Demonstrating the Use of Non-targeted Analysis for Identification of Unknown Chemicals in Rapid Response Scenarios

Supporting Information

John T. Sloop [a, b], Alex Chao [a], Jennifer Gundersen [c], Allison L. Phillips [d], Jon R. Sobus [a], Elin M. Ulrich [a], Antony J. Williams [a], Seth R. Newton [a] *

[a] U.S. Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure, Research Triangle Park, North Carolina 27709, United States

[b] Oak Ridge Institute for Science and Education (ORISE) Participant, Research Triangle Park, North Carolina 27709, United States

[c] U.S. Environmental Protection Agency, Office of Research and Development, Center for Environmental Measurement and Modeling, Narragansett, Rhode Island 02882, United States

[d] U.S. Environmental Protection Agency, Office of Research and Development, Center for Public Health and Environmental Assessment, Research Triangle Park, North Carolina 27709, United States

*Corresponding author; Email address: Newton.Seth@epa.gov

Number of pages: 35; Figures: Figure S1-S4; Tables: Table S1, S4-S8, and S15

1. Materials

Ethanol (ACS spectrophotometric grade) was purchased from Sigma Aldrich (St. Louis, MO, USA). Formic acid was purchased from Fisher Scientific (Hampton, NH, USA). Acetonitrile and methanol (B&J Brand High Purity Solvent) were purchased from Honeywell Burdick & Jackson (Muskegon, MI, USA). Isotopically labeled standards of d_3-thiamethoxam (DTXSID60746816) and d_4-pyriproxyfen (DTXSID20894089) were purchased from CDN Isotopes, Inc. (Pointe-Claire, Quebec, Canada). Isotopically labeled standards of $^{13}C_4$-perfluorooctanoic acid (MPFOA, DTXSID70892999) and $^{13}C_4$-perfluoroontanesulfonate (MPFOS, DTXSID80894101) were purchased from Wellington Laboratories, Inc. (Guelph, Ontario, Canada). Isotopically labeled standards of $^{13}C_3$-atrazine (DTXSID60894088), $^{13}C_6$-methyl paraben (DTXSID30894090), and $^{13}C_4$, $^{15}N_2$-fipronil sulfone (DTXSID10894093) were purchased from Cambridge Isotope Laboratories, Inc. (Tewksbury, MA, USA).

A standard of finasteride (DTXSID3020625) and malathion (DTXSID4020791) were purchased from Sigma Aldrich (St. Louis, MO, USA). A standard of α-hydroxy alprazolam (DTXSID60190613) was purchased from Cayman Chemical Company (Ann Arbor, MI, USA). A commercially available aqueous film forming foam (AFFF) mixture, Solberg Type 6, was selected from the Department of Defense Qualified Products List of aqueous film forming foams and purchased commercially by the National Institute of Environmental Health Science (NIEHS), and a subaliquot was obtained.[1] The carpet squares used in mock scenario 3 were free swatches obtained from a local home improvement store. Ultrapure deionized (DI) water was generated in-house from a Barnsted Easypure Ultraviolet and Ultrafilter (UV/UF) system (Dubuque, IA, USA), coupled with activated charcoal and ion exchange resin canisters.

2. Quality Control

Mass calibration of the instrument was performed according to vendor recommendations prior to analysis in each polarity for each of the mock scenarios. Any drift in the mass accuracy of the time-of-flight (TOF) was continuously corrected for by infusion of two reference compounds, purine (DTXSID5074470, monoisotopic mass = 120.0436) and Hexakis(1H,1H,3H-perfluoropropoxy)phosphazene (DTXSID90880494, monoisotopic mass = 921.0025), via dual-electrospray ionization (ESI) sprayer. Mass accuracy of tracer compounds spiked into each sample and blank were monitored during each mock scenario, and if at any point mass accuracy was > 5 ppm for more than one of the tracer compounds per ionization mode, analysis was paused, and a thorough cleaning of the instrument would take place. Tracer performance is shown in Table S6. Matrix blanks were prepared by performing the appropriate sample preparation (dilution or extraction) on an un-spiked sample of the same matrix as the spiked sample(s) for each mock scenario. Matrix blanks were run for each mock scenario in order to perform blank subtraction of sample spectra and instrumental response. Spectrum blank subtraction was performed by subtracting the blank spectrum from the sample spectrum via Agilent's Qualitative Analysis and by subtracting the blank measured signal from the sample measured signal for a feature of interest using data output from the WebApp.

3. Sample Selection and Preparation

The work presented here attempted to mimic situations where targeted rapid response methods had failed to identify a chemical, and thus NTA would be employed as a logical next step. Selected Analytical Methods for Environmental Remediation and Recovery (SAM) are made

available to labs that aid in rapid response with the EPA.[2] Targeted methods for analysis of specific chemicals can be searched via the SAM Chemical Methods Query.[3, 4] The specific chemicals that the mock scenarios were intended to mimic, the surrogates used in place of those, and the chemicals assigned a structure in mock scenario 3 were searched against the list of chemicals/methods in SAM. While other methods may exist and be at the disposal of rapid response laboratories, one of the primary documents for chemicals routinely targeted did not contain any of the chemicals identified in any of the mock scenarios presented here, thus meeting our aims.

3.1. Mock Scenario 1

The first mock scenario involved identification of a surrogate of a chemical warfare agent (CWA) that was spiked into an alcoholic beverage intended to poison an individual. This scenario was chosen because of the attacks in recent years against multiple foreign operatives, specifically the one against former Russian agent Sergei Skripal and his daughter in the United Kingdom in 2018.[5] The chemical chosen for this scenario was malathion ($C_{10}H_{19}O_6PS_2$, DTXSID4020791), intended to be used as a surrogate for Novichok nerve agents, specifically Novichok A-234 ($C_8H_{18}FN_2O_2P$, DTXSID60896946), the chemical suspected to be used in the 2018 poisoning.[24] Both Novichok nerve agents and malathion are organophosphate acetylcholinesterase inhibitors, and both malathion and the chemicals in the class of Novichok nerve agents have similar structures, specifically the phosphate functional groups. Malathion is also commonly used and referred to in the literature as an acceptable surrogate for nerve agents.[6, 7]

To mimic an attack, the chemical surrogate malathion was spiked into pure ethanol at 20 µg/mL concentration by Analyst 1. Un-spiked ethanol was used as the matrix blank. While pure ethanol solvent is not an ideal matrix to mimic an alcoholic beverage, the scenarios were meant to become progressively more complex, so the sample matrix for this scenario was kept relatively simple. Because the matrix of this sample was amenable for LC-MS analysis, no additional sample pre-treatment was required by Analyst 2 prior to preparing the set of serial dilutions (diluted with acetonitrile) for both the sample and matrix blank.

