Samanipour, S, Reid, M J, Bæk, K, Thomas, K V. 2018. Combining a Deconvolution and a Universal Library Search Algorithm for the Nontarget Analysis of Data-Independent Acquisition Mode Liquid Chromatography−High-Resolution Mass Spectrometry Results. Environmental Science and Technology. 52 (8): 4694-4701.

Combining a deconvolution and a universal library search algorithm for the non-target analysis of data independent LC-HRMS spectra

Saer Samanipour,[*,†] Malcolm J. Reid,[†] Kine Bæk,[†] and Kevin V. Thomas[†,‡]

[†]*Norwegian Institute for Water Research (NIVA), 0349 Oslo, Norway*

[‡]*Queensland Alliance for Environmental Health Science (QAEHS), University of Queensland, 39 Kessels Road, Coopers Plains QLD 4108, Australia*

E-mail: saer.samanipour@niva.no

Abstract

Non-target analysis is considered one of the most comprehensive tools for identification of unknown compounds in a complex sample analyzed via liquid chromatography coupled to high resolution mass spectrometry (LC-HRMS). Due to the complexity of the data generated via LC-HRMS, the data dependent acquisition mode, which produces the MS^2 spectra of a limited number of the precursor ions, has been one of the most common approaches used during non-target screening. On the other hand, data independent acquisition mode produces highly complex spectra that require proper deconvolution and library search algorithms. We have developed a deconvolution algorithm and a universal library search algorithm (ULSA) for the analysis of complex spectra generated via data independent acquisition. These algorithms were validated and tested using both semi-synthetic and real environmental data. Six thousand randomly selected spectra from MassBank were introduced across the total ion chromatograms of 15 sludge extracts at three levels of background complexity for the validation of the algorithms via semi-synthetic data. The deconvolution algorithm successfully extracted more than 60% of the added ions in the analytical signal for 95% of processed spectra (i.e. 3 complexity levels × 6,000 spectra). The ULSA ranked the correct spectra among the top three for more than 95% of cases. We further tested the algorithms with five wastewater effluent extracts for 59 artificial unknown analytes (i.e. their presence or absence was confirmed via target analysis). These algorithms did not produce any cases of false identifications while correctly identifying $\sim$ 70% of the total inquiries. The implications, capabilities, and the limitations of both algorithms are further discussed.

INTRODUCTION

Little is known about the vast majority of the manmade substances released into the environment.[1–4] There are about 8,400,000 compounds commercially available globally.[1,2] Of these, the REACH Regulation has identified around 100,000 chemicals with an annual volume of production greater than one ton.[5] These chemicals may go through chemical transformation processes during their release into the environment, which drastically increases their number.[3,4] For example, a pharmaceutical such as carbamazepine potentially can produce five different metabolites once consumed by a human being (Human Metabolome Database HMDB[6]). Overall, less than 5% of these 100,000 chemicals (excluding transformation products) have been measured in the environment and less than 1% of them are included in monitoring programs and/or are regulated.[7] Environmental monitoring programs designed to measure these chemical footprints are primarily focused on a (relatively) small number of "known" chemicals. This is defined as "targeted analysis" or "analysis of suspects".[8] However, considering the number of chemicals released into the environment, the cost of standards and analysis, the target and suspect analysis approaches are not adequate for comprehensive monitoring of the environment. Furthermore, the application of non-target analysis using liquid chromatography coupled to high resolution mass spectrometry (LC-HRMS) has shown great potential in the comprehensive chemical characterization of complex samples.[8–12]

