



# Using prepared mixtures of ToxCast chemicals to evaluate non-targeted analysis (NTA) method performance

Jon R. Sobus<sup>1</sup> · Jarod N. Grossman<sup>2,3</sup> · Alex Chao<sup>2</sup> · Randolph Singh<sup>4</sup> · Antony J. Williams<sup>5</sup> · Christopher M. Grulke<sup>5</sup> · Ann M. Richard<sup>5</sup> · Seth R. Newton<sup>1</sup> · Andrew D. McEachran<sup>4</sup> · Elin M. Ulrich<sup>1</sup>

Received: 19 September 2018 / Revised: 14 November 2018 / Accepted: 27 November 2018 / Published online: 5 January 2019  
© This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2019

## Abstract

Non-targeted analysis (NTA) methods are increasingly used to discover contaminants of emerging concern (CECs), but the extent to which these methods can support exposure and health studies remains to be determined. EPA's Non-Targeted Analysis Collaborative Trial (ENTACT) was launched in 2016 to address this need. As part of ENTACT, 1269 unique substances from EPA's ToxCast library were combined to make ten synthetic mixtures, with each mixture containing between 95 and 365 substances. As a participant in the trial, we first performed blinded NTA on each mixture using liquid chromatography (LC) coupled with high-resolution mass spectrometry (HRMS). We then performed an unblinded evaluation to identify limitations of our NTA method. Overall, at least 60% of spiked substances could be observed using selected methods. Discounting spiked isomers, true positive rates from the blinded and unblinded analyses reached a maximum of 46% and 65%, respectively. An overall reproducibility rate of 75% was observed for substances spiked into more than one mixture and observed at least once. Considerable discordance in substance identification was observed when comparing a subset of our results derived from two separate reversed-phase chromatography methods. We conclude that a single NTA method, even when optimized, can likely characterize only a subset of ToxCast substances (and, by extension, other CECs). Rigorous quality control and self-evaluation practices should be required of labs generating NTA data to support exposure and health studies. Accurate and transparent communication of performance results will best enable meaningful interpretations and defensible use of NTA data.

**Keywords** ENTACT · Non-targeted analysis · ToxCast · Exposome

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s00216-018-1526-4>) contains supplementary material, which is available to authorized users.

✉ Jon R. Sobus  
sobus.jon@epa.gov

- <sup>1</sup> U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA
- <sup>2</sup> Student Contractor, U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA
- <sup>3</sup> Present address: Agilent Technologies Inc., Santa Clara, CA 95051, USA
- <sup>4</sup> Oak Ridge Institute for Science and Education (ORISE) Participant, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA
- <sup>5</sup> U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA

## Introduction

Since the 1970s, targeted analytical methods have been the primary means of generating monitoring data to support environmental health studies and chemical risk assessments. Monitoring programs around the globe now have the means to measure hundreds of target chemicals in biological specimens (e.g., blood and urine) and/or environmental samples (e.g., surface water and drinking water). Thousands of chemicals, however, are known or suspected to exist in consumer products, household articles, food and food packaging materials, surface and drinking water, household dust, and a host of other media [1–5]. Furthermore, signatures of chemical exposures have been detected in numerous biological media, including human serum, urine, breath, teeth, and breast milk [6–10]. The diversity and magnitude of chemical exposures are poorly characterized, as are the health effects that may result from these exposures. Growing awareness of these information gaps has led researchers to define and explore the

exposome, a concept that emphasizes the importance of all lifetime exposures, rather than exposure to individual target substances [11, 12]. This interest, in turn, has fostered an emphasis on non-targeted analysis (NTA) methods that can characterize many contaminants of emerging concern (CECs), in addition to target substances. To date, NTA methods have been employed to examine a wide variety of media, with most utilizing a combination of chromatography, high-resolution mass spectrometry (HRMS), and data processing workflows [13]. There remains, however, no standardized methods, quality control practices, or acceptance criteria for NTA studies, which may limit the utility of these data for research and decision-making.

Several research projects have examined NTA method performance by focusing on specific aspects of analytical workflows. For example, five Critical Assessment of Small Molecule Identification (CASMI) contests have examined the abilities of different tools to correctly identify unknown substances given experimental MS<sup>1</sup> and MS<sup>2</sup> data (<http://www.casmi-contest.org/2017/index.shtml>). CASMI contests have focused solely on computational aspects of NTA workflows by controlling variation that would otherwise stem from laboratory preparation and instrumental analysis. Other projects have allowed for these additional sources of variation when comparing NTA results across labs. For example, a recent article reported findings from a NTA collaborative trial in which 18 institutes from 12 European countries examined common extracts of a single surface water sample [14]. The trial was not meant to rank labs based on performance, but to better understand the impacts of method variation on substance identification. Results from these trials and contests have been tremendously useful for defining the state-of-the-science for NTA. An exhaustive examination of NTA method performance, however, requires the use of well-characterized samples containing a large and diverse assortment of substances. This need, as articulated during a U.S. EPA workshop in 2015, led to the design of EPA's Non-Targeted Analysis Collaborative Trial (ENTACT) [15].

ENTACT uses over 1200 chemical substances that are maintained by EPA's National Center for Computational Toxicology (NCCT) as part of the ToxCast project [16]. In 2016, a series of ten synthetic mixtures were prepared by combining these ToxCast substances [15]. The ten mixtures, each containing 95–365 substances, were distributed to approximately 25 labs with experience in performing NTA. Each lab was charged with performing a blinded analysis of each mixture, using an NTA method (or methods) of their choosing, followed by an unblinded review to assess method performance. The contents of each mixture were carefully selected by EPA staff. The purpose of the experimental design was to define the boundaries/limitations of each NTA method. Specific design elements, the exact contents of each mixture, and preliminary results across multiple laboratories are

discussed in detail in a companion article [17]. Future articles will carefully examine results of the trial across labs, paying particular attention to which NTA methods/techniques are well suited to individual substances and chemical classes. The purpose of this article is to report on the initial analyses conducted within our own laboratory. Specifically, we describe in detail the following: (1) our approach for analyzing individual mixtures using liquid chromatography (LC) quadrupole time-of-flight (QTOF) HRMS; (2) our methods for assigning candidate structures to observed molecular features; (3) our methods for comparing lists of tentatively assigned structures to lists of spiked substances; and (4) blinded and unblinded results of our experiments. Specific attention is given to factors that affected true positive rates, confidence levels in substance identifications, and method reproducibility across all ten mixtures. This article provides a performance evaluation model for labs involved in ENTACT. It further provides a general blueprint for any NTA lab wishing to conduct self-evaluations using synthetic mixtures. Finally, it provides a point-of-reference for determining the suitability of NTA methods to support exposure monitoring studies and related assessments of health risks.

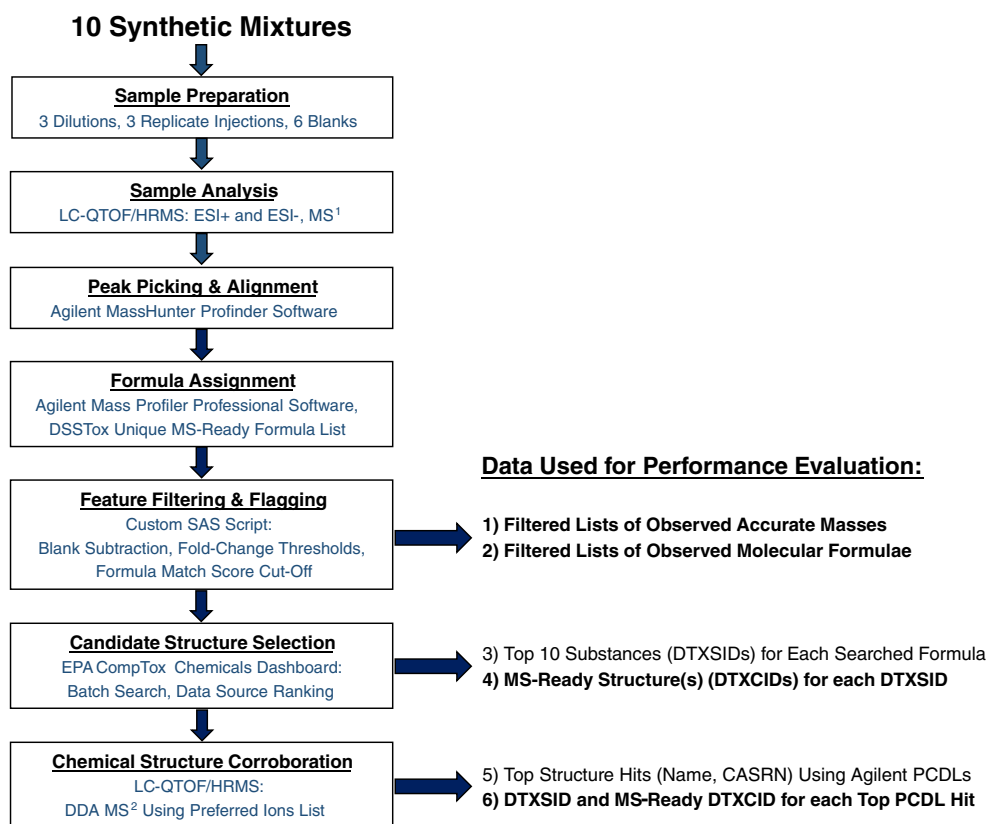
## Materials and methods

The overall NTA workflow is described in Fig. 1. This workflow includes information related to sample preparation and analysis, data processing, and the assignment of formulae and candidate structures to observed sample features.