3.2. Mock Scenario 2

The second mock scenario involved identification of a surrogate of alprazolam ($C_{17}H_{13}ClN_4$, DTXSID4022577) and fentanyl ($C_{22}H_{28}N_2O$, DTXSID9023049) from a surface wipe sample and a carpet sample. This scenario was intended to mimic a situation in which a clandestine drug laboratory (any location where illicit drugs are being illegally manufactured or processed, like an individual's home) was discovered. In this mock scenario, an illicit drug (alprazolam, common brand name "Xanax") was being "cut" (i.e., diluted with a cheaper, more powerful drug to increase potency and stretch the supply) with fentanyl or a fentanyl analog, and on-scene investigators were tasked with determining the identity of the drugs via surface wipe and non-traditional sampling (i.e., any sampling of porous materials) done on-site.

Finding an appropriate surrogate of fentanyl was difficult, due to so many of the "most similar" chemicals being highly regulated by the U.S. Drug Enforcement Administration (U.S. DEA). To select a surrogate for fentanyl, chemical lists of cannabinoids from the CompTox Chemicals Dashboard were filtered based on commercial availability and not being present in the

list of chemicals in the NIST14 database. This was done to find a chemical that would be available for purchase and less likely able to be tentatively identified by current methods used by rapid responders. From the over 8,000 cannabinoids contained in lists on the Dashboard, seven passed the filtering criteria, and those seven were ranked via Tanimoto similarity score. The surrogate for fentanyl, finasteride ($C_{23}H_{36}N_2O_2$, DTXSID3020625), was selected using this approach. Another commercially available standard, α-hydroxy alprazolam ($C_{17}H_{13}ClN_4O$, DTXSID60190613), was selected as the surrogate for alprazolam.

Samples were prepared by Analyst 1 in two different media, the first being a wipe of a non-porous countertop in the lab, to mimic a non-porous surface in a home such as a kitchen counter, and the second being a 3" × 3" square of carpet. Both sample media were spiked with 0.5 mL of 300 µg/mL α-hydroxy alprazolam solution and 0.5 mL of 100 µg/mL finasteride solution. The surface wipe was performed according to Willison et al., by spiking the surface of the countertop with the finasteride and α-hydroxy alprazolam solution, waiting 45-60 minutes for the surface to dry, applying 3-4 mL of methanol to a wipe cloth, and then wiping the surface being sampled with firm pressure, using vertical and horizontal S-strokes.[8] The surface wipe was prepared as a "real" sample, intentionally conducted over an area in the lab that was covered in dust, as to be sure that some background was introduced. The carpet sample was prepared by spiking the carpet directly with the finasteride and α-hydroxy alprazolam solutions. Both the surface wipe and carpet sample were extracted by Analyst 2. The surface wipe sample was extracted by adding 30 mL methanol to a 50-mL centrifuge tube containing the sample, vortexing for 1 minute, and then sonicating in an ultrasonic bath for 30 minutes. The carpet sample was extracted by placing the carpet square in a beaker, adding 30 mL methanol, and sonicating in an ultrasonic bath for 30 minutes. After sonication, the sample extracts were filtered using a 3-mL plastic syringe (BD, Franklin Lakes,

NJ, USA) and 25-mm syringe filter (VWR, Radnor, PA, USA) with a 0.2 µm polypropylene membrane. Matrix blanks of both media were prepared according to the steps described above on an un-spiked dusty surface of the lab in a different location than the spiked surface wipe sample was collected and an un-spiked carpet square that was a different piece of carpet (different brand and color, same style) than the spiked carpet sample. In real-world situations where a matrix blank is not easily obtained from the location of the release, current practices by rapid responders are to use a surrogate (i.e., as similar of a sample matrix as possible from a source known to be uncontaminated). For example, in situations similar to this scenario, a carpet swatch could be obtained from a home improvement store for use as a matrix blank, even if it is not the exact brand/make as the contaminated carpet sample. The set of serial dilutions of samples and matrix blanks were prepared by diluting with acetonitrile.

3.3. Mock Scenario 3

The third mock scenario involved identification of various components of an industrial spill in surface water. Aqueous film forming foam (AFFF) mixtures are considered class B synthetic foams, designed for class B fires (i.e., those involving flammable liquids).[9] While AFFF is highly effective at fighting high-hazard flammable liquid fires, such as gasoline, oil, and jet fuel, these mixtures are typically created by combining foaming agents with fluorinated and non-halogenated surfactants. Notably, per- and polyfluoroalkyl substances (PFAS) are the most well-known components of AFFF, although they only comprise a small volume percentage of most AFFF mixtures (5-10%).[10, 11] This scenario was intended to mimic a situation in which an AFFF mixture was either intentionally or unintentionally spilled and penetrated a body of water. For this

scenario, a commercially available AFFF mixture was used as the industrial mix, and a sample of surface water from a nearby lake was used as the sample matrix.

The sample was created by Analyst 1 by diluting the AFFF mixture 100-fold in surface water, and a surface water sample taken from the same body of water but at a different location than where the sample matrix was collected (approx. 250 meters apart) was treated as the matrix blank. Because the matrix of this sample was amenable for LC-MS analysis, no additional sample pre-treatment was required by Analyst 2 prior to preparing the set of serial dilutions for both the sample and matrix blank.

4. Instrumental Analysis

Liquid Chromatography (LC) – Quadrupole/Time-of-Flight (QToF) High Resolution Mass Spectrometry (HRMS) analysis was carried out using an Agilent 1290 Infinity high pressure liquid chromatography (HPLC) instrument (Agilent Technologies, Palo Alto, CA), interfaced with an Agilent 6530B QToF HRMS. Chromatographic separation was accomplished using an Eclipse Plus C8 column (2.1 × 50 mm, 3.5 μm; Agilent Technologies, Palo Alto, CA). The QToF was fitted with a dual-injection electrospray ionization (ESI) source, which operated in both negative (ESI-) and positive (ESI+) polarity (with a separate injection for each).

Three LC-MS methods were used during this study. The first was a 9-minute, LC-MS "rapid range finding" method, intended to perform quick chromatography for determination of appropriate sample concentration and ionization polarity (ESI+ and/or ESI-). The most dilute samples from the prepared serial dilutions were run first, along with a blank, in each ionization mode. Sample concentration was increased incrementally until an obvious difference between the

sample and blank was visually observed on the chromatograms, without being too concentrated as to saturate the detector. This sample dilution was then chosen as the preferred concentration for subsequent LC-MS analysis. In mock scenarios 1 and 2, only one ionization mode (ESI+) was used in the subsequent runs, based on the results of the rapid range finding method. The second method was a longer, 30-minute LC-MS method, intended to achieve greater chromatographic separation for the selected sample dilution in the chosen ionization mode(s). The third method was a 30-minute LC-MS/MS method, operating under the same LC conditions as the 30-minute LC-MS method, using data dependent acquisition (DDA) with the ion(s) of interest added to the preferred ion list (fragmenting at collision energies of 10, 20, and 40 eV). This was performed to collect MS/MS fragmentation data for selected ions that were found during the rapid range finding method and deemed potentially important for the study, as well as additional ions that the instrument selected for fragmentation based on abundance.

