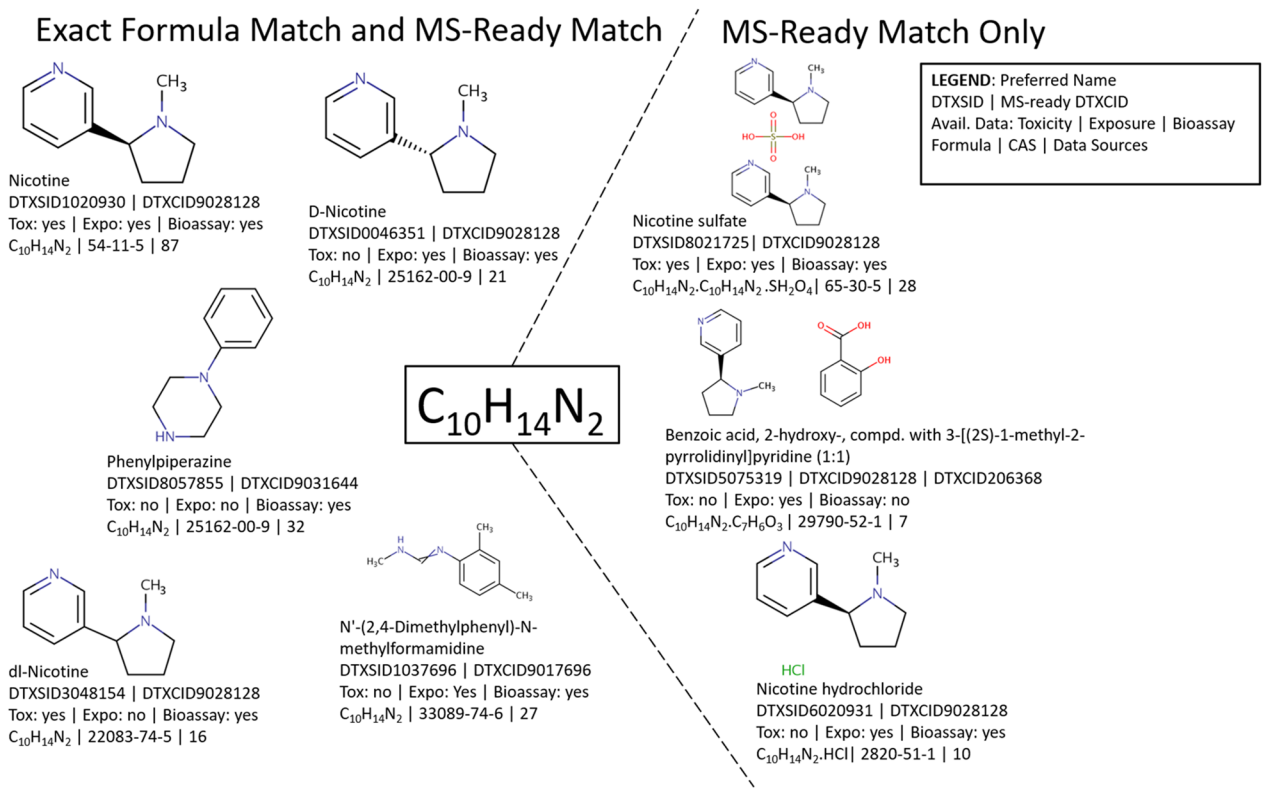

Fig. 3 Results of both an exact formula (left) and MS-Ready formula search (all) demonstrated using the molecular formula of nicotine ($C_{10}H_{14}N_2$), top left. A search of $C_{10}H_{14}N_2$ using the MS-Ready search functionality [45] retrieves all 8 substances while an exact formula search [46] retrieves only the 5 on the left. The MS-Ready DTXCID representing the chemical structure of nicotine is present in 6 of the 8 example substances (DTXCID9028128). Metadata such as toxicity, exposure, and bioactivity data vary for all results. Accessing the data for the mixtures, salts, etc. is unachievable in a single search without linking through the MS-Ready form. Figure based on the concept illustrated by Schymanski and Williams (2017), with permission [22]

headers specify the individual component structure (DTXCID) that was matched to the input as well as the mapped substance (DTXSID) and substance-associated data (Additional file 4: Tables S1 and S2). Additionally, users can include other data from the Dashboard export pane that is relevant to their needs (e.g., exposure data, bioactivity data, property predictions, presence in lists). This MS-Ready batch search option is designed to enable candidate retrieval through searching large numbers of suspect formulae and masses (Additional file 4: Table S2) [9]. By selecting the "MetFrag Input File" option in the Batch search, users can generate a file (including any selected metadata) containing all relevant structural information required for MetFrag to upload and process MS-Ready structures correctly (see below).

An MS-Ready file generated from all chemical structures contained within the DSSTox database is available for download [44]. With this file, users may create their own databases to incorporate into instrument software for screening.

Integration with MetFrag

The export option ("MetFrag Input File (Beta)" under *Metadata*) was added to the Batch Search page to create an MS-Ready export file suitable for direct import into the in silico fragmenter MetFrag [7, 47]. As outlined above, mixtures and salts are excluded in MetFrag by default. However, through the MS-Ready export file, MetFrag can now process the component of the mixture observed at the given input formula (i.e., the MS-Ready form) and retain the metadata and identifiers associated with the substance form (mixture, salt, original substance). Column headers in the Dashboard export were elaborated to distinguish the individual component structure (DTXCID) and associated data from data related to the substance (DTXSID). By default, the export file from the Dashboard contains the fields: INPUT; FOUND_BY; DTXCID_INDIVIDUAL_COMPONENT;

FORMULA_INDIVIDUAL_COMPONENT; SMILES_INDIVIDUAL_COMPONENT; MAPPED_DTXSID; PREFERRED_NAME_DTXSID; CASRN_DTXSID; FORMULA_MAPPED_DTXSID; SMILES_MAPPED_DTXSID; MS_READY_SMILES; INCHI_STRING_DTXCID; INCHIKEY_DTXCID; MONOISOTOPIC_MASS_DTXCID (Additional file 4: Table S3). Users can select any other additional data fields on the Batch Search page to include in the MetFrag scoring (details below). In this export file, MetFrag treats the "DTXSID" (substance identifier) field as the identifier, but takes the structural information (formula, mass, SMILES, InChI, InChIKey) from the fields denoted with DTXCID (which corresponds with the structure observed in MS). The other fields are included in the export file so that users can display the mixture or components. Any additional data fields that contain numeric data are automatically imported by MetFrag and included as an additional "Database scoring term" in the "Candidate filter & Score Settings" tab (Additional file 5: Figure S5).

By default, MetFrag groups all candidates with the same InChIKey first block, reporting only results from the highest scoring member of the group. However, the MS-Ready search involves components of mixtures, where individual components are often also in the Dashboard and contain different metadata. Merging these by the component InChIKey would result in a loss of the metadata obtained from the Dashboard search. To retain all candidates, the "Group candidates" option in the "Fragmentation Settings and Processing" tab should be deselected. Even if candidates are grouped, all substance identifiers within a group are still displayed and hyperlinked to the Dashboard (see Additional file 5: Fig. S6).