The data dependent acquisition (DDA) mode is one of the most commonly employed analysis methods during non-target screening of complex samples employing LC-HRMS.[8–14] In the DDA mode a selection of the detected precursor ions from the full scan MS^1 is fragmented using a high collision energy (i.e. MS^2 spectra). The main shortcoming of this method is the fact that the MS^2 spectra is only available for a limited number of precursor ions. Another less common approach used during the non-target analysis is the data independent acquisition (DIA) mode where all the precursor/parent ions generated at low collision energy are fragmented in the next cycle using a higher collision energy.[15] However, the DIA approach generates spectra, which are complex and difficult to process and moreover these spectra require adequate deconvolution algorithms[15–17] in order to be used during non-target screening. Most of the available deconvolution algorithms rely on peak picking in MS^1 domain[18,19] and are not adequate for handling MS^2 spectra generated during the DIA analysis.[15] Currently, to our knowledge, there are only two open access software for data processing of complex MS^2 spectra generated via DIA.[17,20] The first one, MS-DIAL, developed by Tsugawa et. al. performs peak picking in the MS^2 domain using the second derivative approach.[17] This method has been shown to have difficulties when processing highly complex samples with irregular peak shapes and peak widths.[18] The second software package, MetDIA by Li et. al., takes a metabolite focus approach.[20] In other words, the algorithm searches the whole chromatogram for all the MS^2 spectra present in the library. This approach avoids the peak picking difficulties in the MS^2 domain. However, it becomes extremely time consuming when dealing with a large spectral database, such as MassBank.[21] Therefore, development of a fast, efficient, and reliable algorithm for deconvolution of MS^2 spectra, which does not rely on peak picking is warranted.

Once the clean MS^2 spectrum of a precursor ion is generated, this spectrum is used to provide a tentative identification for that ion.[22–24] The application of public and/or local spectral libraries is one of the most common approaches used during non-target screening for the chemical identification.[24–29] However, difficulties persist due to the high level of instrument dependency of the MS^2 spectra, the limited number of publicly available spectra and the currently available library search algorithms.[24,25,30] Most of the library search algorithms in use are based on the highly reproducible electron ionization (EI) sources and/or a single match factor.[24,25,30,31] These algorithms have been shown to be inadequate in preforming reliable library search using the spectra generated via the less reproducible electrospray ionization source (ESI), hence the continuous development in this area.[24,25,30,32,33]

Herein we report the development and validation of a deconvolution algorithm and a universal library search algorithm (ULSA) for processing of the LC-HRMS data generated via DIA. Both algorithms are comprehensively validated and tested using both semi-synthetic data and real environmental data. In total 18,000 (i.e. 6,000 × 3) ESI+ randomly selected high resolution spectra from MassBank were used for the validation of the combination of these algorithms. Finally, this combination was used to identify 59 artificial unknown analytes in five wastewater effluent extracts employing a local version of MassBank[21,28] as the spectral library. Throughout this manuscript an artificial analyte refers to an anlyte, which has its presence or absence in the sample confirmed via conventional target analysis.

EXPERIMENTAL METHODS

Environmental Sampling and Sample Preparation

Fifteen biosolid samples were collected from three different wastewater treatment plants (five replicates for each treatment plant) in Norway during the spring of 2015. More details regarding these samples and the extraction procedure used for these samples are available elsewhere.[34] The chromatograms of these samples were used for the generation of the semi-synthetic signal, section S4.

One liter of wastewater effluent sample was collected from Aarhus Denmark, Helsinki Finland, Oslo Norway, and Stockholm Sweden in glass containers during September and October of 2015. We created a fifth sample by combining 200 mL of the four effluent samples, hereafter referred to as the mix sample. Two hundred and fifty mL of each sample were extracted using 200 mg Oasis HLB (Waters Milford, MA, US) solid phase extraction cartridges. After washing the cartridges with MilliQ water, the analytes were eluted with three cartridge volumes consisting of 1% formic acid in methanol, methanol, and methanol with 2% ammonium hydroxide. The final extracts of 500μL were reconstituted in methanol following evaporation under a gentile flow of nitrogen. All extracts were stored at -20 °C until analysis. The list of all the chemicals used and their suppliers is provided in the Supporting Information, section S1.

Instrumental Conditions and Analysis

All the samples were separated on an Acquity UPLC (Waters Milford, MA, US) using an Acquity BEH C18 column (100 × 2.1 mm, 1.7 μm) (Waters Milford, MA, US) with a methanol and water (10 mM ammonium acetate) mobile phase. Gradient elution was from 2% to 99% methanol over a 13 minute program. The UPLC system was connected to a high resolution mass spectrometer Xevo G2S QToF (Waters Milford, MA, US) operated in positive ESI mode.