Data was collected in 2 GHz high resolution mode, collecting ions in *m/z* range 100–1000 in both centroid and profile data formats for MS analysis, and 100-1700 *m/z* range for the MS/MS analysis. Specifics on common instrumental parameters for all three methods can be found in Table S7, and details on the LC gradients used in each of the three runs can be found in Table S8.

5. Data processing

A detailed explanation of each of the five data processing approaches used in this work are described below, in sections 5.1-5.5. Specific parameters used for the various tools described below are given in Tables S9-S14 (provided in a separate Excel file).

197 5.1. Formula matching to MS-Ready Formula using MS^1 data

198 Molecular feature extraction and chemical formula assignment were performed according to
199 previously published methods using Agilent Profinder v8.0 and Agilent Mass Profiler Professional
200 (MPP) v15.0, respectively.[12] Molecular features (defined by an exact mass [*m/z*] at a retention time
201 [RT], associated ions, and intensity of an apparent unknown compound) were aligned and
202 extracted using the Batch Recursive Feature Extraction Wizard in Profinder. Extracted and aligned
203 features were saved in .CEF files which were then imported into MPP using the data import wizard,
204 with no additional alignment performed. Specific parameters used for Profinder can be found in
205 Table S9.

206 Chemical formulae were assigned to molecular features via the Compound Identification
207 Wizard in MPP. This tool uses "MS-Ready" formulae for ~ 760 K substances contained within
208 EPA's Distributed Structure-Searchable Toxicity (DSSTox) Database.[13] These formulae are
209 available for download online at the CompTox Chemicals Dashboard page
210 (About/Downloads/DSSTox MS Ready Mapping File).[14] Procedures for generating MS-Ready
211 formulae were described in McEachran et al., and involve desalting, desolvation, removal of
212 stereochemistry, and neutralization.[15] Matching molecular features to MS-Ready formulae was
213 based on isotope presence, abundance, and spacing. Specific parameters used for MPP can be
214 found in Table S10. For each molecular feature, MPP assigned and output a maximum of one MS-
215 Ready formula with the highest match score (maximum score = 100) from all potential candidate
216 formulae, which was then deemed the top formula match from MS-Ready formula matching.

217

218 5.2. Manual Molecular formula prediction on MS^1 data

Another method used for determining a formula for an unknown chemical was the Molecular Formula Generator tool on Agilent MassHunter Qualitative Analysis 10.0. This tool utilizes the *m/z* of the selected peak, as well as the abundance and spacing of any neighboring isotopologue peaks, minimum and maximum numbers of elements to consider, maximum neutral mass allowed, minimum score per charge carrier, maximum hits per charge carrier, and minimum peak heights (both absolute and relative) when generating the list of likely predicted molecular formulae. A total match score (maximum score = 100) is assigned to each predicted formula, and the formula with the highest match score was deemed the top formula match from molecular formula prediction. Formula predictions were made after background subtracting ~0.1 min before and after the peak of interest, which has been observed to give better formula prediction than raw data. The user must specify elements to be included, as well as minimum and maximum numbers for each. The elements included and their ranges were C (3-60), H (0-240), O (0-45), N (0-50), S (0-25), Cl (0-20), P (0-25), F (0-40), and Br (0-10). Specific parameters used can be found in Table S11.

5.3. NTA WebApp Search by Mass using MS^1 data

For the past several years, the U.S. EPA has been developing tools to aid NTA studies in chemical identification. One of these tools is the online NTA WebApp (referred to from here on as "the WebApp"). The MS1 tool of the WebApp automates the process of generating feature candidates by performing batch searches against the contents of the DSSTox database for every feature included in the results of the Compound Identification Wizard in MPP.[13, 16] Parameters can be set on the WebApp that determine the thresholds for filtering and exclusion of individual features. In this study, the WebApp was told to search for features by mass, with a mass accuracy

window of ±5 ppm. Specific parameters used can be found in Table S12. Because the WebApp was told to search by mass, every feature that was included in the MPP output file, regardless of whether a formula was assigned or not, was searched against the contents of the DSSTox database. The WebApp's MS1 tool generates a .CSV file of all features matched to unique chemicals within the DSSTox Database, and the chemical candidates listed for each feature are ordered by number of data source hits. Because it has been shown in NTA studies that the candidate within each feature's candidate list with the greatest number of data source hits is the correct identification ~80% of the time, that candidate was deemed the top chemical match for that feature from the WebApp search by mass.[12, 17]

5.4. Matching MS^2 data to spectral libraries

Seven personal compound database and library (PCDL) files were used to match MS/MS spectra to experimental MS/MS spectra stored in a database or library: Metlin, ForTox, Pesticides, Water, Sulfas, VetDrugs, and Extractables and Leachables. Agilent MassHunter Qualitative Analysis 10.0 was used to perform these matches. Molecular features were first extracted from the MS/MS data files using the molecular feature extractor, and then compounds were identified by matching to PCDLs purchased from Agilent Technologies. Specific parameters used can be found in Table S13. A compound match (maximum forward score = 100) for a given feature was deemed as a potential candidate from matching MS/MS data to spectral libraries, and spectral matches were visually inspected to determine the best match.

5.5. NTA WebApp for Matching MS^2 data to CFM-ID in-silico database

The WebApp also features an MS2 matching tool, separate from the MS1 data processing tool.[16] The MS2 tool matches experimental MS/MS data to a database of pre-predicted MS/MS spectra that was built by applying the Competitive Fragmentation Modeling for Metabolite Identification (CFM-ID) 2.0 algorithms to DSSTox compounds.[18] It has been shown that by using experimental and in silico libraries together, 73% of 377 unique compounds from EPA's Non-Targeted Analysis Collaborative Trial (ENTACT) were correctly identified.[19] In this workflow, an .MGF file of the collected MS/MS data was exported using Agilent MassHunter Qualitative Analysis. That file was then uploaded to the MS2 tool on the WebApp, with parameters for precursor and fragment mass accuracy set (the specific parameters used can be found in Table S14). The WebApp's MS2 tool generated a .CSV file containing every candidate match, with match scores (maximum score = 1) assigned by fragmentation energy. Matches were considered as possible candidates, and the candidate with the highest total match score (summed across scores from all fragmentation energies) was initially deemed the best candidate from matching MS/MS data to the CFM-ID *in silico* database.