MetFrag example calculations

To demonstrate the workflow, the results of an MS-Ready formula search for $C_9H_{16}ClN_5$ (terbutylazine) and $C_7H_{12}ClN_5$ (desethylterbutylazine) were exported as.csv for import into MetFrag. The.csv file was imported into the MetFragBeta web interface [47] and the candidates were selected by molecular formula. Experimental fragmentation data were retrieved from the European MassBank [48] to conduct the queries in MetFrag. Spectral data for terbutylazine (DTXSID4027608 [49]) was collected from record EA028406 [50], recorded at collision energy HCD 75 (higher-energy collisional dissociation) and resolution 7500 (MS/MS) on an LTQ Orbitrap XL (at Eawag, Switzerland). Spectral data for desethylterbutylazine (DTXSID80184211) was also retrieved from MassBank, record EA067106 [51], likewise a MS/MS spectrum measured at HCD 75 and R = 7500 on the LTQ Orbitrap XL at Eawag. Metadata from the Dashboard that were included as scoring terms were: Data Sources, PubMed Reference Count, ToxCast % active and the presence in two lists: Norman Priority [52] and STOFF-IDENT [53]. The use of data sources in the Dashboard for identification of unknowns has been documented [9] and combined ranking schemes using multiple data streams and database presence are being optimized in current research. The metadata selected here should not be considered finalized scoring parameters but primarily to demonstrate functionality. The fragmentation settings were Mzppm = 5, Mzabs = 0.001, Mode = $[M+H]^+$, Tree depth = 2, Group candidates = deselected. In addition to the Dashboard scoring, the MetFrag Scoring Term "Exact Spectral Similarity (MoNA)" was activated [54]. On the MetFrag web interface, the combination of the regular MetFrag Fragmenter score (ranging from 0 to 1), the spectral similarity term (also ranging from 0 to 1) and each metadata field creates an additive score, with the maximum determined by the number of metadata fields selected. For example, the MetFrag Fragmenter score, spectral similarity score and 5 metadata categories mentioned here will result in a maximum score of 7, where the scores for each individual category are automatically scaled between 0 and 1 based on maximum values (no data gives score = 0). While it is possible to perform more sophisticated scoring via the command line version, this is beyond the scope of the current article—the work presented here is intended to demonstrate the potential for the MS-Ready approach to support identification efforts. Additional examples not described in the text are provided in the Additional file 5 (Figures S7–S8 for $C_{10}H_{14}N_2$, the formula of nicotine, and $C_{17}H_{21}NO$, the formula of diphenhydramine, respectively).

Results and discussion

Linking metadata via MS-Ready structures

It has been demonstrated that data sources and other metadata linked to chemical structures improve identification of unknowns [7, 15, 55]. Substances in the Dashboard contain different linked metadata [22], making access to all forms of a chemical structure important for identification (Fig. 3). Beyond data sources alone, chemical functional use and product occurrence data [56, 57] are metadata that can help analysts arrive at the source of a chemical in a sample through mapping via MS-Ready structures. Nicarbazin (DTXSID6034762, $C_{19}H_{18}N_6O_6$ [58]), a coccidiostat used in poultry production, is a two component chemical (with the associated formulae for the two separate structures being $C_{13}H_{10}N_4O_5$ and $C_6H_8N_2O$) whose components would dissociate in the environment, leading to the observation of individual components only via HRMS. Neither of the single components has known commercial uses (yet) that would result in environmental occurrence. By mapping the two

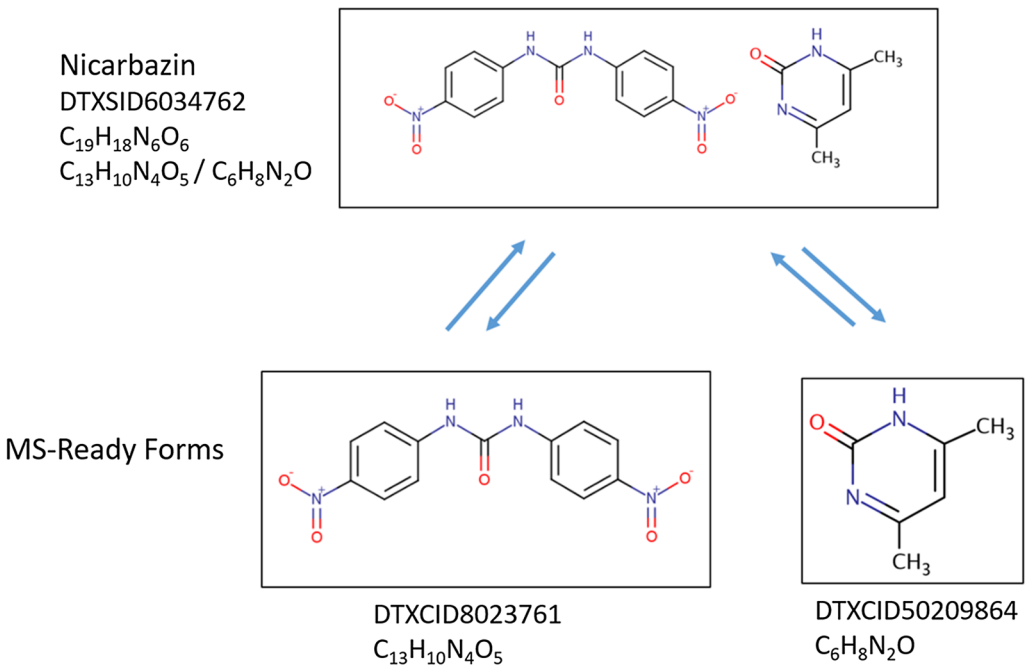

Fig. 4 The substance Nicarbazin (DTXSID6034762) and its two components (DTXCID8023761; DTXCID50209864), separated as a result of the MS-Ready workflow. The MS-Ready forms are linked to the source substance and enable retrieval of associated structures and metadata through a single database query

observable components to the source substance, the analyst is potentially able to identify the substance likely used in commerce with an observed formula search (Fig. 4), thereby improving exposure characterization where accurate identification of source substances is critical. Furthermore, the presence of one part of a component may indicate the presence of the other component in the sample, triggering further identifications. Informing the analyst of the most likely substance, rather than just the chemical structure identified by HRMS, may allow decision makers and risk assessors the ability to link chemical identifications and substances. The application of this during candidate selection in non-target screening is discussed further below.

Non-target collaborative trials

In 2013, the NORMAN Network coordinated a collaborative non-targeted screening trial on a river water sample [2]. Several examples from this trial indicated the need for improved curation of chemical structures as well as better metadata linkage across substances in a sample during non-targeted screening. Participants reported, for instance, mass matches to the salt form of a substance in a suspect list (e.g., tris[4-(diethylamino)phenyl]methylium acetate, $C_{31}H_{42}N_3.C_2H_3O_2$ reported at *m/z* 516.3565 by one participant, which could not be observed in the sample as the acetate would dissociate). Using MS-Ready structures can reduce errors associated with identifying salt forms by searching at the single component level and returning mapped substances. The complex nature of considering metadata and sample context in non-target identification is further demonstrated with the tentative annotations provided for the masses $m/z = 229.1094$ and 201.0781 (see Fig. 5, adapted from Fig. 2 in [2]). For $m/z = 229.1094$, most participants provided the tentative annotation for terbutylazine (DTXSID4027608, which many participants had as a target analyte). Propazine (DTXSID3021196) is not approved for use in Europe and should not be detected in typical environmental samples, yet it was still reported three times due to the high reference count. For $m/z = 201.0781$, the presence of terbutylazine provides strong evidence to support the tentative annotation of desethylterbutylazine (DTXSID80184211), although many participants reported simazine (DTXSID4021268) due to its higher reference count (Fig. 5). Simazine and desethylterbutylazine (with the often co-eluting desethylsebutylazine, DTXSID20407557) can often be distinguished using fragmentation information.