The mass spectrometer was operated in full-scan between 50 Da and 850 Da with a sampling frequency of 2.7 Hz. The MS^1 spectra were acquired with a collision energy of 6 eV whereas the MS^2 spectra (MS^E experiments) were generated using a ramping collision energy between 15 eV and 45 eV. All of the chromatograms were acquired in the DIA mode with a nominal resolving power of 35,000. In other words we did not perform any ion selection during the MS^2 spectra generation.

Identification Criteria

We analyzed the five wastewater effluent extracts for 59 target analytes employing the UNIFI software (Waters Milford, MA, US). The following identification criteria were employed for the target analysis: presence of the accurate mass of parent ion, presence of at least two fragments; good isotopic fit defined as ≤ 5 ppm for the m/z match and $\leq 10\%$ root mean square error of the relative intensity; mass error smaller than 2 mDa for both the parent ion and the fragments; and finally a retention time match with the error smaller than 0.1 min. These criteria showed to be effective in the confident identification (i.e. level one[8]) of target analytes in complex environmental samples.[35]

The identification of the artificial unknown analytes (i.e. their presence or absence was confirmed via target analysis) was performed in the five wastewater effluent extracts using the combination of the deconvolution algorithm and ULSA. For a precursor ion to be identified, a positive match of the accurate mass of the precursor ion, positive match of at least three fragments, and a final score value of ≥ 3.5 was necessary. More details regarding the score calculations are provided in section S3 of the Supporting Information. These criteria enabled us to identify the evaluated precursor ions with the highest level of confidence (i.e. level 2a[8]). During our identification, we employed a local version of MassBank[21,28] as the spectral library.

The 59 artificial analytes consisted of 42 analytes with HRMS spectra available in MassBank whereas the remaining 17 did not have an HRMS spectrum available in MassBank, Table S1. This design of experiment enabled us to verify the tendency of the ULSA in producing false positive identifications for the cases without an HRMS spectrum in the library.

Data Processing

Both the sludge and wastewater effluent samples were acquired in profile mode using MassLynx (Waters Milford, MA, US). These chromatograms were converted to open format, netCDF, employing the DataBridge package included in the MassLynx software. These chromatograms were then imported into Matlab[36] for data processing. The raw data independently from its source went through the deconvolution algorithm first in order to produce a centroided MS^2 spectra and then those spectra were tentatively identified via USLA, Figure 1. The scripts for both deconvolution algorithm and the ULSA are openly available upon request. The chromatograms of the sludge extracts were used for the generation of semi-synthetic data while the chromatograms of wastewater effluent samples were used for the final test of the full workflow of deconvolution and identification via ULSA.

Deconvolution Algorithm

The developed deconvolution algorithm extracts the pure MS^2 spectra of an MS^1 precursor ion from the spectra generated in the high energy channel without performing peak picking in the MS^2 spectra, as explained in detail below and in Figure S1. Throughout this manuscript, we will refer to this feature dependent spectra as pseudo MS^2 spectra. The main inputs to this algorithm are the raw data in an open MS format, the mass-retention time pairs, the evaluation window, the maximum expected peak width in the time domain, the maximum expected peak width in mass domain, mass tolerance, retention time tolerance, minimum ion intensity, and finally the threshold for the correlation coefficient. The raw data goes through the following steps in order for the algorithm to extract the pure pseudo MS^2 spectra: mass calibration, binning, ion chromatogram extraction (XIC), retention matching, XIC correlation, and centroiding the pure pseudo MS^2 spectra. During the mass calibration the observed mass error of the calibrant, continuously infused into the source during the analysis, was used to calculate the necessary mass shift in each scan. After the calibration the mass error observed across the full scan in our dataset was $\leq \pm$ 5 mDa. The mass