### Sample preparation

Substances included in the ENTACT mixtures were from EPA's ToxCast library [16]. A total of 1269 unique substances were spiked across the ten ENTACT mixtures (see Electronic Supplementary Material (ESM) Table S1). Of these substances, five were spiked into all ten mixtures, 57 were spiked into four mixtures, 33 were spiked into three mixtures, 388 were spiked into two mixtures, and 786 were spiked into only one mixture. Five of the mixtures contained a total of 95 spiked substances, two contained 185 spiked substances, and three contained 365 spiked substances. One of the 95-substance mixtures (mixture 507) was designed to contain a large percentage of isomers (the same molecular formula). One of the 365-substance mixtures (mixture 508) was also designed to contain many isomers, as well as substances with lower molecular weights (MW) and lower quality control (QC) ratings (based on purity, stability, and expected vs. observed concentration). Substances included in all other mixtures had acceptable QC ratings based on previously collected experimental data [17].

**Fig. 1** NTA workflow as applied to ten synthetic mixtures from the ENTACT study. Major steps (left) are underlined and bold, and specific details of each step are written in blue font. The NTA workflow generated six data outputs (right), four of which (bold) were compared against spiked substance lists for the purpose of evaluating method performance



Mixtures were initially prepared in dimethyl sulfoxide with individual substances at a concentration of  $\sim 0.05$  mM. Multiple dilutions were then prepared (yielding working concentrations of approximately 0.5, 0.1, and 0.02  $\mu$ M) using Optima grade methanol (Fisher Scientific, Hampton, NH, USA) and 2 mM ammonium formate (Acros Organics, Geel, Belgium) buffer in HPLC grade water (B&J, Muskegon, MI, USA) at a ratio of 1:3 methanol:buffer. For this initial investigation, multiple dilutions were used as a means of filtering features emanating from sample preparation materials (see “[Feature filtering and flagging](#)” below). In future studies, data corresponding to specific dilutions may be further used to investigate the quantitative behavior (e.g., known concentration vs. observed intensity) of specific analytes of interest.

A 400- $\mu$ L aliquot of each diluted mixture was added into an autosampler vial. Each aliquot was then spiked (10  $\mu$ L) with a 1-ng/ $\mu$ L solution of stable isotope-labeled “tracer” compounds (Table 1) in methanol. Each dilution was analyzed in triplicate (15- $\mu$ L injections) using LC-QTOF/MS. Six method blanks (consisting of methanol, buffer, and tracer compounds) were interspersed throughout sample worklists to monitor instrument response and account for background/contaminant peaks. The use of relatively high analyte concentrations and matrix-free samples in this study is meant to provide a “best-case” scenario for evaluating NTA methods. By design, a lack of substance identifications cannot be attributed to matrix effects or insufficient instrument sensitivity.

### LC-QTOF/MS<sup>1</sup> analysis

Analyte separation was accomplished using an Agilent 1290 Infinity II LC system (Santa Clara, CA). Two different chromatographic methods were employed for this study, details of which are summarized in ESM Table S2. Method 1 used an Agilent ZORBAX Eclipse Plus C8 column (2.1  $\times$  50 mm, 1.8  $\mu$ m) and a mobile phase consisting of methanol, water, and ammonium formate buffer. Method 2 used a Waters Acquity UPLC<sup>®</sup> BEH C18 column (2.1  $\times$  50 mm, 1.7  $\mu$ m) and a mobile phase consisting of acetonitrile, water, and formic acid buffer. Unless otherwise noted, reported results are from method 1 only; this method was originally optimized for the detection of per- and polyfluoroalkyl substances (PFAS) in water, but over time proved to be a suitable generic method for a variety of compounds and media. No one method is assumed suitable for all analytes of interest; a key goal of ENTACT is to determine which methods are most suitable for specific chemicals and chemical classes.

An Agilent 6530B Accurate-Mass QTOF/MS (Santa Clara, CA) with a Dual AJS ionization source was operated under positive and negative electrospray ionization (ESI) in full scan mode (100–1000  $m/z$ ). The instrument was tuned using Agilent-specific tuning solution in both ionization modes before the start of any analysis. Reference solution was infused into the electrospray source to correct for mass drifts during a run. Purine (DTXSID5074470) and HP-0921

**Table 1** Descriptive statistics for stable isotope-labeled “tracer” compounds. Each tracer (source indicated by superscript) was observed in all replicate study samples ( $n = 90$ ) and blanks ( $n = 60$ ). Since feature alignment was performed separately for each ENTACT mixture, a mixture-specific monoisotopic mass (averaged across 9 samples and 6 blanks) was determined for each tracer. Minimum, median, and maximum mass accuracy

was then based on the 10 monoisotopic mass estimates for each tracer. Precision, as estimated using relative standard deviation (%-RSD) of peak area, was determined across replicate injections ( $n = 3$ ) at each mixture concentration ( $n = 3$ ). Minimum, median, and maximum precision was then based on 30 RSD estimates (10 samples  $\times$  3 concentrations) for each tracer

Chemical name	DTXSID <sup>d</sup>	Ionization mode	Monoisotopic mass (Da)	Retention time (min)	Mass error (ppm), $n = 10$			Precision (%-RSD), $n = 30$		
					Min.	Med.	Max.	Min. (%)	Med. (%)	Max. (%)
<sup>13</sup> C <sub>6</sub> -Methyl paraben <sup>a</sup>	DTXSID30894090	ESI-	158.0675	2.28	0.80	6.50	8.40	0	5	14
<sup>13</sup> C <sub>6</sub> -Butyl paraben <sup>a</sup>	DTXSID90894091	ESI-	200.1144	7.21	0.38	7.88	9.38	0	5	10
<sup>13</sup> C <sub>4</sub> -Perfluorooctanoic acid <sup>b</sup>	DTXSID70892999	ESI-	417.9871	8.33	0.29	3.66	4.26	0	5	15
<sup>13</sup> C <sub>4</sub> <sup>15</sup> N <sub>2</sub> -Fipronil <sup>a</sup>	DTXSID50894092	ESI-	441.9462	9.76	0.24	2.61	3.41	0	3	14
<sup>13</sup> C <sub>4</sub> <sup>15</sup> N <sub>2</sub> -Fipronil sulfone <sup>a</sup>	DTXSID10894093	ESI-	457.9411	10.56	0.24	3.47	3.69	0	6	20
<sup>13</sup> C <sub>5</sub> -Perfluorononanoic acid <sup>b</sup>	DTXSID70894099	ESI-	468.9873	9.45	0.39	2.70	3.24	0	5	14
<sup>13</sup> C <sub>4</sub> -Perfluorooctanesulfonic acid <sup>b</sup>	DTXSID80894101	ESI-	503.9509	9.50	0.03	3.55	3.94	0	4	15
<sup>13</sup> C <sub>2</sub> -Perfluorodecanoic acid <sup>b</sup>	DTXSID20894100	ESI-	515.9740	10.40	0.34	2.38	3.05	0	2	15
<sup>13</sup> C <sub>3</sub> -Atrazine <sup>a</sup>	DTXSID60894088	ESI+	218.1038	5.33	0.28	2.24	7.51	0	5	38
D <sub>3</sub> -Thiamethoxam <sup>c</sup>	DTXSID60746816	ESI+	294.0381	0.85	1.42	6.46	10.94	0	7	29
D <sub>4</sub> -Pyriproxyfen <sup>c</sup>	DTXSID20894089	ESI+	325.1616	11.68	0.00	0.77	12.92	0	5	65

<sup>a</sup> Cambridge Isotope Laboratories, Inc. (Tewksbury, MA, USA)

<sup>b</sup> Wellington Laboratories Inc. (Guelph, Ontario, Canada)

<sup>c</sup> CDN Isotopes Inc. (Pointe-Claire, Quebec, Canada)

<sup>d</sup> DSSTox Substance Identifier

(DTXSID90880494) ( $m/z$  121.0509 and 922.0098, respectively) were used in positive mode, while trifluoroacetic acid (DTXSID9041578) and the formate adduct of HP-0921 ( $m/z$  112.9856 and 966.0073, respectively) were used in negative mode. According to manufacturer specifications, the Agilent 6530B is expected to perform with better than 2 ppm mass accuracy based on ten repeat injections of Reserpine ( $m/z$  609.2807 when monitoring the  $[M+H]^+$  ion).

## Peak picking and alignment

Procedures were performed separately for ten individual experiments, which each included nine samples (three dilutions of a single ENTACT mixture with triplicate injections of each dilution) and six method blanks. Recursive molecular feature extraction (MFE) and alignment was performed using Agilent Profinder software (version B.08). Each extracted molecular feature was defined as the aggregate of  $m/z$  peaks (including isotopologue, adduct, and fragment peaks) relating to a common analyte and represented using a neutral monoisotopic mass, retention time (RT), and composite spectrum ( $MS^1$ ). Ion species were restricted to  $[M+H]^+$  and  $[M-H]^-$  in ESI+ and ESI- modes, respectively, to minimize incorrect assignments of neutral accurate mass and molecular formula.

Thresholds used for feature extraction and alignment are given in ESM Table S3.