The EPA's Non-Targeted Analysis Collaborative Trial (ENTACT) was initiated following the NORMAN collaborative trial [2]. ENTACT is an inter-laboratory trial where participating laboratories and institutions were provided blinded chemical mixtures and environmental samples for NTA and SSA [59, 60]. The blinded

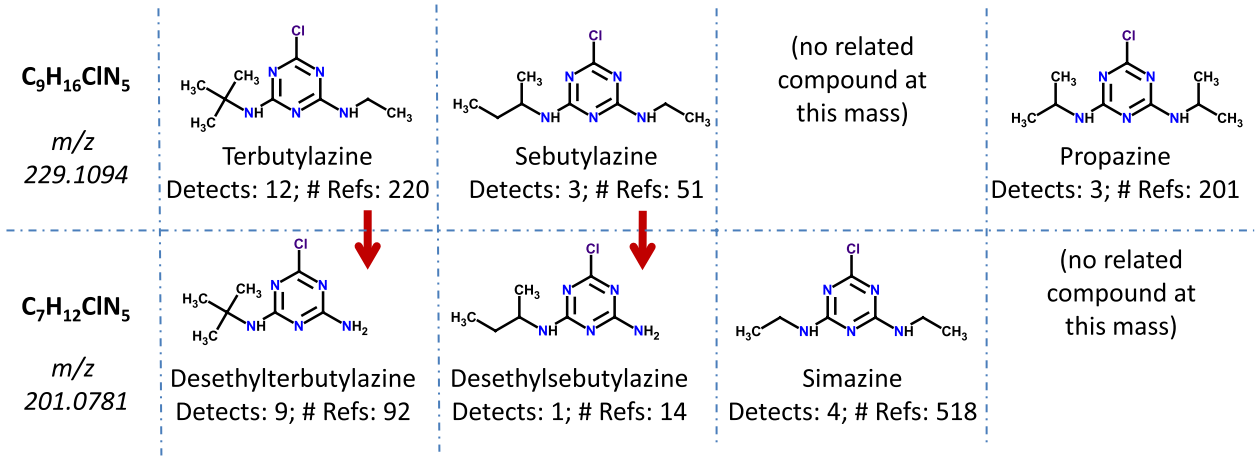

Fig. 5 Tentative annotations of *m/z* 229.1094 (top) and *m/z* 201.0781 provided by NORMAN Collaborative Trial participants. Number of detects indicates the number of participants in the collaborative trial who provided the structural annotation of the selected compound. Reference data is from ChemSpider. Source data and figure modified from Schymanski et al. [2]

chemical mixtures included several multi-component substances that could be either mismatched or unidentified without a linkage between the MS-Ready form of a chemical structure and its multi-component form (e.g., chemical mixtures, salts). For the purposes of ENTACT, identification of the original substances added to the mixtures is critical to the trial evaluation. Methapyrilene fumarate (DTXSID0047404 [61]), for example, is a mixture of two chemical components (in a 3:2 ratio) that would be observed separately (DTXCID003278 [62]; DTXCID8028133 [63]), while raloxifene hydrochloride (DTXSID1034181 [64]) is a substance containing a hydrochloride salt that would be matched incorrectly from MS data without the appropriate standardization and linking. Linking the MS-Ready forms of these chemicals to the substance forms facilitates identification by including all variants in the search results with associated metadata. For example, blinded analysis of one of the ENTACT mixtures resulted in the observation of $m/z = 262.1385$ in ESI+ (Sobus et al. submitted for publication). With this exact mass and associated isotopic peaks, the formula $C_{14}H_{19}N_3S$ was generated. When the formula was searched in the Dashboard ($C_{14}H_{19}N_3S$ [65]) the results included both the single component methapyrilene (DTXSID2023278 [66]) and the multi-component methapyrilene fumarate (DTXSID0047404 [61]) in the top 5 results as ranked by data source count. An exact formula search would not have returned the substance originally added to the ENTACT mixture, which was in fact methapyrilene fumarate. The MS-Ready search in the Dashboard and the linkages are especially beneficial when the structures identified by HRMS differ from the form of the substance initially contained within the mixture (e.g., Fig. 4). In addition to the Dashboard MS-Ready functionality in the user interface, files containing MS-Ready forms of the chemical structures, mapped to the original chemical substances contained within the mixtures, were provided to the participants as part of ENTACT and are available via the Dashboard as an Excel spreadsheet [44].

Enhanced searching: an example with perfluorinated chemicals

With an increasing focus on perfluorinated chemicals and their effects on the environment and public health [67–71], it is not only important to be able to accurately identify perfluorinated structures in environmental samples but also to identify the potential sources of the contaminant for exposure characterization. Perfluorinated chemicals also present a challenge for NTA, as the presence of monoisotopic fluorine renders calculation of possible molecular formulae very challenging [5, 72]. As a result, SSA and compound database searching is advantageous to finding these compounds. Perfluorosulfonic acids (e.g., PFOS, DTXSID3031864 [73]), perfluorocarboxylic acids (e.g., PFOA, DTXSID8031865 [74]), and other similar structures are thought to occur in the environment as anions [67]. Hence, these structures are often reported in the literature as anions, but have also been reported as neutral acids. In chemical databases these structures can be represented in their neutral forms, as a part of chemical mixtures, and as multi-component salts (e.g., PFOS-K, DTXSID8037706 [75]), representing the myriad of chemical forms available in commerce (see the linked MS-Ready substances for PFOS currently in the Dashboard [76]). PFOS would generally be observed by

an analyst via HRMS as a negatively charged *m/z* feature ($C_8F_{17}O_3S^-$), and when a neutral monoisotopic mass is calculated, the analyst is likely to arrive at the molecular formula of the neutral acid form of PFOS ($C_8HF_{17}O_3S$). Searching the neutral formula of PFOS ($C_8HF_{17}O_3S$) in the Dashboard MS-Ready Batch Search option returns the neutral acid, the sulfonate ($C_8F_{17}O_3S^-$), and multiple salts and mixtures containing PFOS in the results list (Fig. 6). These results include the neutral form and the substance forms thought to occur in the environment and used in consumer products/commerce, along with associated metadata. Many forms of PFOS may be contained in other public databases, and other strategies have been developed to counteract the anion/neutral form issue during compound searching (e.g., UC2 by Sakurai et al. [77]). The current MS-Ready functionality in the Dashboard provides mappings to multiple forms of chemicals related via their "MS-Ready" form in a single search, improving researchers' ability to identify sources and improve exposure characterization with increased coverage and access to metadata.

Non-target identification: in silico methods and candidate searching

In this section two examples from the NORMAN Collaborative Trial (Fig. 5) are used to show how the MS-Ready form of a mixture will help analysts combine MS evidence (such as fragments) with mixture metadata for candidate screening in NTA. By crosslinking with the MS-Ready form through the export format described above, the candidates can be processed using MS-Ready structures, with metadata from the mixture in MetFrag. As described in the Methods (*MetFrag Example Calculations*), two MetFrag scoring terms plus five metadata terms were used, which would result in a maximum possible score of 7 for candidates in each example.

The results for the top three candidates from the first example, $C_9H_{16}ClN_5$, using fragmentation data from terbutylazine are shown in Fig. 7. This demonstrates how the combination of fragmentation prediction, MS/MS library matching, and metadata supports the annotation of terbutylazine (MetFrag Score 7.0, including an exact spectral match of 1.0 from MoNA—i.e., a Level 2a identification [24]) above propazine (MetFrag Score 5.5, exact spectral match 0.5774, i.e., a poor match). The presence of the $C_4H_9^+$ fragment at $m/z = 57.0698$, explained by MetFrag, indicates the presence of a butyl substituent, absent from propazine (Fig. 8). Sebutylazine, the third candidate, has a much lower score due to fewer metadata (see Fig. 7), although the fragmentation data is very similar to terbutylazine (Fig. 8).