5.6. General data processing guidelines

The results from all five data processing approaches were considered when assigning a chemical identification to any given feature in each of the mock scenarios. (It should also be reiterated that before this stage of data processing began, the initial features of interest were already prioritized and selected for further analysis by the work done during the rapid range finding method.) As a general rule, priority was given to the results from the WebApp's MS1 tool, with the remaining approaches serving to further corroborate these results. Furthermore, situational information was considered when assigning a tentative chemical identity. The conclusion of

chemical identity for a feature required the analyst's judgement when weighing the evidence from all five workflows. Chemical identifications were assigned a level of confidence based on the Identification Confidence scale by Schymanski et al., ranked from levels 1-5.[20]

6. Mock scenario 2: Additional results

For the first feature investigated ($C_{17}H_{13}ClN_4O$ at 324.0783 Da), using the MFG tool (SI 5.2) gave multiple possible formula matches, with $C_{17}H_{13}ClN_4O$ being scored second highest at 72.46 (the highest scored formula match was $C_{19}H_{15}ClNO_2$ at 83.22). Mass search results of the WebApp MS1 tool (SI 5.3) showed 10 candidates with at least 5 data source hits, three of which had the same molecular formula as the one matched via MPP. The highest scoring match based on number of data source hits was α-hydroxy alprazolam (n=29), with the next two hits having n=17 and n=12 data source hits, but a different molecular formula ($C_{11}H_{14}F_6O_4$ and $C_{14}H_{16}N_2O_5S$, respectively). While there was no hit for the MS/MS results based on PCDL matching (SI 5.4), there were multiple candidates scored via CFM-ID predicted spectra (SI 5.5). From the original list of candidates from the WebApp MS1 results, α-hydroxy alprazolam scored second highest via CFM-ID at 0.8347 (the highest scored match via CFM-ID predicted spectra was anti-Benzo(a)chrysene-11,12-diol-13,14-epoxide, with formula $C_{22}H_{12}O_3$, n=6 data source hits from the WebApp MS1 tool, and CFM-ID score 0.8414). Even though it ranked 2nd based on MFG and CFM-ID scoring, considering all the evidence gathered (MPP match, top ranked by data source hits, and highly ranked by MFG and CFM-ID), Analyst 2 correctly reported that the chemical identification was α-hydroxy alprazolam, at a Level 2B.

The second feature investigated was the feature with MPP formula match $C_{23}H_{36}N_2O_2$ (at 372.2718 Da). Using the MFG tool (SI 5.2) gave multiple possible formula matches, with $C_{23}H_{36}N_2O_2$ being scored second highest at 87.90 (the highest scored formula match was $C_{26}H_{38}$ at 90.65). Results of the mass search using the WebApp MS1 tool (SI 5.3) showed 7 candidates with at least 5 data source hits, and two of these had the same molecular formula matched via MPP (the top hit, finasteride, with n=123, and the 6th hit, 2-pentadecyl-3H-benzimidazole-5-carboxylic acid, with n=5). The highest scoring match based on number of data source hits was finasteride (n=123), with the next two hits having much fewer number of data source hits (n=39 and n=20). While none of the spectral matches from the WebApp MS2 tool (SI 5.5) were on the list of candidates from the WebApp MS1 results, there was a PCDL match (SI 5.4) for finasteride scored at 76.80, shown in Figure S2. Based on the evidence gathered from all five data processing approaches, Analyst 2 correctly reported that the chemical identification was finasteride, at Level 2A.

The third feature investigated was the feature with MPP formula match $C_{11}H_{15}NO_2$ (at 193.1110 Da). The MFG tool (SI 5.2) gave multiple possible formula matches, with $C_{11}H_{15}NO_2$ scoring second highest at 86.31 (the highest scored formula match was $C_9H_{13}N_4O$ at 87.12). Results of the mass search using the WebApp MS1 tool (SI 5.3) showed many candidates (>10) with at least 20 data source hits, and the majority having the same molecular formula matched via MPP. The highest scoring match based on number of data source hits was isoprocarb (n=106), with the next two having a similar number of data source hits (butyl 4-aminobenzoate with n=95 and parbenate with n=82), and all three had the same formula that was matched via MPP ($C_{11}H_{15}NO_2$). There was both PCDL (SI 5.4) and CFM-ID (SI 5.5) predicted spectra matches based on MS/MS data for feature candidates, with parbenate being the best PCDL match and second

highest scoring CFM-ID match at 0.6718 (out of a maximum possible score of 1). Based on the evidence gathered from all five data processing approaches, Analyst 2 reported that the chemical identification was parbenate, at Level 2A.

7. Mock scenario 3: Additional results

It was determined that both the 50-fold and 10-fold matrix blank/sample dilutions would need to be analyzed via the longer MS instrumental method. To capture good data for both (i) features that were present in the 10-fold diluted sample but not in the 50-fold diluted samples, and (ii) features that were saturated in the 10-fold diluted sample, it was necessary to analyze both the 10-fold and 50-fold diluted samples via the general MS method and the DDA MS/MS method. A minimum threshold of sample:matrix blank ratio was then set at 10:1, and the top 3 from ESI+ and top 5 from ESI- were then chosen for further inspection. A secondary set of features of interest were also chosen, to intentionally select features that may be halogenated. Because of the recent increased interest in identifying halogenated compounds (such as PFAS compounds and other fluorinated chemicals), features that had a negative mass defect (which is common with halogenated chemicals) were also specifically sought out when analyzing the results. To not miss any potentially important halogenated compounds, features that satisfied the sample:matrix blank ratio requirement and had a negative mass defect (i.e., *m/z* = XXX.7, XXX.8, or XXX.9) were identified. The top 3 from ESI+ and top 6 from ESI- were then added to the list for further inspection, for a total of 17 features of interest for further investigation.

From ESI+ results, there were two features assigned at a Level 4. There were no features investigated from ESI- results that resulted in a Level 4 assignment. The chemical formulae

assigned were $C_5H_5Cl_2N_3S$ and $C_4H_3ClN_2O_3$. From both ESI+ and ESI- results, there were a total of six features of interest that remained at a Level 5. Two of these features were observed via ESI+, and four of these features were observed via ESI-. The masses of interest observed in ESI+ were *m/z* 100.9915 and 185.1159, and the masses of interest observed in ESI- were *m/z* 134.9874, 256.9545, 306.9832, and 334.9557. For these assignments, there was either only enough supporting evidence to allow for a formula assignment, or not enough evidence to assign anything other than the *m/z* seen in the data.

Of note, there were an additional 3 features found in the ESI- results that were incorrectly assigned as unique features during the Batch Recursive Feature Extraction Wizard on Agilent Profinder. Upon further manual inspection of the chromatogram and MS spectra of these 3 features, it was determined that 2 of them were isotopologues of other features and therefore were not unique features. After extracting and inspecting the MS spectrum from the regions in the chromatogram to the left and right of the third feature, it was determined that it was a spike in the background that was also incorrectly identified as its own unique feature. Of the 17 features originally selected for further analysis, after removing features that were not "real", 14 features of interest remained.