## Formula assignment

Aligned features were searched against a database of chemical formulae using the Compound Identification Wizard in Agilent’s Mass Profiler Professional (MPP) software (version 14.9). First, “MS-Ready” structures and their associated molecular formulae were generated for  $\sim 760$  K substances contained within EPA’s Distributed Structure-Searchable Toxicity (DSSTox) Database [18]. Procedures for generating MS-Ready formulae are described in McEachran et al. [19] and involve desalting, desolvation, removal of stereochemistry, and neutralization. Next, predicted MS-Ready formulae were deduplicated, yielding a final list of 142,507 unique formulae that were included in a compound database file (.csv format) used by MPP. Matching of molecular features to MS-Ready formulae was based on similarities in mass and isotopic fidelity (i.e., isotope presence, abundance, and spacing). For a given molecular feature, MPP would assign and output a maximum of one MS-Ready formula with the highest match score (maximum score = 100). Criteria and parameters for feature matching are given in ESM Table S4.

## Feature filtering and flagging

Matched and unmatched molecular features were further processed using custom scripts (available upon request) written in the SAS programming language (SAS Institute Inc., Cary, NC, USA). Briefly, SAS scripts were used to (1) track instrument mass accuracy, RT drift, and signal variability (using tracers); (2) remove duplicate features that were artifacts of recursive MFE; (3) filter out features that were not reproducible across replicate injections; (4) filter out features that were found at similar intensities across various sample dilutions and method blanks; (5) calculate blank-subtracted feature intensities; (6) flag features that were observed using both positive and negative ionization modes; and (7) prepare lists of unique formulae (i.e., no duplicate formulae) to be used for structure searching on the US EPA CompTox Chemicals Dashboard (<https://comptox.epa.gov/dashboard>) (hereafter referred to as the “Dashboard”) [18]. All features searched on the Dashboard (1) were observed in two out of three replicate injections of the highest dilution (~0.5  $\mu\text{M}$ ), (2) had a median peak area (across replicates) 3 $\times$  larger in the highest dilution than in the blanks, (3) had a median peak area 1.5 $\times$  larger in the highest dilution than in the middle dilution (~0.1  $\mu\text{M}$ ), and (4) were assigned a formula (in MPP) with a match score  $\geq 90$  or had a negative mass defect and a formula match (no score threshold) containing a halogen.

## Candidate structure selection

MS-Ready structures, denoted by DSSTox Chemical Identifiers (DTXCIDs), are mapped to registered DSSTox substances (with their associated DTXSIDs) within the DSSTox Database. Formula batch searches ([https://comptox.epa.gov/dashboard/dsstoxdb/batch\\_search](https://comptox.epa.gov/dashboard/dsstoxdb/batch_search)) against the content of the database retrieve DTXSIDs (which can be mixtures, polymers, or single chemicals) based on mappings to MS-Ready structure representations (ESM Figs. S1 and S2). The mapping of MS-Ready DTXCIDs to DTXSIDs enables identification and performance evaluation at both the observed chemical level (DTXCID) and the substance level (DTXSID). For this analysis, performance was evaluated based on the identification of the corresponding MS-Ready DTXCID of all spiked substances.

Following formula batch searches on the Dashboard, the top candidate substances (up to ten, sorted by descending data source counts) corresponding to each searched formula were downloaded (as .xls files) for further examination. Substances in the download files were defined using DTXSIDs and included salts, mixtures, ionic species, and neutral molecules. The top-ranked substance, based on “data source” counts, for each searched formula was deemed the most likely candidate. According to McEachran et al. [20], “data sources” in the Dashboard represent the number of registered lists in the

DSSTox Database that contains a particular chemical. Prevalence across many registered lists is potentially indicative of a chemical’s relative likelihood of occurrence. Additional details and performance metrics of the data source ranking method can be found in McEachran et al. [20].

## Structure corroboration using MS<sup>2</sup> analysis

Feature lists from MS<sup>1</sup> analyses were manually converted into preferred ion lists for automated data-dependent acquisition (DDA) MS<sup>2</sup> (Auto MSMS) using the Agilent LC-QTOF/MS. Specifically, all features that were searched on the Dashboard were ultimately included on a preferred ion list. Selected ions were fragmented at 10, 20, and 40 eV in both positive and negative ionization modes. The  $m/z$  of reference compounds and anticipated contaminants were added to an exclusion list. Windows for mass and RT were set at 20 ppm and 0.3 min, respectively (except for reference masses which were excluded for the entire run). An abundance threshold of 1000 was used with active exclusion of 5 spectra and a 0.3-min release time. The isolation window of the quadrupole was set to medium (4 Da), and data collection for MS<sup>1</sup> and MS<sup>2</sup> were set to 3 spectra/s and 4 spectra/s, respectively. Acquired MS<sup>2</sup> data for preferred ions were matched to a combined Agilent Personal Compound Database and Library (PCDL) using MassHunter Qualitative Analysis software (version B.07). The combined PCDL contained reference mass spectra at 10, 20, and 40 eV for compounds found within the following four Agilent PCDLs: Environmental Water Screening, Pesticides, Forensic Toxicology, and Veterinary Drugs. PCDL matches were identified using forward and reverse scoring thresholds of 25 and 85, respectively. Compound hits, generally identified by chemical name and CASRN, were included in output files and assigned DTXSIDs and MS-Ready DTXCIDs.

## Evaluation of spiked substances vs. observed features

After unblinding, NTA method performance evaluation began with processing of each ENTACT mixture list (Fig. 2). Specifically, an MS-Ready DTXCID, MS-Ready formula, and MS-Ready monoisotopic mass were first assigned to each spiked substance (DTXSID). It was next determined (1) how often each DTXSID occurred across mixtures, (2) how often each MS-Ready DTXCID occurred within and across mixtures, (3) whether each MS-Ready formula occurred more than once per mixture (isomers), and (4) whether each MS-Ready monoisotopic mass occurred (within 0.005 amu) more than once per mixture.

After processing and annotation, each list of spiked substances was compared to the respective list of observed features. Matches between spiked substances and observed features occurred at the MS-Ready DTXCID, MS-Ready



confidence in the formula assignment and spiked substance match, the feature was given a “pass” rating. If the evidence was not strong enough to support a match, the feature was given an “undetermined” rating. Finally, if the evidence supported a mismatch, the feature was given a “fail” rating. All features assigned an “undetermined” or “fail” rating were removed from the final dataset. All features assigned a “pass” rating were maintained in the final dataset and aligned using ionization mode and RT information. This alignment yielded a master list of observed ENTACT substances to be used as a matching library in future experiments.

Method performance was determined for each ENTACT mixture using true positive rate (TPR) estimates. For this study, TPR is defined as the number of correctly identified substances divided by the number of spiked substances. Within the final dataset for each ENTACT mixture, it was common for multiple features with “pass” ratings to match to a single spiked substance. This occurred when spiked substances were observed in both ionization modes or when multiple isomer peaks were observed for a given spiked substance. To reconcile these instances with TPR calculations, only the highest ranking matched feature (based on star ratings) was considered for each spiked substance. The “Blinded TPR” (Eq. 1) was then defined as the sum of 4- and 5-star “pass” features divided by the number of spiked substances. This statistic communicates how often a substance could be correctly identified, at the MS-Ready DTXCID level, using the NTA workflow. A second statistic, called the “unblinded TPR” (Eq. 2), was defined as the sum of all “pass” features divided by the number of spiked substances. This statistic communicates how often a substance could be correctly identified given that its presence in the sample was known.

$$\text{Blinded TPR} = \frac{5 \text{ Star Passes} + 4 \text{ Star Passes}}{\text{Spiked Substances}} \quad (1)$$

$$\text{Unblinded TPR} = \frac{\text{All Passes}}{\text{Spiked Substances}} \quad (2)$$

By design, different ENTACT mixtures had different numbers of spiked isomers [17]. In many instances, it was not possible to resolve spiked isomers, particularly when feature identification was based on MS<sup>1</sup> data. As such, adjustments were made to the TPR calculations to allow for better comparability in method performance statistics across the mixtures. Equations 3 and 4 show how adjustments were made to blinded TPRs and unblinded TPRs, respectively.

$$\begin{aligned} &\text{Blinded Adj. TPR} \\ &= \frac{5 \text{ Star Passes} + 4 \text{ Star Passes} - \text{Spiked Isomer Passes}}{\text{Spiked Substances} - \text{Spiked Isomers}} \end{aligned} \quad (3)$$

$$\begin{aligned} &\text{Unblinded Adj. TPR} \\ &= \frac{\text{All Passes} - \text{Spiked Isomer Passes}}{\text{Spiked Substances} - \text{Spiked Isomers}} \end{aligned} \quad (4)$$

## Results

### Evaluation of tracer compounds

#### Mass accuracy

Eleven stable isotope-labeled tracer compounds (Table 1) were monitored across all experimental runs ( $n = 150$ ). Tracers ranged in monoisotopic mass from 158.0675 amu (<sup>13</sup>C<sub>6</sub>-methyl paraben) to 515.9740 amu (<sup>13</sup>C<sub>2</sub>-perfluorodecanoic acid) and in RT from 0.85 min (D<sub>3</sub>-thiamethoxam) to 11.68 min (D<sub>4</sub>-pyriproxyfen) (solvent interference was observed out to ~0.4 min). Median mass error estimates for individual tracers ranged from 0.77 to 7.88 ppm (global median = 3.05 ppm). Several tracers had a maximum mass error that approached or exceeded 10 ppm. Nearly all tracers were observed at < 1 ppm mass error in at least one ENTACT mixture.