The second example, the MS-Ready search for $C_7H_{12}ClN_5$ with the spectral data of desethylterbutylazine, was run with the same settings, but with the candidate grouping activated. The top three candidates from the MetFrag web interface [47] are given in Fig. 9 and detailed scores are provided in Additional file 5: Table S4. The top-ranked candidate with the selected metadata and default scoring is simazine (Score 4.98 of maximum 7.0). It is also clear from the numerous DTXSID values

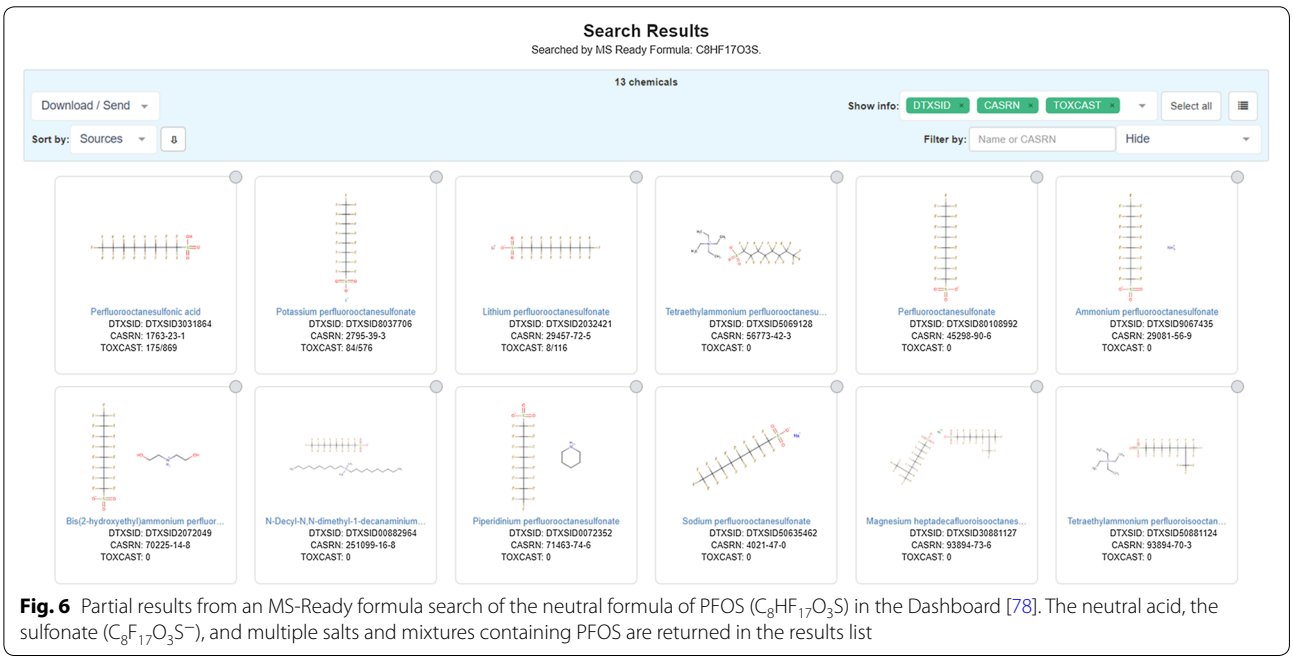

Fig. 6 Partial results from an MS-Ready formula search of the neutral formula of PFOS ($C_8HF_{17}O_3S$) in the Dashboard [78]. The neutral acid, the sulfonate ($C_8F_{17}O_3S^-$), and multiple salts and mixtures containing PFOS are returned in the results list

Fig. 7 MetFrag combined results (top) and results for the top 3 candidates (bottom) retrieved with the MS-Ready search for $C_9H_{16}ClN_5$. The score categories are (1st to 7th): MetFrag Fragmentation, Exact Spectral Similarity, Data Sources, Presence in NORMAN Priority list, Number of PubMed Articles, Presence in STOFF-IDENT, and Percent Active ToxCast Assays. Terbutylazine had the highest score, above propazine. Sebutylazine (which, if present, often co-elutes with terbutylazine in common NTA methods) has a lower score due to fewer metadata values (absent from NORMAN list and no ToxCast bioassay data)

displayed in the "Identifier" column for simazine that there are many substances (mixtures, salts) in the Dashboard that contain simazine as one component (11 of the 21 candidates returned in the MS-Ready search). Desethylterbutylazine is in second place with a score of 4.26. Additional file 5: Figs. S7 and S8 show MetFrag results for additional searches correctly placing nicotine (DTXSID1020930) and diphenhydramine (DTXSID4022949) as the top result, respectively, with the same metadata options included and candidate grouping activated.

The example in Fig. 9 demonstrates how users must think critically about the impact of the metadata on the results. While simazine (Score 4.98) outranks desethylterbutylazine (Score 4.26), closer inspection reveals this result is due to metadata score influence. The experimental data (fragmentation prediction, peaks explained, spectral similarity, exact spectral similarity) matches better for desethylterbutylazine (6/8 peaks explained and scores close to or equal to 1 for the other experimental fields) than for simazine. Desethylterbutylazine does not