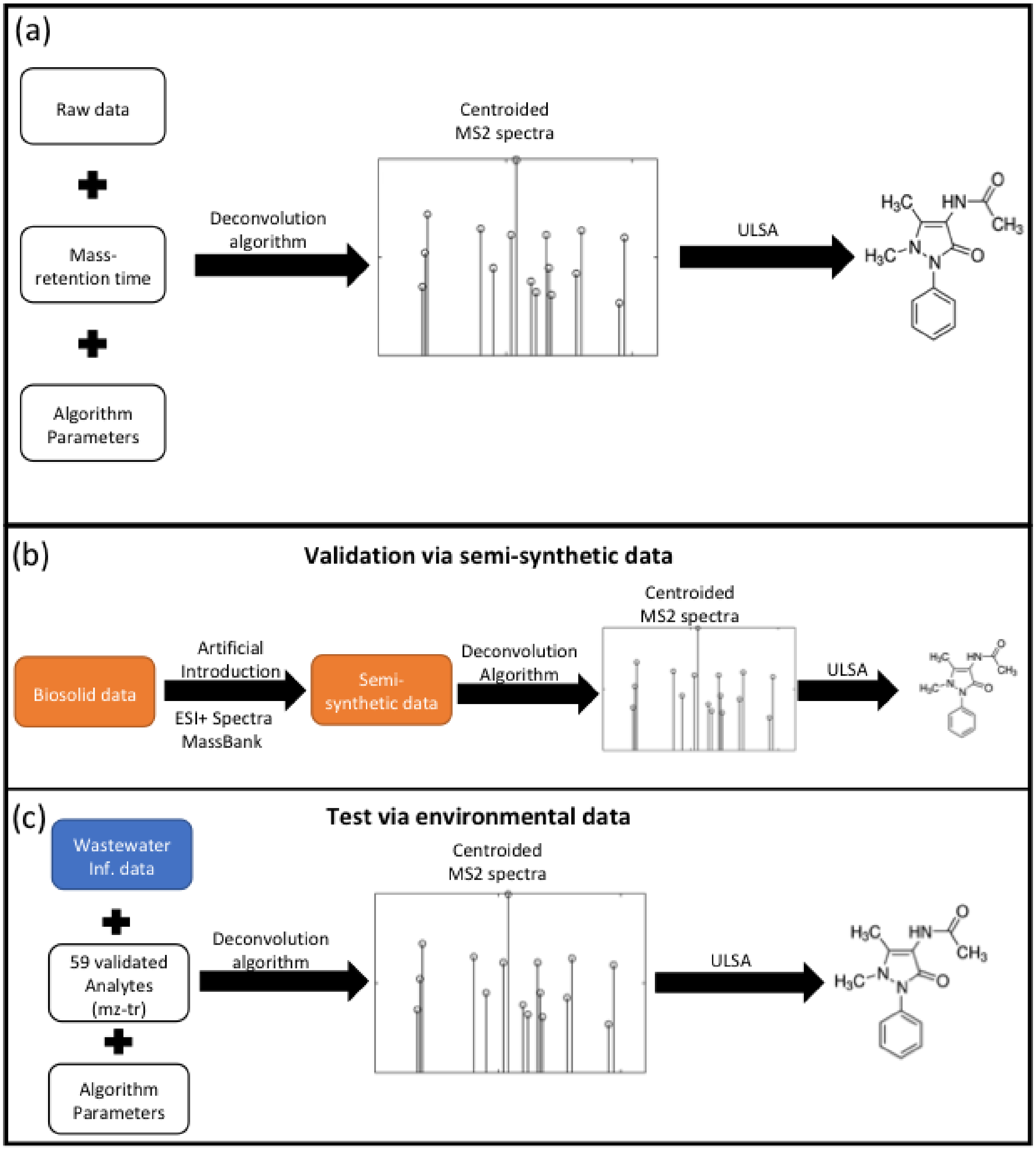


Figure 1: Showing the workflow of (a) the combination of deconvolution algorithm and ULSA, (b) the validation via semi-synthetic data, and (c) the final test using real environmental data. All three workflows depict the overall process from raw data to the final chemical identification.