#### Measurement precision

Precision estimates are reported as relative standard deviations (%-RSD = [standard deviation/mean × 100]) and reflect the variability in observed feature intensities across three replicate measures (Table 1). For eight out of eleven tracers, all RSD estimates (30 per tracer) were ≤ 20%. An overall median RSD of 5% was observed across 330 RSD estimates (3 concentrations × 11 tracers × 10 mixtures). The minimum observed RSD for each of the 11 tracers was < 1%.

### Feature characterization and identification

#### Observed and filtered features

Between 1822 and 4168 molecular features were observed in the individual mixtures (Table 3), considering data from both positive and negative ionization modes (positive mode features generally outnumbered negative mode features by a factor of 4). After removing non-reproducible features and markers of background contamination, between 547 and 2179 filtered features remained in the datasets (Table 3). Figure 3a shows an example comparison of features that were kept vs. those that were removed via data filtering. In this example, 4168 total features were observed for mixture 505 before filtering, which contained 365 spiked substances. Only 52% of these features were

**Table 3** Descriptive statistics, based on MS<sup>1</sup> and MS<sup>2</sup> results, for ten ENTACT mixtures

ENTACT mixture ID	Analysis sequence <sup>a</sup>	Spiked substances	MS <sup>1</sup> results			MS <sup>2</sup> results		
			Observed features <sup>b</sup>	Filtered features <sup>c</sup>	Searched formulae <sup>d</sup>	Library chemicals <sup>e</sup>	Library matches	Match percentage <sup>f</sup>
499	4	95	1914	619	179	38	10	26
500	10	95	1862	547	120	36	0	0
501	3	95	1822	956	245	40	11	28
502	2	95	2091	1208	241	40	10	25
503	5	185	2458	999	256	70	15	21
504	6	185	2678	1128	298	68	10	15
505	1	365	4168	2179	426	143	30	21
506	7	365	3507	2084	478	157	18	11
507	8	95	2201	646	133	30	4	13
508	9	364	2797	1377	249	132	4	3

<sup>a</sup> The order in which mixtures were analyzed using LC-QTOF/MS

<sup>b</sup> The number of molecular features observed via MFE and alignment

<sup>c</sup> The number of observed features that passed through custom filtering scripts

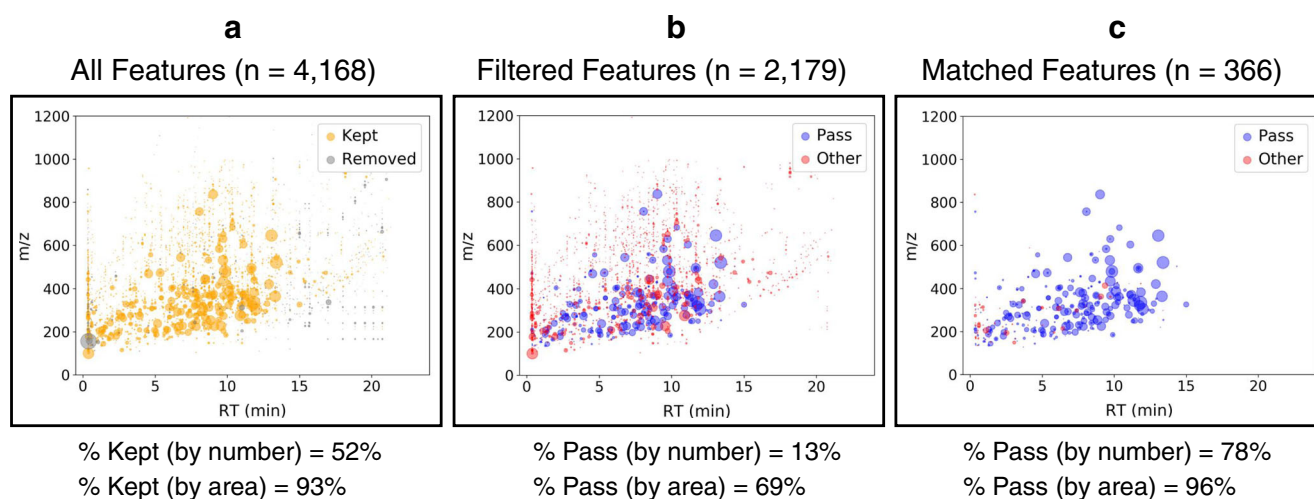
<sup>d</sup> The number of de-duplicated formulae, assigned to observed features, that were flagged and searched on the Dashboard

<sup>e</sup> The number of compounds in the combined PCDL library that had (1) MS<sup>2</sup> data and (2) enough information to map to a MS-ready DTXCID

<sup>f</sup> Match percentage = (library matches/library chemicals) × 100

ultimately kept in analysis. Yet, the summed peak area of these kept features comprised 93% of the total peak area for mixture 505. Percentages of features kept in the

analysis, calculated by peak number and area, are presented in ESM Table S5 for all ENTACT mixtures. Bubble plots for all ENTACT mixtures are given in ESM Table S6.



**Fig. 3** Bubble plots for molecular features observed in ENTACT mixture 505. Each feature is represented by a bubble positioned on the x- and y-axis according to observed retention time and *m/z*, respectively. The diameter of each bubble is proportional to the peak area of the observed feature. Panel **a** shows a comparison of features that were kept in the analyses after initial filtering (yellow bubbles) vs. those that were removed (gray bubbles). Here, 52% of the total features (comprising 93% of the total peak area) were kept. Panel **b** shows all filtered features (*n* = 2179) with “pass” features (i.e., those that *correctly* matched to a spiked substance) represented by blue bubbles, and all

other features (i.e., those that could not be matched to spiked substance) represented by red bubbles. Here, 13% of the filtered features (comprising 69% of the total peak area) were “pass” features. Panel **c** shows all filtered features that matched to a spiked substance, based on MS-Ready monoisotopic mass, within a 15-ppm mass error window (*n* = 366). Here, “pass” features, or *correct* matches, are again shown as blue bubbles. Seventy-eight percent of mass matches were ultimately “pass” features. The summed peak area of these correct matches was 96% of the total peak area for all mass matches



## Molecular formula assignments

Approximately one-quarter of filtered features were assigned a MS-Ready formula with an acceptable level of confidence (i.e., feature had a formula match score  $\geq 90$  or a negative mass defect and a formula match [no score threshold] containing a halogen). For each mixture, between 120 and 478 unique matched formulae were ultimately searched on the Dashboard and utilized as the basis for DDA MS<sup>2</sup> preferred ion lists (Table 3). The numbers of searched formulae were generally comparable to the numbers of spiked substances, with ratios of searched formulae to spiked substances ranging from 0.68 to 2.58.

## PCDL (MS<sup>2</sup>) matches

Across four Agilent PCDLs, MS<sup>2</sup> data were identified as being available for 497 out of 1269 unique substances that were spiked into the ENTACT mixtures. Initial DDA MS<sup>2</sup> experiments yielded between 0 and 30 hits for each ENTACT mixture (Table 3). This corresponds to match percentages ranging from 0 to 28%. In other words, the selected DDA MS<sup>2</sup> method correctly identified no more than one third of the substances that were spiked into a mixture and contained (with MS<sup>2</sup> data) in the Agilent PCDLs.

## Performance on individual ENTACT mixtures

### True positive rates

Table 4 shows the number of spiked isomers for each ENTACT mixture, along with sensitivity estimates with and without adjustment for isomers. Only two mixtures (501 and 502) were free of spiked isomers. Six mixtures (499, 500, 503, 504, 505, and 506) had a modest proportion of spiked isomers ( $\leq 5\%$ ), and the remaining two mixtures (507 and 508) had a higher proportion ( $\sim 55\%$ ) of spiked isomers. Without adjustment for the presence of isomers, blinded TPRs ranged from 5 to 46%. After unblinding, the number of spiked substances that could be correctly identified (via manual review) rose considerably. Specifically, unblinded TPRs ranged from 9 to 62%. When isomers were not considered in sensitivity calculations, the unblinded Adj. TPR estimates ranged from 19 to 65%. Estimates for eight of the ten mixtures were generally comparable (44–65%), whereas estimates for two mixtures (500 and 508) were considerably lower ( $\sim 20\%$ ).

The order of analysis (Table 4) indicates diminished performance for ENTACT mixtures that were run towards the end of the initial experiments. Specifically, TPR estimates were lowest for mixtures 508 and 500, which were the last two mixtures to be analyzed in the sequence. Results for labeled tracers (ESM Table S7) show modest decreases in mass accuracy, measurement precision, and

signal strength for these two mixtures. Thus, a decline in instrument cleanliness over the run sequence likely contributed to a decrease in performance. A second analysis, at a later time, was therefore performed on mixtures 500 and 508. Mixture 502 was also included in the second analysis as a QC sample, as some method parameters were altered to improve identifications (method 2, ESM Table S2). Results of the second analysis are given in parentheses in Table 4. Unblinded results for mixture 502 were identical when comparing results of the first and second analyses. Blinded and unblinded results for mixtures 500 and 508 were substantially better in the second analysis compared to those in the first analysis. The updated results (Table 4) suggest that 56 to 59% of spiked substances (excluding isomers) could be observed in these samples, after unblinding. While results of the second analysis are clearly superior for mixtures 500 and 508, results of the first analysis are the basis for all summary statistics (in subsequent sections) relating to method performance.