8. "Known unknowns" vs. "Unknown unknowns"

It is necessary to consider that in each of the three mock scenarios presented in this work, not all chemicals of interest were truly "unknown" chemicals (i.e., never-before discovered or documented), but instead were known chemicals whose structures mimicked those unable to be identified prior to performing NTA. The chemicals used in this study were also spiked at relatively

high concentrations with the assumption that in real rapid response scenarios, samples can be collected near the original source, where the chemical(s) are present at much higher concentrations than the background matrix and therefore easy to identify during rapid range finding. Should this not be the case, future analysts should consider concentrating the sample extracts rather than diluting for range finding exercises. Considering the amount of chemical present and whether it is a known chemical or truly unidentified and undocumented prior to the analysis (i.e., a "known unknown" or "unknown unknown"), there are four situations in which an analyst could find themselves. These situations are detailed in Table S15. The difficulty of the resulting analysis, and therefore the time required for said analysis, increases as the concentration of chemical decreases and if it is an undocumented chemical versus one that is previously discovered.

9. Hazard comparison module discussion

9.1. Mock scenario 1

The hazard report for mock scenario 1 is shown in Figure S3. In this scenario, there are many human health effect concerns for malathion, and its predicted transformation product, ethanol, including oral, inhalation, and dermal acute mammalian toxicity, genotoxicity mutagenicity, single exposure neurotoxicity and systemic toxicity, skin sensitization and irritation, and eye irritation. Because the chemical in this scenario was spiked into a beverage, the main concern would be the individual who consumed the beverage, and any individuals nearby when the incident occurred, due to the inhalation toxicity concerns. Responders would address the individual who consumed the beverage, as well as likely close off the area the incident occurred from others until remediation efforts were finished. Even though the chemical and its predicted

transformation product have acute aquatic toxicity concerns, because this chemical was not released into a body of water, this piece of information is not relevant to this situation.

9.2. Mock scenario 2

The hazard report for mock scenario 2 is shown in Figure S4. In this scenario, there are some human health effect concerns for the three identified compounds (alpha-hydroxy alprazolam, finasteride, and parbenate) and their predicted transformation products, including oral acute mammalian toxicity and genotoxicity mutagenicity. Specifically, for one transformation product of finasteride (tert-butylamine) and one transformation product of parbenate (ethanol), there are other concerns, including inhalation and dermal acute mammalian toxicity, single exposure systemic toxicity, skin sensitization and irritation, and eye irritation. It can be assumed that the location of this scenario would already be closed to the public (since it was a raid on a "drug house" and now likely a crime scene), but individuals investigating and conducting remediation efforts would need to wear the appropriate personal protective equipment (PPE) used when at the scene of clandestine fentanyl laboratories, taking care to not orally ingest, inhale, or come into contact via skin or eyes any of the surfaces potentially impacted at the location.

References

1. Ruyle, B. J.; Thackray, C. P.; McCord, J. P.; Strynar, M. J.; Mauge-Lewis, K. A.; Fenton, S. E.; Sunderland, E. M., Reconstructing the Composition of Per- and Polyfluoroalkyl Substances in Contemporary Aqueous Film-Forming Foams. *Environmental Science & Technology Letters* **2021,** *8* (1), 59-65.

2. Agency, U. S. E. P. Environmental Sampling and Analytical Methods (ESAM) Program. Selected Analytical Methods for Environmental Remediation and Recovery (SAM). https://www.epa.gov/esam/selected-analytical-methods-environmental-remediation-and-recovery-sam.

3. Agency, U. S. E. P., Environmental Sampling and Analytical Methods (ESAM) Program. SAM Chemical Methods Query. 2022.

4. Agency, U. S. E. P., Selected Analytical Methods for Environmental Remediation and Recovery (SAM) 2017. 2017; Vol. EPA/600/R-17/356.

5. Vale, J. A.; Marrs, T. C.; Maynard, R. L., Novichok: a murderous nerve agent attack in the UK. *Clinical Toxicology* **2018,** *56* (11), 1093-1097.

6. Noradoun, C. E.; Mekmaysy, C. S.; Hutcheson, R. M.; Cheng, I. F., Detoxification of malathion a chemical warfare agent analog using oxygen activation at room temperature and pressure. *Green Chemistry* **2005,** *7* (6), 426-430.

7. de A. Cavalcante, F. S.; Simas, B. C. A.; Kuča, K., Nerve Agents' Surrogates: Invaluable Tools for Development of Acetylcholinesterase Reactivators. *Current Organic Chemistry* **2019,** *23* (14), 1539-1559.

8. Willison, S. A.; Daniel Stout, I. I.; Mysz, A.; Starr, J.; Tabor, D.; Wyrzykowska-Ceradini, B.; Nardin, J.; Morris, E.; Snyder, E. G., The impact of wipe sampling variables on

437 method performance associated with indoor pesticide misuse and highly contaminated areas.
438 *Science of The Total Environment* **2019,** *655*, 539-546.
439 9. Magrabi, S. A.; Dlugogorski, B. Z.; Jameson, G. J., A comparative study of drainage
440 characteristics in AFFF and FFFP compressed-air fire-fighting foams. *Fire Safety Journal* **2002,**
441 *37* (1), 21-52.
442 10. Moody, C. A.; Field, J. A., Perfluorinated Surfactants and the Environmental Implications
443 of Their Use in Fire-Fighting Foams. *Environmental Science & Technology* **2000,** *34* (18), 3864-
444 3870.
445 11. Houtz, E. F.; Higgins, C. P.; Field, J. A.; Sedlak, D. L., Persistence of Perfluoroalkyl Acid
446 Precursors in AFFF-Impacted Groundwater and Soil. *Environmental Science & Technology* **2013,**
447 *47* (15), 8187-8195.
448 12. Sobus, J. R.; Grossman, J. N.; Chao, A.; Singh, R.; Williams, A. J.; Grulke, C. M.;
449 Richard, A. M.; Newton, S. R.; McEachran, A. D.; Ulrich, E. M., Using prepared mixtures of
450 ToxCast chemicals to evaluate non-targeted analysis (NTA) method performance. *Analytical and*
451 *Bioanalytical Chemistry* **2019,** *411* (4), 835-851.
452 13. Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker,
453 N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M., The CompTox
454 Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of*
455 *Cheminformatics* **2017,** *9* (1), 61.
456 14. Agency, U. S. E. P. CompTox Chemicals Dashboard: Downloads.
457 https://comptox.epa.gov/dashboard/downloads.

15. McEachran, A. D.; Mansouri, K.; Grulke, C.; Schymanski, E. L.; Ruttkies, C.; Williams, A. J., "MS-Ready" structures for non-targeted high-resolution mass spectrometry screening studies. *Journal of Cheminformatics* **2018,** *10* (1), 45.

16. Minucci, J. S., Deron; Purucker, Tom; Boyce, Matthew; Grulke, Christopher M. EPA NTA WebApp. https://github.com/quanted/nta_app.

17. McEachran, A. D.; Sobus, J. R.; Williams, A. J., Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. *Analytical and Bioanalytical Chemistry* **2017,** *409* (7), 1729-1735.

18. McEachran, A. D.; Balabin, I.; Cathey, T.; Transue, T. R.; Al-Ghoul, H.; Grulke, C.; Sobus, J. R.; Williams, A. J., Linking in silico MS/MS spectra with chemistry data to improve identification of unknowns. *Scientific Data* **2019,** *6* (1), 141.