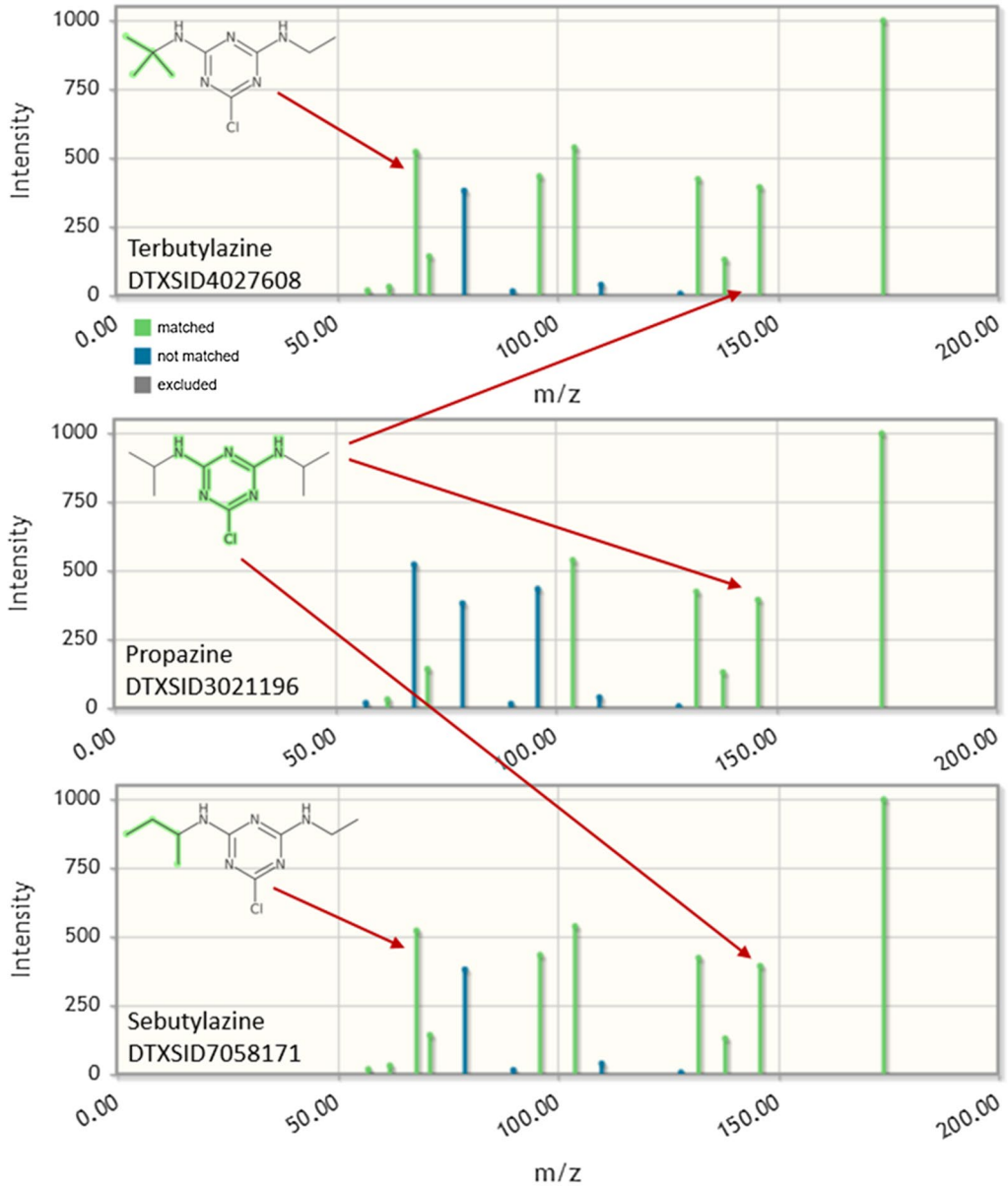

Fig. 8 MetFrag Fragmentation results for the top three candidates retrieved with the MS-Ready search for $C_9H_{16}ClN_5$. Terbutylazine (top) has the highest score and includes the $C_4H_9^+$ fragment at $m/z = 57.0698$ indicating the presence of a butyl substituent, absent from propazine (middle)

have a ToxCast Bioassay score and has no PubMed references, resulting in two zero scores, while simazine has a score of 1 for both of these metadata categories. Furthermore, while the MetFrag website [47] provides users with a convenient interface to score with a tick-box, users must be aware of the limitations inherent in providing a convenient interface. The data in each external category is imported and scaled between 0 and 1 using the minimum and maximum values, which is not meaningful for all metadata categories (such as predicted properties). Note that it is possible to adjust the weighting and relative contributions of the scores by adjusting the bars on the "Weights" field at the top of the results page (once candidates are processed), while additional scoring possibilities are available via the command line version.

Improvements and future work

Beyond access to structures and workflows via the Dashboard, future functionality of the Dashboard will allow for users to upload structure files and receive back the

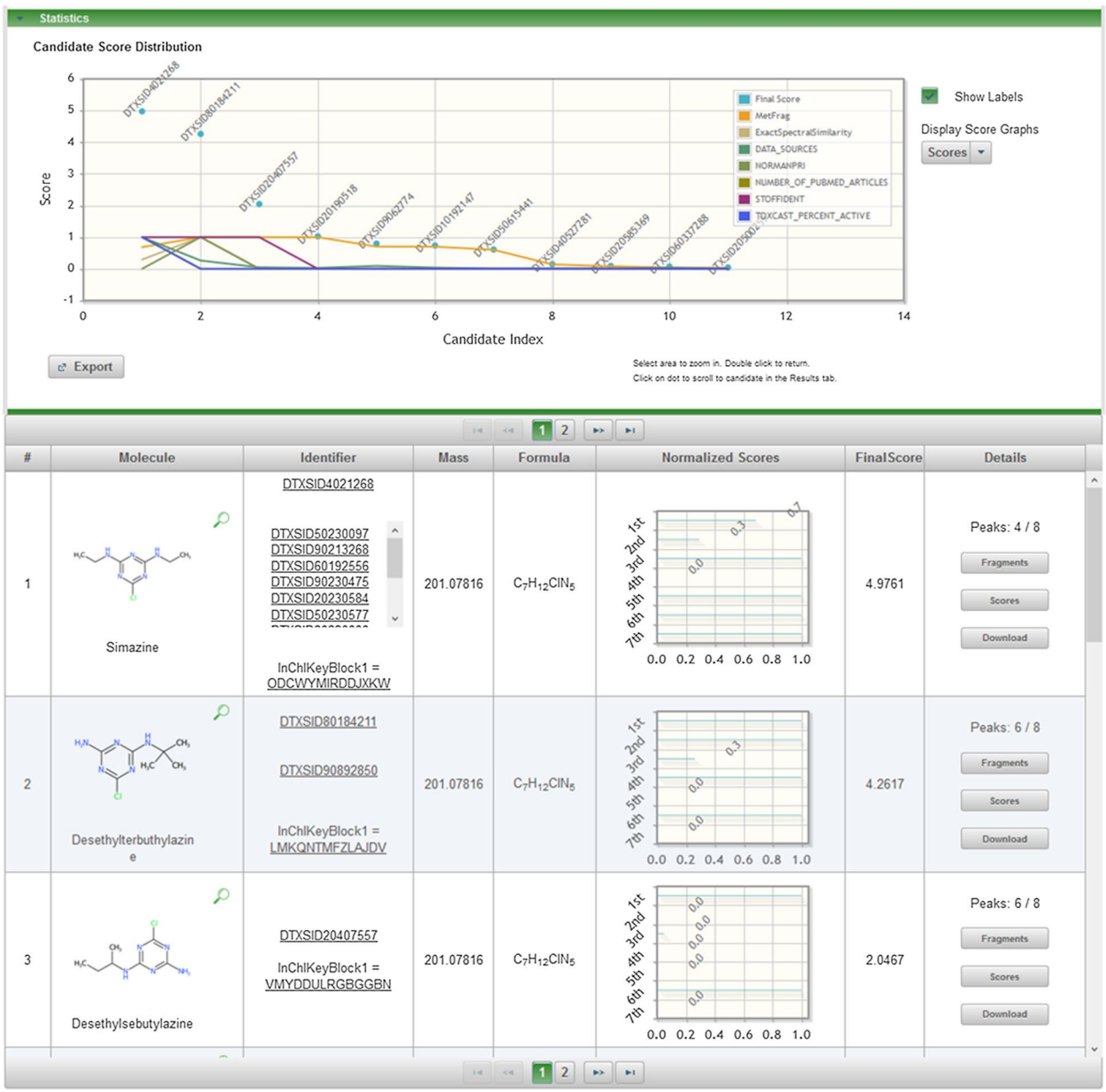

Fig. 9 MetFrag combined results (top) and results for the top 3 candidates retrieved with the MS-Ready search for $C_7H_{12}ClN_5$ (as displayed in the web interface). The score categories are (1st to 7th): MetFrag Fragmentation, Exact Spectral Similarity, Data Sources, Presence in NORMAN Priority list, Number of PubMed Articles, Presence in STOFF-IDENT, and Percent Active ToxCast Assays. Candidate merging was activated and the 10 forms of simazine have been merged into one result (with metadata from the highest scoring entry)

MS-Ready version of the structures of interest, increasing standardization across database searching and compound identification. Alterations to the output format (as described in the Methods) will enable other in silico fragmentation and compound identification tools, methods, and software to use the work described here. Further flexibility in file formats will be implemented to achieve broader usability. As with any chemical structure standardization workflow, algorithms are modified to deal with edge cases as they are identified. As the database content continues to expand, the algorithm is improved as failures are identified. While the MS-Ready approach may lead to potentially confusing results sets containing structures with different formulae and masses than specified in the original search parameters, communication, education, and transparency within the

Dashboard interface, download files, and publications will serve to clarify and provide guidance. Finally, to facilitate access to the underlying data for structure identification on the broadest scale, an application programming interface (API) and associated web services to allow instrument software integration is forthcoming. These will enable access via applications such as Python, R, and Matlab to facilitate integration of Dashboard data into user-specific applications.

Conclusions

Database searching is a vital part of NTA and SSA workflows. The accurate mapping of MS-Ready structures to chemical substances improves accessibility to structure metadata and improves searching of the represented chemical space. By providing access to MS-Ready data from DSSTox, both via the Dashboard and as downloadable datasets, users of HRMS instrumentation who perform NTA/SSA experiments will benefit from this approach as an enhancement to other online databases that do not support MS-Ready structural forms. The integration into the in silico fragmenter MetFrag lets users further explore the use of this approach in identification of unknowns. The openly available workflow for generation of MS-Ready structures allows others to process their own data for preparation of MS-Ready data files and extend the data handling to account for errors and specific cases that we have not yet identified.