## Star ratings for observed substances

Between 19 and 225 spiked substances were observed, per mixture, after unblinding and manual evaluation of matched features (Table 4). Figure 4 shows the percentages of these observed substances that were matched at the 1-through 5-star levels. On average, 14% of observed substances were matched at the 5-star level; the MS-Ready structures of these substances were correctly identified based on matching of observed vs. library (PCDL) MS<sup>2</sup> spectra. Fifty-six percent of observed substances were matched at the 4-star level; the MS-Ready structures of these substances were correctly identified via formula matching (based on MS<sup>1</sup> data) and assignment of the top-ranked Dashboard structure. Twelve percent of observed substances were matched at the 3-star level; the MS-Ready formulae of these substances were correctly identified, but the corresponding MS-Ready structures were not Dashboard top hits (they were, however, amongst the top 10 Dashboard hits). No observed substances were matched at the 2-star level. In other words, every time the MS-Ready formula of a spiked substance was correctly identified, the corresponding MS-Ready structure was amongst the top 10 Dashboard hits. Eighteen percent of observed substances were matched at the 1-star level; the MS-Ready monoisotopic masses of these substances were correctly observed, but formulae (and ultimately structures) were incorrectly assigned using automated techniques. Pooling results across star bins,  $\sim 70\%$  of observed substances were correctly identified at the MS-Ready structure level (4- and 5-star hits), and 82% were correctly identified at the MS-Ready formula level (3-, 4-, and 5-star hits), on average.

**Table 4** Blinded and unblinded analysis results for ENTACT mixtures. Estimates of “true positives,” “TPR,” and “Adj. TPR” are based on method 1 (Supplemental Table 2). Values given in parentheses are based on a second analysis of mixtures 500, 502, and 508 using method 2 (Supplemental Table 2)

ENTACT mixture	Order of analysis	Spiked substances	Spiked isomers <sup>a</sup>	Blinded results			Unblinded results		
				True positives <sup>b</sup>	TPR <sup>c</sup>	Adj. TPR <sup>d</sup>	True positives <sup>e</sup>	TPR <sup>f</sup>	Adj. TPR <sup>g</sup>
499	4	95	2	33	0.35	0.35	46	0.48	0.49
500	10	95	2	12 (35)	0.13 (0.37)	0.13 (0.37)	19 (53)	0.20 (0.56)	0.20 (0.56)
501	3	95	0	26	0.27	0.27	47	0.49	0.49
502	2	95	0	44 (36)	0.46 (0.38)	0.46 (0.38)	58 (58)	0.61 (0.61)	0.61 (0.61)
503	5	185	2	66	0.36	0.36	103	0.56	0.56
504	6	185	2	81	0.44	0.44	103	0.56	0.56
505	1	365	18	156	0.43	0.45	225	0.62	0.65
506	7	365	2	144	0.39	0.40	195	0.53	0.54
507	8	95	52	18	0.19	0.42	19	0.20	0.44
508	9	364	207	19 (80)	0.05 (0.22)	0.11 (0.42)	31 (107)	0.09 (0.29)	0.19 (0.59)

<sup>a</sup> Determined when two spiked substances shared the same MS-Ready formula

<sup>b</sup> Number of spiked substances for which a “pass” feature was observed at the 4- or 5-star level

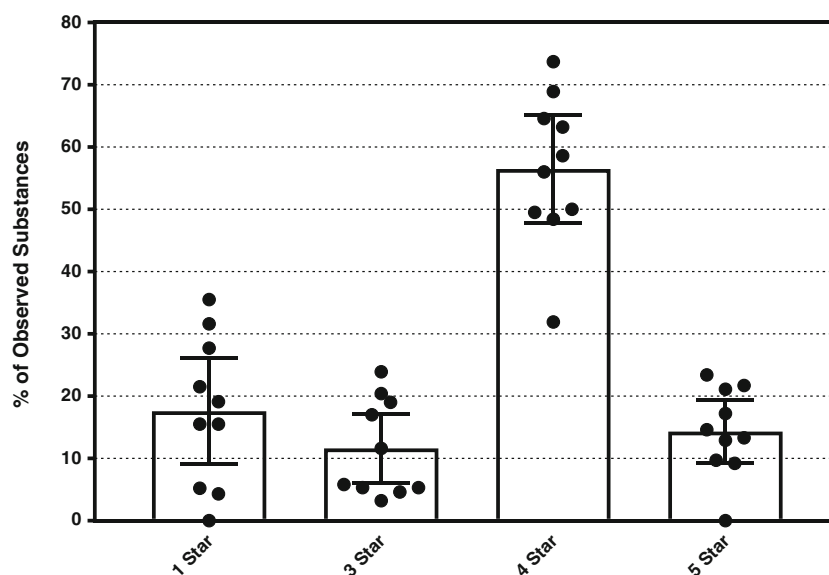
<sup>c</sup> Calculated according to Eq. 1

<sup>d</sup> Calculated according to Eq. 3

<sup>e</sup> Number of spiked substances for which a “pass” feature was observed at any star level

<sup>f</sup> Calculated according to Eq. 2

<sup>g</sup> Calculated according to Eq. 4



**Fig. 4** Percentage of observed substances according to star ratings. ENTACT mixtures ( $n = 10$ ) contained between 95 and 365 spiked substances. Molecular features were observed upon analysis of each mixture and matched to spiked substances according to MS-Ready monoisotopic mass, MS-Ready formula, and/or MS-Ready structure. Each feature match was manually reviewed and assigned a rating of 1–5 stars as described in Table 2. Multiple feature matches were observed

for certain spiked substances due to chromatographic separation of isomers or identification in both ESI+ and ESI-. Only the highest ranking feature match for each spiked substance was used in percentage calculations. Each black circle represents the percentage of observed substances, for an individual mixture, that was assigned a specific star rating. Bar height represents the mean percentage for each star-rating category, and error bars represent the 95% confidence intervals

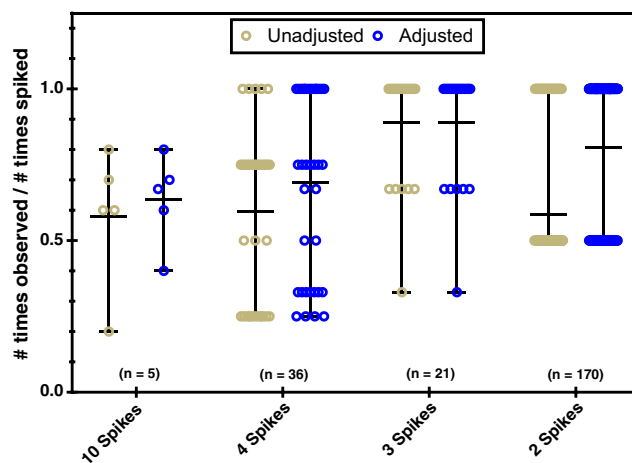
## Reproducibility across mixtures

Four hundred eighty-three substances were spiked into more than one ENTACT mixture, and 232 (48%) of these substances were observed at least once. For these 232 substances, a total of 597 substance hits were possible given the study design ([5 substances  $\times$  10 spikes] + [36 substances  $\times$  4 spikes] + [21 substances  $\times$  3 spikes] + [170 substances  $\times$  2 spikes]). Out of a possible 597 hits, 370 were ultimately recorded, meaning that reoccurring substances were observed 62% of the time.

Many of the substances that were spiked across multiple mixtures were, at times, concurrently spiked with isomers (ESM Table S8). For example, diethyl phthalate was spiked into all ten mixtures—in five mixtures, it was the only spiked substance with the MS-Ready formula  $C_{12}H_{14}O_4$ ; in five other mixtures, it was spiked alongside another substance that shared the same MS-Ready formula. Since  $MS^1$  data could not be used to differentiate isomers, and since most substance hits were based on  $MS^1$  data (Fig. 4), reproducibility statistics were recalculated after subtracting isomer occurrences (597 possible hits – 102 isomer occurrences = 495 possible hits). After adjustment, reoccurring substances were observed 75% of the time. Figure 5 shows reproducibility statistics for groups of substances based on the number of spikes (two, three, four, or ten). ESM Table S8 shows full reproducibility statistics for all 1269 unique ENTACT substances.

## Comparison across methods

Three of the ten ENTACT mixtures were analyzed a second time using different chromatography methods (ESM Table S2). Mixtures 500 and 508 were reanalyzed due to poor results from the first analysis (Tables 3 and 4), and mixture 502 was reanalyzed as a QC check. Whereas unblinded Adj. TPRs for mixture 502 were identical across the first and second analyses (61%), differences were observed in the blinded results, indicating some effect of the selected method on NTA performance. Ninety-five substances were spiked in mixture 502; a side-by-side comparison showed that ten substances were observed in the second analysis but not the first and that another ten were observed in the first analysis but not the second. In six of these 20 instances (e.g., diphenyl phthalate and oxycarboxin), an  $m/z$  for the substance was observed, but the correct formula was not confidently assigned. In two instances (veratraldehyde and benzocaine), the  $m/z$  for the substance was observed, but the feature was filtered out after not meeting fold-change thresholds. In the remaining 12 instances (e.g., 8-hydroxyquinoline and dimethyl phthalate),  $m/z$  values corresponding to the spiked substances could not be observed (either truly not present or missed by the peak picker) using either the first or second analysis method.