19. Chao, A.; Al-Ghoul, H.; McEachran, A. D.; Balabin, I.; Transue, T.; Cathey, T.; Grossman, J. N.; Singh, R. R.; Ulrich, E. M.; Williams, A. J.; Sobus, J. R., In silico MS/MS spectra for identifying unknowns: a critical examination using CFM-ID algorithms and ENTACT mixture samples. *Analytical and Bioanalytical Chemistry* **2020,** *412* (6), 1303-1315.

20. Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J., Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environmental Science & Technology* **2014,** *48* (4), 2097-2098.

Table S1. Detailed description of the five steps in the data processing workflow. Each step is listed in column 1, with a description of each provided in column 2, and the results for the first mock scenario (as well as the thinking/rationale behind certain steps) provided in column 3. In the "Description" column, a general description is provided in italic font, with a more detailed description provided in plain text beneath.

Data Processing Approach	Description	Results from Mock Scenario 1
(1) Feature extraction and formula matching	*Molecular features were extracted via Agilent Profinder v8.0 and assigned chemical formulas from compounds within DSSTox via Agilent Mass Profiler Professional (MPP) v15.0.* If a feature was matched to at least one formula, the highest scoring formula match (max=100) was assigned to that feature. The formula assignment provided for each feature was considered the "best" formula assignment, based on mass, isotope presence, abundance, and spacing.	MPP formula assignment was $C_{10}H_{19}O_6PS_2$ (score=89.2)
(2) Molecular Formula Generator (MFG) tool	*A list of likely molecular formulae was predicted via Agilent MassHunter Qualitative Analysis 10.0.* The formula predictions for each feature were ranked based on total match score (max=100), which is based on the *m/z* of the selected peak, isotope presence, abundance, and spacing, a min/max number of elements to consider, and min peak height thresholds. The highest scoring predicted formula was considered the "best" formula prediction.	Top scoring MFG formula prediction was $C_{10}H_{19}O_6PS_2$ (score=99.11)
(3) WebApp MS1 tool	*Candidate lists for each feature output from MPP were generated by an automated search against the contents of DSSTox via the NTA WebApp's MS1 tool.* Features for further investigation were prioritized by sorting by sample intensity after blank subtraction. Candidates were searched by mass, so multiple candidates with potentially different molecular formulae	Total of 49 potential candidates for the feature of interest. Top 3 candidates by data source hits were malathion (n=250), isomalathion (n=33), and becampanel (n=17). Because the formula of both malathion and isomalathion

	were returned. Candidates were initially ranked by data source hits (i.e., the number of times a unique chemical appears in the various lists and libraries that compose the DSSTox database). However, consideration was also given to lesser ranked candidates if they had the same molecular formula as the matched and/or predicted formula hits from steps (1) and (2). Following this step, a top candidate from the MS1 approach was selected.	were the same as the formulas from steps (1) and (2) and becampanel was not, and malathion had a significantly greater number of data source hits than isomalathion, malathion was selected as the top candidate from the MS1 approach.
(4) Matching MS2 spectra to PCDL(s)	*Experimentally collected MS/MS spectra were matched against contents of seven different personal compound databases and libraries (PCDLs) via Agilent MassHunter Qualitative Analysis 10.0.* Molecular features were first extracted from the MS/MS files using the molecular feature extractor in Qualitative Analysis. Compounds were then identified by matching to PCDLs and provided a score (max=100). The highest scoring compound was considered the "best" PCDL match.	Matching experimentally collected MS/MS spectra to PCDL spectra yielded two compounds scored very low (25.48 and 27.32). Malathion was returned as a match from the PCDL approach, but the PCDLs did not contain a malathion mass spectrum, so it did not receive a match score.
(5) Matching MS2 spectra to CFM-ID database	*Experimentally collected MS/MS spectra were matched to a database of pre-predicted MS/MS spectra built by applying the CFM-ID 2.0 algorithm to chemicals within DSSTox via the NTA WebApp's MS2 tool.* An .MGF file of the experimentally collected MS/MS data was exported using Agilent MassHunter Qualitative Analysis, and then uploaded to the WebApp's MS2 tool. This generates a .CSV file containing every candidate match for every feature, with match scores (max=1) assigned by fragmentation energy. A total match score was generated by summing across the scores for all fragmentation energies (max=3). All matches were considered as possible candidates, and the candidate with the highest total match score was deemed the "best" candidate from CFM-ID matching.	Matching MS/MS spectra to a database of pre-predicted spectra yielded 55 potential matches. While scored low relative to the remaining matches, malathion was one of the potential matches. However, malathion had the greatest number of data source hits (n=250) when compared to any of the remaining highest scoring candidates on the list (n=28, n=16, etc.).
Final assignment of chemical identity	*Considering the results from all 5 data processing approaches, a final assignment is made on the chemical identity.*	Given that all MS1 approaches pointed towards the same molecular formula and the relatively large

	Typically, the overall assignment is performed by first only considering the MS1 results (database formula matching, formula prediction, and candidates ranked by data source from the WebApp's MS1 tool). Assuming that at least one of the formulas from steps (1) and (2) agree with the top candidate from the WebApp, that is typically considered the best match from the MS1 approach. Then, the MS2 data is used as a way to further increase the confidence in the best match from the MS1 results. Assuming that one of the MS2 approaches from steps (4) and (5) return a match for the top ranked candidate via MS1 data (and ideally, the MS2 match is also ranked high), then that candidate is viewed as the best match and reported as the final assigned chemical identity.	number of data source hits when compared to the next highest ranked candidates, malathion was considered the best match from the MS1 data. While there was no spectrum of malathion to match to in the PCDL, matching to pre-predicted spectra yielded a match with malathion (though scored low). Considering all 5 points of data, the analyst reported that malathion was the chemical identification at a Level 2b on the Schymanski et al. scale, based on parent compound information (match based on the matched formula, predicted formula, measured *m/z*, and data source hits) as well as diagnostic MS/MS fragmentation (observed via the CFM-ID match).

482

483

Table S4. Top 10 candidate features for mock scenario 2 via MPP output (SI 4.1). Experimental abundance (counts, sorted high to low after blank subtraction), measured accurate mass (Da), and RT (min) are shown. The three features investigated further (based on RT < 20 min) have their "Feature ID" listed in bold (either formula match or exact mass at RT for features not matched to a formula).