calibrated date then went through the binning process, which employed a bin thickness of 10 mDa (i.e. ± 5 mDa), considering the observed mass accuracy in our dataset. An area of the binned chromatogram (i.e. for both MS^1 and MS^2 domains) around the retention time of the precursor ion with a width of two times the evaluation window plus one scan is isolated. In the next step the XIC of the precursor ion is extracted (or XIC^1), using the mass-retention time pair provided by the user. It should be noted that the mass-retention time pairs may come from different sources, for example conventional peak picking in the MS^1 domain, statistical variable selection,[34] and/or a suspect list, which enables the analysts to use this algorithm as a complementary tool to their own workflows. The Apex detection algorithm (explained in detail elsewhere[34]), at this point, is used to find the apex and the baseline of the peak for the precursor ion in the XIC^1. This process is repeated for each MS^2 ion with an intensity larger than the user defined minimum intensity, thus resulting in XIC^2 (i.e. XIC of the fragment ions in the MS^2 domain). At this stage, the algorithm uses two complementary criteria for inclusion of ions present in the MS^2. The first criterion is that the retention time of the apex for XIC^2s must match the retention time of XIC^1. Once the retention time criterion is met, then the profile of XIC^1 is correlated to each XIC^2. If the correlation coefficient for these two XICs is larger than a user defined threshold (i.e. in this study 0.9), then that XIC^2 is considered to be a true fragment of the initial precursor ion. Finally, during the last stage, the algorithm converts the previously generated pseudo MS^2 spectra (i.e. keeping only the MS^2 ions, which met the selection criteria) to a centroided spectra for storage and/or library search.

For both the semi-synthetic data and the wastewater effluent sample data, we used a bin thickness of 10 mDa, an evaluation window of 15 scans (i.e. 5.6 s), a maximum expected peak width of 30 scans (i.e. 11 s), mass tolerance of 10 mDa, retention tolerance of ± 1.2 s, minimum ion intensity of 800 counts, and a correlation coefficient threshold of 0.9. These parameters, which are dataset dependent, were optimized for our dataset and produced the best results for the evaluated dataset in this study. The mass-retention time pairs used for the 59 artificial analytes in wastewater effluent samples were implemented as suspect list.

Universal Library Search Algorithm (ULSA)

The pure pseudo MS^2 spectra via the developed deconvolution algorithm are annotated employing a universal library search algorithm (ULSA) for LC-HRMS. The ULSA produces a list of potential candidates with a final score associated to each candidate defining the similarity of that candidate to the user spectra (i.e. pure pseudo MS^2) through three main steps. In the first step, the ULSA takes advantage of the measured accurate mass of the precursor ion, a user defined error window (e.g. 50 mDa for our analysis) for the measured mass, and the list of possible adducts and isotopes to isolate the library entries (e.g. MassBank) that may be potential candidates. This wide mass error window was used to further test the ULSA capability for identifying the precursor ions. This algorithm, differently from the other available approaches, does not make any assumptions about the nature of precursor ion. In other words, for a certain measured precursor ion of A, the algorithm does not assume an $[M+H]^+$ structure. The algorithm first calculates the measured accurate mass of the potential neutral precursor ions from A, by removing the exact masses of all potential adducts and isotopes from the mass of that precursor ion (in the positive case). Then those accurate neutral masses are used for isolating the potential library entries relevant to that precursor ion. For example, if due to issues during the feature creation (i.e. grouping the precursor ion with the adducts and isotopes), the mass of 326.1363, which is the $[M+Na]^+$ structure for cocaine is considered as a potential precursor, this algorithm, differently from the others, does not assume the $[M+H]^+$ structure, which would cause a miss-identification of that precursor ion. This approach enables the identification of the measured precursor ions which are only present as an adduct or isotope with a structure different from $[M+H]^+$ and/or cases where there is a larger mass error than the expected values for the precursor ion. By increasing the mass error window, the number of potential candidates to be evaluated increases exponentially. It should be noted that the isolation step proved to be essential in order to process a large spectral library in a timely manner. During the second step, the ULSA calculates the score values for seven complementary parameters: the number of the matched fragments in the user spectra, the number of fragments matched in the library spectra, mass error of the precursor ion, the average mass error of the matched fragments in the user spectra, the standard deviation of the mass error for the matched fragments in the user spectra, and finally the direct and reverse similarity values calculated via Dot-product.[35,37] More detailed information regarding the score calculations for each parameter is provided in section S3, Supporting Information. It should be noted that fragment related parameters were scored taking into account the total number of fragments in the deconvoluted spectra and/or the reference spectra rather than only the matched fragments. This approach reduced the likelihood of generating large final scores based on only one or two matched fragments, section S3. A weighting function is applied to these seven scores and the results are summed up to create the final score for each potential candidate during the third step. The weighting function is a vector of seven elements, where each element can vary between zero and one, defining the weight of each of the seven parameters in the final score. In other words, if the weighting function is set to one for all seven parameters, a perfect match would result in a final score of seven while for an orthogonal candidate (i.e. a candidate with no similarity to the user spectra) the final score would be zero. Finally, the candidates are sorted based on their final scores with the most similar potential candidate to the user spectra on top of the list.