**Fig. 5** Reproducibility of NTA feature identification methods. Four hundred eighty-three substances were spiked more than once across the ten ENTACT mixtures. Three hundred eighty-eight substances were spiked in two mixtures; 170 (44%) of these substances were observed at least once. Thirty-three substances were spiked into three mixtures; 21 of these substances (64%) were observed at least once. Fifty-seven substances were spiked into four mixtures; 36 (63%) of these substances were observed at least once. Five substances were spiked into all 10 mixtures; all five of these substances were observed at least once. The quotient of # times observed / # times spiked is displayed separately for substances that were spiked two, three, four, and ten times and that were observed at least once. Here, circles represent individual substances, and horizontal lines represent the minimum, mean, and maximum quotient for each group. Some substances were included in mixtures with isomers. Since identification at the  $MS^1$  level could not distinguish isomers, an adjustment was made to the quotient calculation to account for this limitation (i.e., adjusted quotient = # times observed / [# times spiked – # times spiked with isomer])

Four hundred fifty-six unique substances were spiked into mixtures 500, 502, and 508. Two hundred fifteen (47%) of these substances were observed in the first analysis (considering results across all ten mixtures), and 193 (42%) were observed in the second analysis. Discordance between results of the first and second analyses was suspected for at least 93 out of the 456 unique substances (criteria for discordance given in ESM Table S8). Of the 93 discordant features, 31 were observed in the first analysis but not the second, and 62 were observed in the second analysis but not the first. All discordant features are listed in ESM Table S8.

## Overall performance

Seven hundred fifty-eight substances could be identified, as part of the first or second analysis, out of 1269 unique spiked substances (ESM Table S8). It can therefore be concluded that at least 60% of the spiked substances are well suited to NTA using LC-QTOF/MS instrumentation with electrospray ionization. One hundred thirty-one of these compounds were observed in both the first and second analyses, suggesting that they are amenable to LC-MS-based NTA under a variety of experimental conditions. In contrast, 179 substances (out of

456) could not be observed despite being screened with two LC methods. Ten of these substances (e.g., monobutyl phthalate, acetaminophen, and estrone) were always spiked alongside an isomer, making them particularly difficult to identify at the structure level. An additional 332 substances were screened in the first analysis but not the second, and not observed. Six of these substances were always spiked alongside an isomer. Combining non-detect lists, 495 substances ( $[179 - 10] + [332 - 6]$ ) are suspected of not being amenable to the *utilized* LC-MS-based NTA methods. Importantly, 178 of these substances were classified as “undetermined” upon manual evaluation (ESM Table S8), meaning that some evidence exists for their presence in the mixtures. More work is needed, however, examining individual standards of each substance, to definitively categorize each substance as a “pass” or “fail.” Until then, lists of substances that were ( $n = 758$ ) and were not ( $n = 495$ ) observed are given in ESM Figs. S3 and S4, respectively.

## Discussion

For decades, environmental monitoring data have provided a basis for regulatory decisions that impact public health. Monitoring studies have, in large part, focused on small lists of target substances and utilized robust methods with demonstrable accuracy, precision, sensitivity, and reliability. Recent applications of NTA methods have highlighted numerous CECs in both outdoor and indoor environments and in products [1, 4, 5]. Targeted methods cannot keep pace with the discovery of these CECs [21]. NTA methods must therefore be developed to a point of providing defensible, albeit preliminary, monitoring data to enable risk-based prioritizations and provisional assessments. The ENTACT project was developed as a large litmus test for the NTA research community with the overarching goal of establishing performance baselines and benchmarks [15, 17]. This article represents a broad and transparent examination of ten ENTACT mixtures, each containing 95–365 substances, using LC-QTOF/HRMS. The discussion material provided herein focuses on strengths and limitations of our adopted methods, as well as broader NTA challenges that must be addressed by the research community.

## Summary of findings

A total of 1269 unique substances were spiked across the ten ENTACT mixtures. Taking the best results for each mixture (from either the first or second analysis), blinded TPRs ranged from 19 to 46% (Table 4). This means that, for all mixtures, less than half of the spiked substances could be correctly identified when they were unknown. Even after adjusting for spiked isomers, no more than 46% of a mixture’s contents were correctly identified at the blinded stage of the analysis.

These results highlight deficiencies of NTA for environmental monitoring. Compounds that are truly present in a medium can and will go undetected; which chemicals go undetected appears to be a function of method selection and implementation. Hence, practitioners must develop strategies for characterizing and communicating false negative and false positive rates in monitoring studies.

After unblinding, 758 substances (60%) were ultimately observed using the selected methods. Only 14% (averaged across samples) of these observed substances were identified using MS<sup>2</sup> data (Fig. 4). This low percentage stems, in part, from limited overlap between spiked substances and library compounds with MS<sup>2</sup> data (Table 3). Eighty-two percent of the observed substances were assigned the correct molecular formula using MS<sup>1</sup> data and automated matching techniques. Most of these substances were further assigned the correct MS-Ready DTXCID using functions within EPA’s Chemicals Dashboard. Only 18% of observed substances were initially assigned an incorrect formula or assigned the correct formula with a low match score (Fig. 4). These results highlight a need for more comprehensive MS<sup>2</sup> libraries, the further development and utilization of predicted MS<sup>2</sup> spectra, the implementation of optimized formula prediction algorithms, and the discovery and use of optimized data processing settings.

To our knowledge, ENTACT is the first NTA trial to allow assessment of method performance and reproducibility across synthetic mixtures with varying numbers of spiked substances and isomers. Considering the unblinded results (Table 4), the presence of isomers had a much stronger effect on performance than did the number of spiked substances. Mixtures 507 and 508 had, by far, the largest proportion of spiked isomers; these two mixtures yielded an average unblinded TPR of 0.25 vs. 0.55 for the remaining eight mixtures (using the highest TPRs for each mixture). After adjusting for isomers, these two mixtures yielded an average unblinded Adj. TPR of 0.52, which is nearly equal to that of the remaining mixtures. Regarding the number of spiked substances, the 95-, 185-, and 365-substance mixtures yielded average unblinded Adj. TPRs of 0.52, 0.56, and 0.59, respectively. This result clearly shows no negative impact of the number of substances on method performance. A reproducibility estimate of 75% was calculated for compounds that were repeatedly spiked (but not alongside isomers) and observed at least once. Importantly, this estimate reflects results from the first analysis only, where poor performance was observed for mixtures 500 and 508. A higher estimate is expected under optimal working conditions. The statistics reported here, however, reflect an accurate and honest accounting of our initial performance in this trial.

Forty percent of spiked substances ultimately could not be identified at the MS-Ready DTXCID level. Many of these substances simply were not observed as a molecular feature, that is, they likely were not amenable to LC-MS using ESI,

selected chromatography, and/or identification criteria. Others were observed as a molecular feature but filtered out during data processing. Finally, many substances were correctly identified at the MS-Ready formula level but could not be annotated at the structure level due to the presence of a spiked isomer and a lack of MS<sup>2</sup> library data. Future research will closely examine the basis for missed substances to help optimize and broaden our NTA workflow. The following sections provide details on issues that may have limited correct identifications of ENTACT substances and that will be the focus of future investigations.

## Data analysis considerations

### Data filtering and mass assignment

Custom scripts were written to filter out features that represented background contaminants and artifacts of peak picking and alignment processes. Results in ESM Table S5 indicate that processing scripts generally removed about half of the features for a given mixture, which comprised a small portion (~10% on average) of the total peak area. These results indicate the highly effective removal of uninformative features. Filtering criteria, however, may have been too restrictive in some instances. For example, veratraldehyde and benzocaine were both erroneously filtered out in the first analysis of mixture 502 after not meeting fold-change requirements. Less restrictive criteria would have allowed for their ultimate identification, but at the cost of other potential false positives. This example highlights the delicate balance of false positive and false negatives when performing NTA and underscores the care that should be taken when establishing filtering criteria.

Around 10% of filtered features were ultimately matched to spiked substances, suggesting that other real compounds were present in the mixtures (ESM Table S5). Features that were matched to spiked substances generally comprised about half of the total peak area. This indicates that certain unmatched features had large peak intensities (ESM Table S6). It is very likely that some unmatched features represent unknown interaction products, degradants, or impurities of the spiked substances. It is also likely that some represent spiked substances that were misidentified at the monoisotopic mass level. Our feature extraction method (ESM Table S3) allowed for the identification of only two types of ion species, i.e., +H and -H; early experiments with a separate mixture highlighted many incorrect mass and formula assignments when allowing for ion species other than +H and -H. It is likely that certain spiked substances were observed only as other ion species (e.g., +NH<sub>4</sub>, +Na, +HCOO) or after a neutral loss (e.g., H<sub>2</sub>O, CO<sub>2</sub>). In both instances, the neutral monoisotopic mass of the feature would be incorrectly assigned, and the spiked substance never correctly identified. Considering these findings, we have updated our processing scripts to flag features

that were likely identified as multiple ion species in a given mode. More work is needed, however, to improve the identification of features that are not observable as +H or -H ion species.

### Formula matching

Our NTA workflow depended heavily on matching molecular features to MS-Ready formulae. This step made use of MS<sup>1</sup> data and was necessary since so few matches were based on MS<sup>2</sup> data. Nearly one fifth of the observed substances were initially assigned an incorrect formula or assigned the correct formula with limited confidence. Moreover, for several ENTACT mixtures (namely, 500, 507, and 508), only a small number of spiked substances observed by mass (within 15 ppm) could be identified at the formula level and assigned a “pass” rating (ESM Table S5). Poor formula matching was often a consequence of episodic drops in mass accuracy (>5 ppm), likely due to a decline in instrument cleanliness over the run sequence. In other cases, competing formulae were given similar high scores, and only the top-ranked formula was included in the output. The first scenario will be addressed in future experiments via careful monitoring of instrument performance (using tracer compounds) and adhering to well-defined acceptance criteria. The second scenario will be addressed by migrating to mass-based searches on the Dashboard ([https://comptox.epa.gov/dashboard/dsstoxdb/batch\\_search](https://comptox.epa.gov/dashboard/dsstoxdb/batch_search)) and making use of additional formula generation tools and MS<sup>2</sup> data. Efforts are underway to examine the benefits of alternative formula matching strategies using compiled ENTACT data from collaborating labs.