Additional files

Additional file 1. MS-Ready exclusion list.

Additional file 2. QSAR-Ready exclusion list.

Additional file 3. CompTox Chemistry Dashboard search interfaces (**Figures S1–S4**).

Additional file 4. Download file column header descriptions and example output files for MS-Ready and MetFrag Input File batch searches (**Tables S1–S3**).

Additional file 5. Additional MetFrag results and data (**Figures S5–S8**, **Table S4**).

Abbreviations

HRMS: high-resolution mass spectrometry; DSSTox: distributed structure-searchable toxicity; ENTACT: EPA's non-targeted analysis collaborative trial; QSAR: quantitative structure activity relationship; NTA: non-targeted analysis; SSA: suspect screening analysis.

Authors' contributions

AJW is the project lead for the CompTox Chemistry Dashboard. CG is responsible for the development of the DSSTox database and chemical registration software that houses the chemistry data (including all integrated modules). ADM supports the development of the mass spectrometry structure identification aspects of the Dashboard. KM is the developer of the OPERA prediction models and QSAR/MS-Ready structure generation workflows. ELS tested the application of the MS-Ready approach in a research setting, its application to support data integration with online resources and provided feedback to optimize the Dashboard interface and MetFrag output files. CR is the lead developer for the MetFrag application and collaborated with ELS and AJW on the MetFrag/Dashboard integration. CR implemented all necessary code changes. ADM, AJW, CG, CR, ELS and KM all participated in manuscript preparation. All authors read and approved the final manuscript.

Authors' information

Dr. Andrew McEachran received a B.S. from NC State University, an M.S. from Texas Tech University, and a Ph.D. from NC State University, focusing on environmental analytical chemistry. Currently he is an ORISE Postdoctoral Fellow in the National Center for Computational Toxicology within the U. S. Environmental Protection Agency. There his research focuses on improving the identification of unknown chemicals in environmental samples using non-targeted analyses.

Dr. Kamel Mansouri is a computational chemist who obtained an engineering degree in analytical chemistry from the University of Tunis, Tunisia, an M.S. degree in cheminformatics from the University of Strasbourg, France, and a Ph.D. in computational chemistry from the University of Milano Bicocca, Italy. He joined the National Center for Computational Toxicology at the U.S. Environmental Protection Agency as an ORISE Postdoctoral Fellow in 2013 and worked on several projects involving QSAR modeling, cheminformatics, and data-mining, and has collaborated and led projects in the QSAR field with renowned international scientists. He is presently leading the computational chemistry efforts at Integrated Laboratory Systems, Inc. supporting the NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) at the U. S. National Institute of Environmental Health Sciences (NIEHS).

Dr. Chris Grulke received a B.S.E. from the University of Michigan in Chemical Engineering (2003) and a Ph.D. in Pharmaceutical Science, Medicinal Chemistry, and Biophysics from the University of North Carolina at Chapel Hill (2011). He is currently employed as a Cheminformatician at the U.S. Environmental Protection Agency's National Center for Computational Toxicology. Dr. Grulke is applying advanced database and software development skills to building a cheminformatics infrastructure for integrating chemical and biological data to support the development of predictive models pertaining to exposure, pharmacokinetics, and toxicity.

Dr. Emma Schymanski received her Ph.D. from the Technical University Bergakadamie Freiberg in 2011, undertaking research located at the Helmholtz Centre for Environmental Research (UFZ) in Leipzig. Following 6 years of postdoctoral research at the Swiss Federal Institute for Aquatic Science and Technology (Eawag), she recently joined the Luxembourg Centre for Systems Biomedicine, University of Luxembourg to establish the Environmental Cheminformatics group. Her research combines cheminformatics and computational mass spectrometry approaches to elucidate the unknowns in complex samples, primarily with non-target screening. An advocate for open science, she is involved in and organizes several activities to improve the exchange of data, information and ideas between scientists.

Christoph Ruttkies received his diploma in bioinformatics from the Martin Luther University of Halle-Wittenberg, Germany in a collaborative effort with the Institute of Pharmacy. He joined the Mass Spectrometry and Bioinformatics research group at the Leibniz Institute of Plant Biochemistry in Halle, Germany as a Ph.D. student to develop computational methods for compound annotation and identification mainly based on mass spectral data. He is presently part of a European DevOps team in the PhenoMeNal-H2020 project to integrate software tools into workflows for a cloud-based metabolomics data analysis platform.

Dr. Antony Williams received a Ph.D. in analytical chemistry (NMR) from the University of London, UK in 1988. He ran NMR facilities in both academia and US-based Fortune 500 companies. He joined ACD/Labs as their Chief Science Officer with a focus on structure representation, nomenclature, and analytical data management. He was a founder of the ChemSpider chemistry database, later acquired by the Royal Society of Chemistry. In 2015, he joined the National Center for Computational Toxicology within the U.S. Environmental Protection Agency as a computational chemist and is presently focused on the development of Web-based applications to access chemistry data.

Author details
[1] Oak Ridge Institute for Science and Education (ORISE) Research Participation Program, U.S. Environmental Protection Agency, 109 T.W. Alexander Dr., Research Triangle Park, NC 27711, USA. [2] National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Mail Drop D143-02, 109 T.W. Alexander Dr., Research Triangle Park, NC 27711, USA. [3] Present Address: Integrated Laboratory Systems, Inc., 601 Keystone Dr., Morrisville, NC 27650, USA. [4] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6, avenue du Swing, 4367 Belvaux, Luxembourg. [5] Department of Stress and Development Biology, Leibniz Institute of Plant Biochemistry (IPB), Weinberg 3, 06120 Halle (Saale), Germany.

Acknowledgements
We are indebted to the NCCT development team and IT support staff who are involved with the day to day development of the Dashboard. Specifically, we acknowledge Jeff Edwards and Jeremy Dunne. We acknowledge all curators of the DSSTox chemistry database underlying the dashboard that have contributed to over 15 years of curation efforts. We thank the management of our center, Russell Thomas (Director), Kevin Crofton and Sandra Roberts for their belief in our efforts and support to create a difference with this developing architecture and application. We have the pleasure of working with scientists in the National Environmental Research Laboratory (NERL) on the ENTACT project and acknowledge Jon Sobus, Elin Ulrich and Seth Newton for their feedback on the Dashboard and this manuscript. This work was supported in part by an appointment to the ORISE Research Participation Program at the Office of Research and Development, U.S. EPA, through an interagency agreement between the U.S. EPA and U.S. Department of Energy. This work was also supported in part by the Pathfinder Innovation Project (PIP) awarded by the EPA Office of Research and Development. This work has been internally reviewed at the US EPA and has been approved for publication. ES would like to acknowledge those involved in the NORMAN Suspect Exchange initiative, especially Reza Aalizadeh, for discussions. ELS and CR gratefully acknowledge the efforts of Steffen Neumann in the development of MetFrag. 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.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
The dataset(s) supporting the conclusions of this article are available via the CompTox Chemistry Dashboard Downloads Page (https://comptox.epa.gov/dashboard/downloads) and MS-Ready GitHub repository (https://github.com/kmansouri/MS-ready). The MetFrag functionality is available through the web interface (https://msbi.ipb-halle.de/MetFragBeta/) and the command line version (http://c-ruttkies.github.io/MetFrag/projects/metfragcl/). All additional data supporting the conclusions of this article are included within the article and its additional files.