Feature ID	Abundance (counts)	Measured Accurate Mass (Da)	RT (min)
$C_{32}H_{27}N_5O_8$	4.80E+07	609.1752	21.379
$C_{17}H_{13}ClN_4O$	3.97E+07	324.0783	7.485
$C_{23}H_{36}N_2O_2$	2.00E+07	372.2718	9.058
928.2114@21.363	5.68E+06	928.2114	21.363
$C_{46}H_{38}F_{12}P$	3.95E+06	849.2383	20.993
$C_{11}H_{15}NO_2$	3.90E+06	193.111	10.192
$C_{21}H_{35}N_9O_{12}S$	2.87E+06	637.2035	21.362
833.2077@20.197	2.48E+06	833.2077	20.197
$C_{12}HBr_4Cl_2NOS$	1.76E+06	592.5844	20.089
$C_{42}H_{45}NO_{23}$	1.52E+06	931.2102	21.366
$CH_3NO_3S_2$	1.50E+06	140.9516	21.827

Table S5. The 14 features further inspected during mock scenario 3, sorted by final identification level. Details given for each feature include the polarity in which it was observed (ESI+/ESI-), measured accurate mass (Da), RT (min), and ultimate identification level. Note that Feature ID 5 and 6 correspond to the same chemical, observed in both ESI+ and ESI- polarity.

Feature ID	Polarity (ESI+/ESI-)	Measured accurate mass (Da)	RT (min)	Final Identification Level (DTXSID, for those with structure assignments)
1	ESI+	162.1256	6.255	Level 2 (DTXSID8021519)
2	ESI-	210.0924	8.447	Level 2 (DTXSID7042433)
3	ESI-	238.1244	10.234	Level 2 (DTXSID8042428)
4	ESI-	427.9752	9.802	Level 2 (DTXSID6067331)
5	ESI+	528.0757	8.917	Level 3 (DTXSID80880983 or DTXSID10868577)
6	ESI-	528.0750	8.915	Level 3 (DTXSID80880983 or DTXSID10868577)
7	ESI+	208.9575	8.449	Level 4
8	ESI+	323.9655	6.275	Level 4
9	ESI+	99.9837	6.325	Level 5
10	ESI+	184.1077	6.255	Level 5
11	ESI-	135.9952	6.256	Level 5
12	ESI-	257.9623	6.320	Level 5
13	ESI-	307.9910	8.447	Level 5
14	ESI-	335.9635	8.447	Level 5

Table S6. Tracer compound QC results for each mock scenario. Because mock scenario 2 was carried out by a different individual assuming the role of Analyst 2 than for the other mock scenarios, the specific mix of tracer compounds prepared was slightly different than for mock scenarios 1 and 3. In all mock scenarios, for all but one tracer in mock scenario 3, the average mass error (ppm) for each tracer compound was < 5 ppm.

Compound name	ESI Polarity	Avg. mass error (ppm)	Avg. RT (min)
Mock scenario 1			
$^{13}C_3$-atrazine	ESI+	2.75	8.4
d_4-pyriproxyfen	ESI+	0.00	13.3
Mock scenario 2			
d_3-thiamethoxam	ESI+	3.68	4.7
d_4-pyriproxyfen	ESI+	1.05	13.3
Mock scenario 3			
$^{13}C_3$-atrazine	ESI+	1.83	9.0
d_4-pyriproxyfen	ESI+	2.77	13.8
$^{13}C_4$-MPFOA	ESI-	0.72	10.4
$^{13}C_4$-MPFOS	ESI-	0.40	13.8
$^{13}C_6$-methyl paraben	ESI-	5.69	7.0
$^{13}C_4$, $^{15}N_2$-fipronil sulfone	ESI-	1.53	12.7

504 **Table S7.** Common LC-MS instrumental parameters used in all three LC-MS methods.

505

Instrumental parameter (units)	**Value**
Gas temperature (°C)	300
Gas flow rate (L/min)	7
Sheath gas temperature (°C)	350
Sheath gas flow rate (L/min)	11
Fragmentor voltage (V)	135
Injection volume (µL)	10.00
Binary pump flow rate (mL/min)	0.200

506

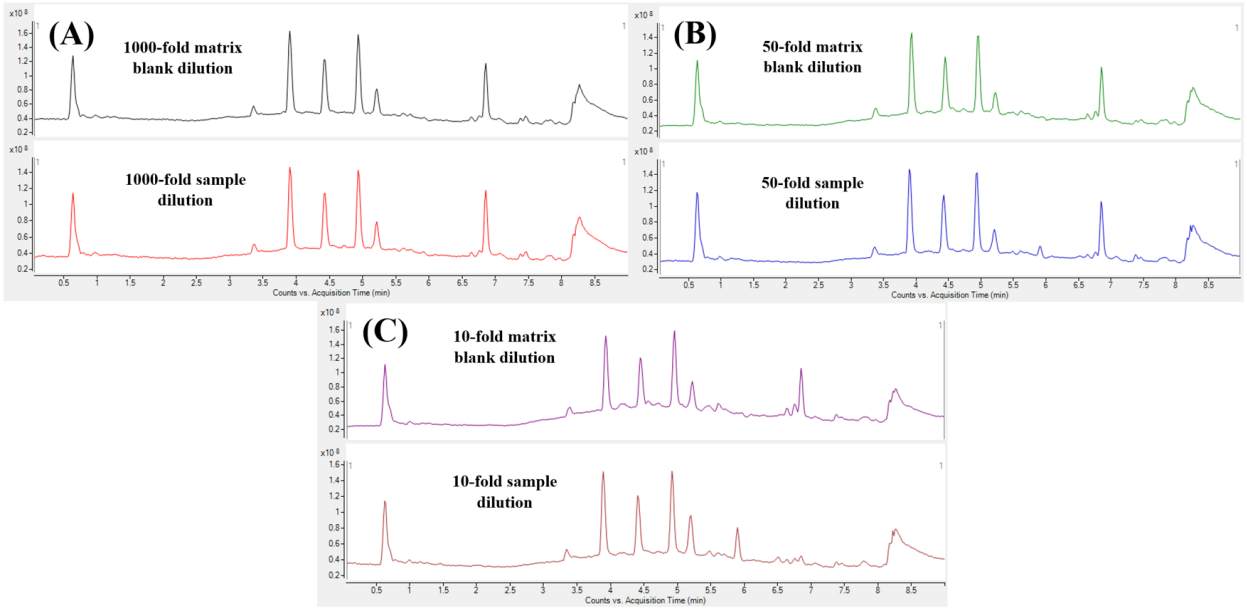
Table S8. LC mobile phase gradients used in each of the LC-MS instrumental methods. Solvent A was 0.1% (v/v) formic acid prepared in DI H_2O, and solvent B was 0.1% (v/v) formic acid prepared in acetonitrile.

Time (min)	Solvent A (%)	Solvent B (%)
Rapid range finding LC-MS method		
1.00	90.00	10.00
6.00	0.00	100.00
7.00	0.00	100.00
7.01	90.00	10.00
9.00	90.00	10.00
Longer, general LC-MS method and DDA LC-MS/MS method		
2.00	90.00	10.00
14.27	5.00	95.00
18.75	5.00	95.00
19.00	0.00	100.00
20.00	0.00	100.00
21.00	90.00	10.00
30.00	90.00	10.00

512 **Table S15.** The four situations encountered in any NTA study, based on the amount of the
513 "unknown" chemical of interest, and if the chemical is undocumented or not.