### Structure searching using MS<sup>1</sup> data

Well over half of the observed substances were correctly identified at the 4- and 5-star levels, and only 12%, on average, were identified at the 3-star level (Fig. 4). These results highlight that top-ranked Dashboard hits were most often good matches for spiked substances when molecular formulae were correctly assigned. They further reflect the merits of metadata-based ranking schemes coupled with a large, high-quality chemical database for selecting amongst candidate compounds. The results are potentially misleading, however, in that all spiked substances are ToxCast chemicals and, therefore, are likely to have high data source counts within the Dashboard. Yet, they mirror those of McEachran et al. [20] who observed an 88% success rate for substance identification when searching the Dashboard by formula (for 162 compounds) and ranking by data source counts. The level of success in the current and prior investigations, again, likely reflects the types of chemicals being examined. To date, our evaluations have mostly considered well-known substances, including pharmaceutical drugs, industrial chemicals, pesticides, personal care product chemicals, and perfluorochemicals. As more

“emerging” contaminants are encountered in NTA studies, more rigorous sorting schemes will be required for accurate identifications, particularly when queries are based entirely on MS<sup>1</sup> data. Work on the Dashboard continues in terms of adding new data and developing new functionality to support NTA. Future developments are intended to allow for weighted sorting of candidate substances based on several metrics, including data sources and reference counts from other databases (e.g., PubChem, PubMed), presence in priority lists (e.g., ToxCast, NORMAN Suspect List Exchange [<http://www.norman-network.com/?q=node/236>]), and likelihood of being measured in a given medium (e.g., water, house dust) and with a given method (e.g., LC-MS ESI+). Retention time prediction has recently been evaluated as both a scoring term and threshold cutoff for candidate selection [22]; broader implementation of RT predictions is underway.

### Library matching using MS<sup>2</sup> data

About 40% of spiked substances were found in Agilent PCDLs with MS<sup>2</sup> reference data (other Agilent PCDLs contain data on additional ToxCast substances). For those substances with MS<sup>2</sup> data, less than one third were ultimately matched at the 5-star level using DDA and Agilent’s Auto MSMS feature extraction (Table 3). These statistics highlight two limitations of our current workflow. First, there was too little coverage of ToxCast chemicals in the utilized PCDLs. ENTACT is attempting to shrink this limitation by making the full ToxCast chemical library (over 4000 substances) available to several instrument vendors to enable the expansion of reference libraries [15, 17]. Obviously, this will provide coverage for only a small portion of the known chemical universe. As such, we are working to make use of other open MS<sup>2</sup> libraries, as well as predicted MS<sup>2</sup> spectra. Recent integration between the Dashboard and the *in silico* fragmentation tool MetFrag [23] has been enabled via the Batch Search menu ([https://comptox.epa.gov/dashboard/dsstoxdb/batch\\_search](https://comptox.epa.gov/dashboard/dsstoxdb/batch_search)), allowing users to input results and metadata from the Dashboard directly into MetFrag [19]. Predicted MS<sup>2</sup> spectra for all structures in DSSTox have also been generated using the *in silico* fragmentation tool CFM-ID and command line utilities [24, 25]. Predicted spectra have been stored, and programmatic access was developed for spectra retrieval and matching to experimental spectra. Future work will make these data and services available via the Dashboard.

The second limitation of our workflow, as it pertains to MS<sup>2</sup> spectral matching, appears to be the use of overly-stringent match criteria. Automated library matching was performed by searching experimental MS<sup>2</sup> spectra against reference MS<sup>2</sup> spectra, with matches meeting or exceeding user-defined thresholds. Matching thresholds were set at stringent levels to yield only confident matches. It was observed that reference MS<sup>2</sup> spectra present in the Agilent PCDLs varied

significantly in complexity, most likely due to differences in compound structure and associated fragmentation pathways. This variability in spectral complexity can affect the quality of library match scores; matches with complex reference MS<sup>2</sup> spectra will be penalized for peaks not identified in experimental MS<sup>2</sup> spectra and will thus have reduced scores. Poor matching is exacerbated when experimental spectra include matrix/noise peaks and when instrument mass accuracy is not optimal (as was the case for mixtures 500 and 508 [Table 3]). Overly-stringent scoring thresholds can prevent accurate compound identifications even in the face of confident fragment matches. The use of lower scoring thresholds can lead to more correct matches but necessitates careful manual review of each match to ensure confident identifications.

### Manual evaluation

Extracted ion chromatograms and observed spectra were manually reviewed for all molecular features that matched to a spiked substance (based on MS-Ready DTXCID, formula, or mass). The manual evaluation was performed to minimize false positives and negatives, while allowing an assessment of blinded vs. unblinded performance. The determination as to whether a feature was classified as “pass,” “undetermined,” or “fail” was somewhat subjective and, therefore, not considered 100% reliable. In other words, it was impossible to assign a classification with absolute certainty since results of individual standards were not generated and compared against those of the mixtures. Moving forward, all “pass” features from this analysis are referred to as “probable structures” using the criteria of Schymanski and colleagues [26]. Specifically, no substances will be considered “confirmed” unless existing results are examined against those of a single reference standard. All substances given a “pass” rating at the 5-star level are considered *level 2a* hits since matching was based on MS<sup>2</sup> reference spectra. All substances given a “pass” rating at the 1- through 4-star levels are considered *level 2b* hits since enough diagnostic evidence exists to support a substance match. We note that confidence levels proposed by Schymanski and colleagues [26] were not used in our analysis, as they are meant to communicate confidence in structure assignments when examining features observed in “real” samples. The current analysis focused only on prepared mixtures, where sample contents were known, and success rates ascertained via comparison of spiked substances vs. observed features. Our scoring criteria were therefore developed to reflect the strength of a given match, rather than the confidence in an initial structure assignment.

### Analytical chemistry considerations

An overarching goal of ENTACT is to determine how analytical chemistry affects NTA performance [17]. In general, our

applied method(s) identified substances that are known to ionize well in ESI like amides, amines, esters, and carbamates (ESM Fig. S3). Conversely, our method(s) generally did not identify substances that possess the aniline, phenol, and nitroso moieties (ESM Fig. S4). Several key factors likely influenced our inability to observe certain substances. First, in-source fragmentation may have caused neutral losses leading to misidentification at the monoisotopic mass level. Second, some analytes did not possess ionizable functional groups that could improve the acquisition or loss of a proton during the ionization process. Third, some compounds are known to have poor ionization efficiencies, like estrogens and sugars, which makes them hard to observe, particularly at lower concentrations. Fourth, very polar compounds that elute right after void volume suffer from ionization suppression. Fifth, some spiked substances have a permanent positive charge, meaning that monoisotopic mass would be incorrectly assigned when assuming detection as  $[M+H]^+$ . Finally, small molecules (mass < 200 Da) that possess multiple ionizable groups may be doubly charged and, therefore, fall below the  $m/z$  range set for the analysis.

Some apparent anomalies were observed throughout our result files. For example, we did not observe 2,2'-(4-methylphenyl)imino]diethanol, diphenylamine, and thidiazuron, all of which are expected to be LC-MS amenable due to the ease by which nitrogen-containing compounds acquire a proton in positive mode. On the other hand, we did observe 2,2'-(oxydimethanediyl)bis(2-ethylpropane-1,3-diol) and pentaethylene glycol, which are presumed unlikely to ionize using ESI because of the lack of ionizable functional groups. It is possible that these anomalies represent a small pool of false negative or positive findings. Careful examination and interpretation of each finding at the substance level is beyond the scope of the current investigation. However, more exhaustive analyses, comparing ENTACT results across multiple laboratories, will be the focus of future efforts designed to match individual substances, and classes of substances, to suitable NTA detection methods.

Discordance was observed for a surprising number of substances when comparing results of the first and second analyses. Notable features observed in the first analysis but not the second include TDCPP, propylparaben, and ethionamide. Example features observed in the second analysis but not the first include amiloride hydrochloride, 2H-3,1-benzoxazine-2,4(1H)-dione, and dl-norgestrel. The side-by-side comparison of mixture 502 results showed that discordance stemmed from (1) real features being missed by the peak picker, (2) inaccurate or weak formula assignments, and (3) overly-stringent feature filtering criteria. It is very likely that discordant features across mixtures 500 and 508 stemmed from these same issues. It is further likely that discordance across methods is driven by ionization efficiency issues related to selected mobile phase solvents and buffers. For example, the

use of methanol in negative mode has been shown to improve spray stability leading to better sensitivity [27]. Effects of mobile phases and buffers on substance identification will be further considered in future investigations of larger ENTACT datasets.