Funding
The United States Environmental Protection Agency, through its Office of Research and Development, funded and managed the research described here for ADM, AJW, CG, and KM. It has been subjected to Agency administrative review and approved for publication. AM and KM were supported by an appointment to the Internship/Research Participation Program at the Office of Research and Development, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA. CR acknowledges funding from EU H2020 project PhenoMeNal under Grant Agreement No. 654241, CR and ES acknowledge funding from EU FP7 project SOLUTIONS under Grant Agreement No. 603437.

Publisher's Note

Received: 16 May 2018 Accepted: 21 August 2018
Published online: 30 August 2018

References

1. Hollender J, Schymanski EL, Singer HP, Ferguson PL (2017) Nontarget screening with high resolution mass spectrometry in the environment: ready to go? Environ Sci Technol 51(20):11505–11512. https://doi.org/10.1021/acs.est.7b02184
2. Schymanski EL, Singer HP, Slobodnik J, Ipolyi IM, Oswald P, Krauss M et al (2015) Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. Anal Bioanal Chem. https://doi.org/10.1007/s00216-015-8681-7
3. Rager JE, Strynar MJ, Liang S, McMahen RL, Richard AM, Grulke CM et al (2016) Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring. Environ Int 88:269–280. https://doi.org/10.1016/j.envint.2015.12.008
4. Warth B, Spangler S, Fang M, Johnson CH, Forsberg EM, Granados A et al (2017) Exposome-scale investigations guided by global metabolomics, pathway analysis, and cognitive computing. Anal Chem 89(21):11505–11513. https://doi.org/10.1021/acs.analchem.7b02759
5. Newton SR, McMahen RL, Sobus JR, Mansouri K, Williams AJ, McEachran AD et al (2018) Suspect screening and non-targeted analysis of drinking water using point-of-use filters. Environ Pollut 234:297–306. https://doi.org/10.1016/j.envpol.2017.11.033
6. Krauss M, Singer H, Hollender J (2010) LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns. Anal Bioanal Chem 397(3):943–951. https://doi.org/10.1007/s00216-010-3608-9
7. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminform 8(1):1–16. https://doi.org/10.1186/s13321-016-0115-9
8. Little J, Williams A, Pshenichnov A, Tkachenko V (2012) Identification of known unknowns utilizing accurate mass data and ChemSpider. J Am Soc Mass Spectrom. https://doi.org/10.1007/s13361-011-0265-y
9. McEachran AD, Sobus JR, Williams AJ (2017) Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. Anal Bioanal Chem 409(7):1729–1735. https://doi.org/10.1007/s00216-016-0139-z
10. Stein S (2012) Mass spectral reference libraries: an ever-expanding resource for chemical identification. Anal Chem. https://doi.org/10.1021/ac301205z
11. 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(7):703–714. https://doi.org/10.1002/jms.1777
12. Aalizadeh R, Thomaidis NS, Bletsou AA, Gago-Ferrero P (2016) Quantitative structure–retention relationship models to support nontarget high-resolution mass spectrometric screening of emerging contaminants in environmental samples. J Chem Inf Model 56(7):1384–1398. https://doi.org/10.1021/acs.jcim.5b00752
13. Bade R, Bijlsma L, Sancho JV, Hernández F (2015) Critical evaluation of a simple retention time predictor based on LogKow as a complementary tool in the identification of emerging contaminants in water. Talanta. https://doi.org/10.1016/j.talanta.2015.02.055
14. McEachran AD, Mansouri K, Newton SR, Beverly BEJ, Sobus JR, Williams AJ (2018) A comparison of three liquid chromatography (LC) retention time prediction models. Talanta. https://doi.org/10.1016/j.talanta.2018.01.022
15. Blaženović I, Kind T, Torbašinović H, Obrenović S, Mehta SS, Tsugawa H et al (2017) Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: database boosting is needed to achieve 93% accuracy. J Cheminform 9(1):32
16. Pence HE, Williams A (2010) ChemSpider: an online chemical information resource. J Chem Educ 87(11):1123–1124
17. CompTox Chemistry Dashboard: DTXSID4022949. https://comptox.epa.gov/dashboard/DTXSID4022949. Accessed 1 Mar 2018
18. CompTox Chemistry Dashboard: DTXCID802949. https://comptox.epa.gov/dashboard/dsstoxdb/results?utf8=✓&search=DTXCID802949. Accessed 1 Mar 2018
19. CompTox Chemistry Dashboard: DTXSID80237211. https://comptox.epa.gov/dashboard/DTXSID80237211. Accessed 1 Mar 2018
20. CompTox Chemistry Dashboard: DTXSID4020537. https://comptox.epa.gov/dashboard/DTXSID4020537. Accessed 1 Mar 2018
21. CompTox Chemistry Dashboard: DTXSID10225883. https://comptox.epa.gov/dashboard/DTXSID10225883. Accessed 1 Mar 2018