	Medium/High Concentration	**Trace Concentration**
Known Chemical	**Easy** - chemicals of interest can be identified using rapid range finding	**Easy** - if information about chemical(s) of interest are available (e.g., the masses of the compounds)
Undocumented Chemical	**Medium Difficulty** - focus can be placed on selected features; correct identification is not guaranteed	**Difficult** – situational information is needed; chances of identification are lower

514

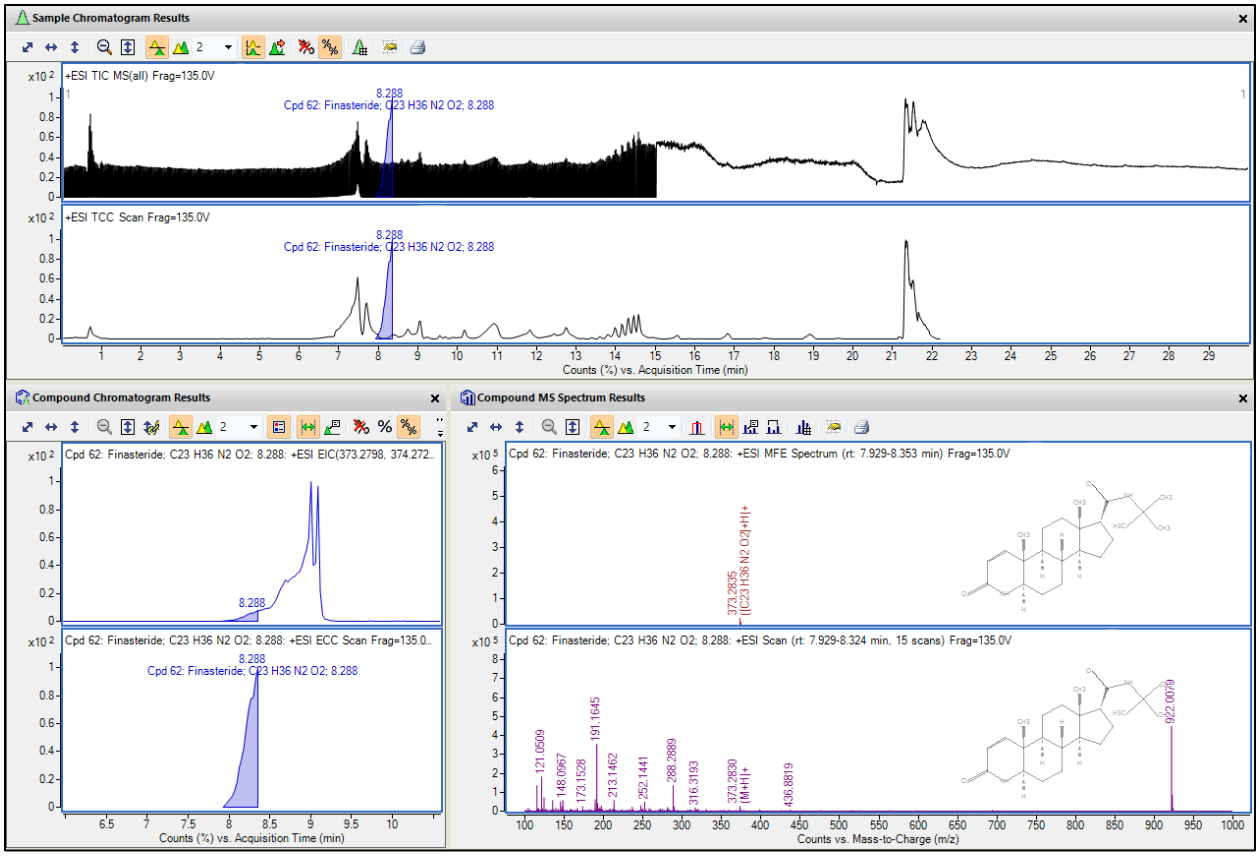


515

516 **Figure S1.** Chromatogram of the 1000-fold (A), 50-fold (B), and 10-fold (C) matrix blank (top)

517 and sample (bottom) dilutions for mock scenario 1.

518



519

520 **Figure S2.** Screenshot of MS/MS compound identification results from Agilent MassHunter
521 Qualitative Analysis for mock scenario 2. The matching for the compound finasteride is shown,
522 with a PCDL match score of 76.80.

523

Chemicals: 2

Toxicity: VH - Very High H - High M - Medium L - Low I - Inconclusive N/A - Not Applicable *Authority:* **Authoritative** Screening *QSAR Model*

Skipped (0)
Unlikely (2)
Filters (0)
Sorting (0)
Structure

CAS / Name	Human Health Effects									Ecotoxicity
	Acute Mammalian Toxicity			Genotoxicity Mutagenicit	Neurotoxicity	Systemic Toxicity				
	Oral	Inhalation	Dermal		Single Exposure	Single Exposure	Skin Sensitization	Skin Irritation	Eye Irritation	Acute Aquatic Toxicity
121-75-5 AIGBT Malathion	H	VH	L	H	H		H	M	H	VH
64-17-5 GBT Ethanol LIKELY	VH	VH	L	L		M	I	L	H	M

524

525 **Figure S3.** Hazard report generated via the Hazard Comparison Module for Mock Scenario 1. The
526 identified chemical, malathion, and 1 generation of breakdown products for it are shown, based on
527 the "emergency response" hazard assessment profile.

528

Chemicals: 9

Toxicity: VH - Very High H - High M - Medium L - Low I - Inconclusive N/A - Not Applicable *Authority:* **Authoritative** Screening *QSAR Model*

Skipped (0)
Unlikely (6)
Filters (0)
Sorting (0)
Structure

CAS / Name	Human Health Effects									Ecotoxicity
	Acute Mammalian Toxicity			Genotoxicity Mutagenicity	Neurotoxicity	Systemic Toxicity				
	Oral	Inhalation	Dermal		Single Exposure	Single Exposure	Skin Sensitization	Skin Irritation	Eye Irritation	Acute Aquatic Toxicity
37115-43-8 alpha-Hydroxyalp…	*M*			*L*						*VH*
No CAS No Name LIKELY	*M*			*L*						*H*
No CAS No Name LIKELY	*M*			*L*						*H*
98319-26-7 GBT Finasteride	M			*L*						M
75-64-9 GBT tert-Butylamine LIKELY	VH	H	L	L	I	I	I	VH	H	**M**
No CAS No Name LIKELY	*L*			*L*						*H*
10287-53-3 GBT Parbenate	*M*			*VH*						L
64-17-5 GBT Ethanol LIKELY	VH	VH	L	**L**		M	I	L	H	**M**
619-84-1 Benzoic acid, 4-(… LIKELY	*M*			*VH*			*H*			*M*

529

530 **Figure S4.** Hazard report generated via the Hazard Comparison Module for Mock Scenario 2. The
531 three chemicals identified (α-hydroxy alprazolam, finasteride, and parbenate) and 1 generation of
532 breakdown products for them are shown, based on the "emergency response" hazard assessment
533 profile.