### Sample integrity considerations

One hundred sixty-nine spiked substances could not be identified despite screening with two analytical chemistry methods (ESM Fig. S5). The possibility exists that individual substances may not have been observed for reasons other than method suitability. First, sample purity and stability may have affected detection rates, particularly for mixture 508. Second, individual substances may have interacted within a mixture and formed compounds that went uncharacterized. Indeed, ESM Table S5 indicates that many features observed in the mixtures were not associated with spiked substances. The effects of sample purity and stability will be the focus of future efforts that consider ENTACT results across multiple laboratories. Focus will be given to substances that were never observed, regardless of the applied NTA techniques and instrumentation, and to substances that were only observed using specific methods. The goal is to use ENTACT data to calculate the likelihood that a given substance would be observed using a specific NTA method. This will help laboratories optimize methods for specific chemical classes, guide comprehensive strategies for broad-scale chemical characterizations, and inform the use of NTA data in decision contexts.

### Conclusions

Based on the collective findings of this study, it is clear that (1) a single NTA method, even when optimized, can likely characterize only a subset of ToxCast substances (and, by extension, other CECs), (2) the selection of an NTA method will affect which substances are observed as true positives and missed as false negatives, (3)  $MS^2$  reference libraries must be expanded to enable improved compound identifications, (4) rigorous quality control is required to achieve acceptable reliability in NTA studies, and (5) performance evaluation using well-characterized samples should be a requirement of labs generating NTA data to support exposure and health studies. This investigation, although not exhaustive, serves as a guide for those looking to perform NTA self-evaluations using synthetic mixtures or other well-characterized samples. Our statistics further provide a baseline against which results of future experiments can be compared to quantify gains in NTA performance.

**Acknowledgements** The authors thank Annette Guiseppi-Elie, Jennifer Orne-Zavaleta, and Russell Thomas for supporting ENTACT; Katherine

Coutros for her assistance in acquiring ToxCast substances, Kamel Mansouri for his role in developing and implementing MS-Ready processing algorithms; Risa Sayre for her assistance in comparing spiked substances against compounds in Agilent reference libraries; and Sarah Laughlin, Aurelie Marcotte, Dawn Mills, James McCord, Mark Strynar, and Carol Ball (Agilent Technologies) for their contributions to the methods used for sample analysis and data processing. The authors further thank James McCord and Mark Strynar for their thoughtful reviews of this manuscript.

**Funding information** The United States Environmental Protection Agency (U.S. EPA), through its Office of Research and Development (ORD), funded and managed the research described here. Partial support for this work was provided by an award from ORD's Pathfinder Innovation Program. The work has been subjected to Agency administrative review and approved for publication. Randolph Singh and Andrew McEachran 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.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflicts of interest.

**Disclaimer** The views expressed in this paper are those of the authors and do not necessarily represent the views or policies of the U.S. EPA.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

- Phillips KA, Yau A, Favela KA, Isaacs KK, McEachran A, Grulke C, et al. Suspect screening analysis of chemicals in consumer products. *Environ Sci Technol*. 2018;52(5):3125–35. <https://doi.org/10.1021/acs.est.7b04781>.
- Biryol D, Nicolas CI, Wambaugh J, Phillips K, Isaacs K. High-throughput dietary exposure predictions for chemical migrants from food contact substances for use in chemical prioritization. *Environ Int*. 2017;108:185–94. <https://doi.org/10.1016/j.envint.2017.08.004>.
- Newton SR, McMahan RL, Sobus JR, Mansouri K, Williams AJ, McEachran AD, et al. Suspect screening and non-targeted analysis of drinking water using point-of-use filters. *Environ Pollut*. 2018;234:297–306. <https://doi.org/10.1016/j.envpol.2017.11.033>.
- Rager JE, Strynar MJ, Liang S, McMahan RL, Richard AM, Grulke CM, et al. Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring. *Environ Int*. 2016;88:269–80. <https://doi.org/10.1016/j.envint.2015.12.008>.
- Alygizakis NA, Samanipour S, Hollender J, Ibanez M, Kaserzon S, Kokkali V, et al. Exploring the potential of a global emerging contaminant early warning network through the use of retrospective suspect screening with high-resolution mass spectrometry. *Environ Sci Technol*. 2018. <https://doi.org/10.1021/acs.est.8b00365>.
- Gerona RR, Schwartz JM, Pan J, Friesen MM, Lin T, Woodruff TJ. Suspect screening of maternal serum to identify new environmental chemical biomonitoring targets using liquid chromatography-quadrupole time-of-flight mass spectrometry. *J Expo Sci Environ Epidemiol*. 2018;28(2):101–8. <https://doi.org/10.1038/s41370-017-28>.
- Colby JM, Thoren KL, Lynch KL. Suspect screening using LC-QqTOF is a useful tool for detecting drugs in biological samples. *J Anal Toxicol*. 2018. <https://doi.org/10.1093/jat/bkx107>.
- Moritz F, Janicka M, Zygler A, Forcisi S, Kot-Wasik A, Kot J, et al. The compositional space of exhaled breath condensate and its link to the human breath volatilome. *J Breath Res*. 2015;9(2):027105. <https://doi.org/10.1088/1752-7155/9/2/027105>.
- Andra SS, Austin C, Wright RO, Arora M. Reconstructing pre-natal and early childhood exposure to multi-class organic chemicals using teeth: towards a retrospective temporal exposome. *Environ Int*. 2015;83:137–45. <https://doi.org/10.1016/j.envint.2015.05.010>.
- Rubert J, Leon N, Saez C, Martins CP, Godula M, Yusa V, et al. Evaluation of mycotoxins and their metabolites in human breast milk using liquid chromatography coupled to high resolution mass spectrometry. *Anal Chim Acta*. 2014;820:39–46. <https://doi.org/10.1016/j.aca.2014.02.009>.
- Wild CP. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomark Prev*. 2005;14(8):1847–50. <https://doi.org/10.1158/1055-9965.EPI-05-0456>.
- Rappaport SM, Smith MT. Epidemiology. Environment and disease risks. *Science*. 2010;330(6003):460–1. <https://doi.org/10.1126/science.1192603>.
- Andra SS, Austin C, Patel D, Dolios G, Awawda M, Arora M. Trends in the application of high-resolution mass spectrometry for human biomonitoring: an analytical primer to studying the environmental chemical space of the human exposome. *Environ Int*. 2017;100:32–61. <https://doi.org/10.1016/j.envint.2016.11.026>.
- Schymanski EL, Singer HP, Slobodnik J, Ipolyi IM, Oswald P, Krauss M, et al. Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. *Anal Bioanal Chem*. 2015;407(21):6237–55. <https://doi.org/10.1007/s00216-015-8681-7>.
- Sobus JR, Wambaugh JF, Isaacs KK, Williams AJ, McEachran AD, Richard AM, et al. Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. *J Expo Sci Environ Epidemiol*. 2017. <https://doi.org/10.1038/s41370-017-0012-y>.
- Richard AM, Judson RS, Houck KA, Grulke CM, Volarath P, Thillainadarajah I, et al. ToxCast chemical landscape: paving the road to 21st century toxicology. *Chem Res Toxicol*. 2016;29(8):1225–51. <https://doi.org/10.1021/acs.chemrestox.6b00135>.
- Ulrich EM, Sobus JR, Grulke CM, Richard AM, Newton SR, Strynar MJ, Mansouri K, Williams AJ. EPA's Non-Targeted Analysis Collaborative Trial (ENTACT): genesis, design, and initial findings. *Anal Bioanal Chem*. 2018. <https://doi.org/10.1007/s00216-018-1435-6>.
- Williams AJ, Grulke CM, Edwards J, McEachran AD, Mansouri K, Baker NC, et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminform*. 2017;9(1):61. <https://doi.org/10.1186/s13321-017-0247-6>.
- McEachran AD, Mansouri K, Grulke CM, Schymanski EL, Ruttkies C, Williams AJ. “MS-Ready” structures for non-targeted high resolution mass spectrometry screening studies. *J Cheminform*. 2018;10(1):45. <https://doi.org/10.1186/s13321-018-0299-2>.
- McEachran AD, Sobus JR, Williams AJ. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. *Anal Bioanal Chem*. 2017;409(7):1729–35. <https://doi.org/10.1007/s00216-016-0139-z>.
- Sobus JR, DeWoskin RS, Tan YM, Pleil JD, Phillips MB, George BJ, et al. Uses of NHANES biomarker data for chemical risk assessment: trends, challenges, and opportunities. *Environ Health*



- Perspect. 2015;123(10):919–27. <https://doi.org/10.1289/ehp.1409177>.
22. McEachran AD, Mansouri K, Newton SR, Beverly BEJ, Sobus JR, Williams AJ. A comparison of three liquid chromatography (LC) retention time prediction models. *Talanta*. 2018;182:371–9. <https://doi.org/10.1016/j.talanta.2018.01.022>.
  23. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform*. 2016;8:3. <https://doi.org/10.1186/s13321-016-0115-9>.
  24. Allen F, Pon A, Wilson M, Greiner R, Wishart D. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res*. 2014;42(Web Server issue):W94–9. <https://doi.org/10.1093/nar/gku436>.
  25. Allen F, Pon A, Greiner R, Wishart D. Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification. *Anal Chem*. 2016;88(15):7689–97. <https://doi.org/10.1021/acs.analchem.6b01622>.
  26. Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP, et al. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol*. 2014;48(4):2097–8. <https://doi.org/10.1021/es5002105>.
  27. Straub RF, Voyksner RD. Negative ion formation in electrospray mass spectrometry. *J Am Soc Mass Spectrom*. 1993;4(7):578–87. [https://doi.org/10.1016/1044-0305\(93\)85019-T](https://doi.org/10.1016/1044-0305(93)85019-T).