22. Schymanski EL, Williams AJ (2017) Open science for identifying "known unknown" chemicals. Environ Sci Technol. https://doi.org/10.1021/acs.est.7b01908
23. Kind T, Scholz M, Fiehn O (2009) How large is the metabolome? A critical analysis of data exchange practices in chemistry. PLoS ONE 4(5):e5440
24. Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP et al (2014) Identifying small molecules via high resolution mass spectrometry: communicating confidence. Environ Sci Technol 48(4):2097–2098
25. Williams AJ, Ekins S, Tkachenko V (2012) Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. Drug Discov Today 17(13):685–701. https://doi.org/10.1016/j.drudis.2012.02.013
26. Karapetyan K, Batchelor C, Sharpe D, Tkachenko V, Williams AJ (2015) The chemical validation and standardization platform (CVSP): large-scale automated validation of chemical structure datasets. J Cheminform 7(1):30. https://doi.org/10.1186/s13321-015-0072-8
27. Fourches D, Muratov E, Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 50(7):1189–1204
28. Mansouri K, Grulke CM, Richard AM, Judson RS, Williams AJ (2016) An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling. SAR QSAR Environ Res. https://doi.org/10.1080/1062936x.2016.1253611
29. Young D, Martin T, Venkatapathy R, Harten P (2008) Are the chemical structures in your QSAR correct? QSAR Comb Sci 27(11–12):1337–1345. https://doi.org/10.1002/qsar.200810084
30. Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A et al (2016) CERAPP: collaborative estrogen receptor activity prediction project. J Environ Health Perspect. https://doi.org/10.1289/ehp.1510267
31. Mansouri K. MS-Ready GitHub repository. https://github.com/kmansouri/MS-ready. Accessed 30 Apr 2018
32. Richard AM, Williams CR (2002) Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. Mutat Res/Fundam Mol Mech Mutagen 499(1):27–52. https://doi.org/10.1016/S0027-5107(01)00289-5
33. Williams AJ, Grulke CM, Edwards J, McEachran AD, Mansouri K, Baker NC et al (2017) The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J Cheminform 9(1):61. https://doi.org/10.1186/s13321-017-0247-6
34. Mansouri K. QSAR-Ready GitHub repository. https://github.com/kmansouri/QSAR-Ready. Accessed 30 Apr 2018
35. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T et al (2008) KNIME: the Konstanz information miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R (eds) Data analysis, machine learning and applications: proceedings of the 31st annual conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007. Springer, Berlin, pp 319–326
36. EPAM (2016) INDIGO toolkit. http://lifescience.opensource.epam.com/indigo/
37. Sitzmann M, Ihlenfeldt W-D, Nicklaus MC (2010) Tautomerism in large databases. J Comput Aided Mol Des 24(6–7):521–551
38. ChemAxon (2014) Standardizer. Structure canonicalization and more. https://chemaxon.com/products/chemical-structure-representation-toolkit
39. Reusch W (2013) Reaction examples: examples of organic reactions. http://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/react2.htm
40. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I (2013) InChI—the worldwide chemical structure identifier standard. J Cheminform. https://doi.org/10.1186/1758-2946-5-7
41. ChemAxon Java JChem API. https://apidocs.chemaxon.com/jchem/doc/dev/java/api/. Accessed 18 Sept 2017
42. CompTox Chemistry Dashboard: Advanced Search. https://comptox.epa.gov/dashboard/dsstoxdb/advanced_search. Accessed 3 Apr 2018
43. CompTox Chemistry Dashboard: Batch Search. https://comptox.epa.gov/dashboard/dsstoxdb/batch_search. Accessed 3 Apr 2018
44. CompTox Chemistry Dashboard: Downloads. https://comptox.epa.gov/dashboard/downloads. Accessed 5 Feb 2018
45. CompTox Chemistry Dashboard: MS-Ready Search of C10H14N2. https://comptox.epa.gov/dashboard/dsstoxdb/multiple_results?utf8=%E2%9C%93&inputs%5B%5D=C10H14N2&input_type=ms_ready_formula. Accessed 1 Mar 2018
46. CompTox Chemistry Dashboard: Exact Formula Search of C10H14N2. https://comptox.epa.gov/dashboard/dsstoxdb/multiple_results?utf8=%E2%9C%93&inputs%5B%5D=C10H14N2&input_type=exact_formula. Accessed 1 Mar 2018
47. MetFrag. https://msbi.ipb-halle.de/MetFragBeta/. Accessed 30 Mar 2018
48. European MassBank. http://www.massbank.eu/. Accessed 30 Jan 2018
49. CompTox Chemistry Dashboard: DTXSID4027608. https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXSID4027608. Accessed 30 Jan 2018
50. MassBank Record EA028496. https://massbank.eu/MassBank/jsp/RecordDisplay.jsp?id=EA028406&dsn=Eawag. Accessed 30 Jan 2018
51. MassBank Record EA067106. https://massbank.eu/MassBank/jsp/RecordDisplay.jsp?id=EA067106&dsn=Eawag. Accessed 30 Jan 2018
52. CompTox Chemistry Dashboard: Norman Priority List. https://comptox.epa.gov/dashboard/chemical_lists/normanpri. Accessed 30 Jan 2018
53. CompTox Chemistry Dashboard: STOFF-IDENT List. https://comptox.epa.gov/dashboard/chemical_lists/stoffident. Accessed 30 Jan 2018
54. MoNA-MassBank of North America. http://mona.fiehnlab.ucdavis.edu/. Accessed 30 Mar 2018
55. Schymanski EL, Ruttkies C, Krauss M, Brouard C, Kind T, Dührkop K et al (2017) Critical assessment of small molecule identification 2016: automated methods. J Cheminform 9(1):22. https://doi.org/10.1186/s13321-017-0207-1
56. Dionisio KL, Phillips K, Price PS, Grulke CM, Williams A, Biryol D et al (2018) The chemical and products database, a resource for exposure-relevant data on chemicals in consumer products. Nat Sci Data 5:180125
57. Phillips K, Yau AY, Favela KA, Isaacs KK, McEachran A, Grulke CM et al (2018) Suspect screening analysis of chemicals in consumer products. Environ Sci Technol. https://doi.org/10.1021/acs.est.7b04781
58. CompTox Chemistry Dashboard: DTXSID6034762. https://comptox.epa.gov/dashboard/dsstoxdb/results?utf8=%E2%9C%93&search=DTXSID6034762. Accessed 30 Jan 2018
59. Sobus JR, Wambaugh J, Isaacs K, Williams A, McEachran A, Richard A et al (2017) Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. J Expo Sci Environ Epidemiol. https://doi.org/10.1038/s41370-017-0012-y
60. Ulrich EM, Sobus JR, Grulke CM, Richard A, Newton S, Strynar MJ et al (2018) EPA's non-targeted analysis collaborative trial (ENTACT): genesis, design, and initial findings. Anal Bioanal Chem (in press)
61. CompTox Chemistry Dashboard: DTXSID0047404. https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXSID0047404. Accessed 15 Jan 2018
62. CompTox Chemistry Dashboard: DTXCID003278. https://comptox.epa.gov/dashboard/dsstoxdb/results?utf8=%E2%9C%93&search=DTXCID003278. Accessed 15 Jan 2018
63. CompTox Chemistry Dashboard: DTXCID8028133. https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXCID8028133. Accessed 15 Jan 2018
64. CompTox Chemistry Dashboard: DTXSID1034181. https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXSID1034181. Accessed 15 Jan 2018
65. CompTox Chemistry Dashboard: MS-Ready Search of C14H19N3S. https://comptox.epa.gov/dashboard/dsstoxdb/multiple_results?utf8=%E2%9C%93&search_inputs=C14H19N3S&search_type=ms_ready_formula. Accessed 15 Jan 2018
66. CompTox Chemistry Dashboard: DTXSID2023278. https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXSID2023278. Accessed 15 Jan 2018
67. Lindstrom AB, Strynar MJ, Libelo EL (2011) Polyfluorinated compounds: past, present, and future. Environ Sci Technol 45(19):7954–7961. https://doi.org/10.1021/es2011622
68. Sun M, Arevalo E, Strynar M, Lindstrom A, Richardson M, Kearns B et al (2016) Legacy and emerging perfluoroalkyl substances are important drinking water contaminants in the Cape Fear River Watershed of North Carolina. Environ Sci Technol Lett 3(12):415–419. https://doi.org/10.1021/acs.estlett.6b00398
69. Newton S, McMahen R, Stoeckel JA, Chislock M, Lindstrom A, Strynar M (2017) Novel polyfluorinated compounds identified using high resolution mass spectrometry downstream of manufacturing facilities near Decatur,

Alabama. Environ Sci Technol 51(3):1544–1552. https://doi.org/10.1021/acs.est.6b05330
70. United States Environmental Protection Agency (2016) Drinking water health advisory for perfluorooctane sulfonate (PFOS). Office of Water, Washington, DC. https://www.epa.gov/sites/production/files/2016-05/documents/pfos_health_advisory_final_508.pdf
71. United States Environmental Protection Agency (2016) Drinking water health advisory for perfluorooctanoic acid (PFOA). Office of Water, Washington, DC. https://www.epa.gov/sites/production/files/2016-05/documents/pfoa_health_advisory_final_508.pdf
72. Trier X, Granby K, Christensen JH (2011) Tools to discover anionic and nonionic polyfluorinated alkyl surfactants by liquid chromatography electrospray ionisation mass spectrometry. J Chromatogr A 1218(40):7094–7104. https://doi.org/10.1016/j.chroma.2011.07.057
73. CompTox Chemistry Dashboard: DTXSID3031864. https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXSID3031864. Accessed 1 Mar 2018
74. CompTox Chemistry Dashboard: DTXSID8031865. https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXSID8031865. Accessed 1 Mar 2018
75. CompTox Chemistry Dashboard: DTXSID8037706. https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXSID8037706. Accessed 1 Mar 2018
76. CompTox Chemistry Dashboard: MS-Ready Mappings of Perfluorooctanesulfonic acid. https://comptox.epa.gov/dashboard/dsstoxdb/ms_ready_mixture?cid=11864&gsid=31864&name=Perfluorooctanesulfonic%20acid. Accessed 1 Mar 2018
77. Sakurai N, Narise T, Sim J-S, Lee C-M, Ikeda C, Akimoto N et al (2017) UC2 search: using unique connectivity of uncharged compounds for metabolite annotation by database searching in mass spectrometry-based metabolomics. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx649
78. CompTox Chemistry Dashboard: MS-Ready Search of C8HF17O3S. https://comptox.epa.gov/dashboard/dsstoxdb/multiple_results?utf8=%E2%9C%93&search_inputs=C8HF17O3S&search_type=ms_ready_formula. Accessed 10 Apr 2018