During our analysis we employed a 0.5 weight value for the parameters the number of the matched fragments in the user spectra and the number of fragments matched in the library spectra while using a weight value of 1 for other five parameters. This implied that the final score for these analysis can vary between 0 for orthogonal spectra and 6 for maximum similarity (i.e. a perfect match).

It should be noted that the deconvolution algorithm and ULSA are completely independent from each other and can be operated individually without relying on the other algorithm. In other words, the deconvoluted spectra can be identified using any other library search algorithm and vice versa.

Computations

All the calculations and data analysis were performed employing Matlab R2015b[36] with a Windows 7 Professional version (Microsoft Inc., USA) workstation computer with 12 CPUs and 128 GB of memory.

RESULTS AND DISCUSSION

The deconvolution algorithm and the ULSA were validated and tested employing semi-synthetic data as well as real environmental data. We utilized 6,000 randomly selected LC-HRMS spectra in positive mode from MassBank for the validation of both deconvolution and library search algorithms at three different levels of background complexity or noise. Finally, five samples of wastewater effluents were analyzed for 59 analytes via both developed algorithms and the conventional target analysis. This final test demonstrated the applicability of the developed algorithms for the feature identification during the suspect and non-target analysis of complex environmental samples.

Validation and test of the deconvolution algorithm

We artificially introduced the signal of 6,000 randomly ESI+ selected LC-HRMS spectra from MassBank, here referred to as the analytical signal, into three different complexity level background signal or noise coming from real environmental samples (i.e. 15 sludge samples). The analytical signal was converted to profile data having m/z peak width of

30 mDa whereas the peak width in the retention dimension was 5 scans (i.e. around 2 S). This continuum analytical signal was added at a random location in a predefined area of the sludge chromatograms at an intensity equivalent of 10% of the highest intensity ion in the background signal. The relative ratios of the ion intensities in the analytical signal were kept as the MassBank entry. This experimental design enabled us to identify the fragments correctly extracted (i.e. true positive ions (TPI)), the fragments which were missed (i.e. false negative ions (FNI)), and the fragments that were wrongly extracted (i.e. false positive ions (FPI)) for the total of 18,000 cases. The detailed procedure for generation of the semi-synthetic dataset is provided in the Supporting Information, section S4.

The deconvolution algorithm was able to successfully extract 100% of introduced ions for $\geq$ 60% of the processed spectra at both low and medium noise levels whereas for the high noise levels this was limited to $\simeq$ 35% of the processed spectra, Figure 2. For all three noise levels this algorithm produced less than 0.01% of FPIs. The small number of cases of the FPIs were caused by the complexity of the background signal, Figure S2. Minimizing the number of FPIs is essential in order to lower the likelihood of the false identification of a feature. At low and medium background complexity levels the deconvolution algorithm performed in a similar way producing a small number FNIs when compared to the high background complexity. For the cases of FNIs, more than 92% of the cases were caused by the fact that added signal of these fragments were smaller than the predefined minimum threshold of intensity (i.e. 800 counts), Figures S3 and S2. The remaining 8% of FNIs were caused by the complexity of the background signal which was translated into an irregular peak shape for the XICs, Figure S4. Thus, the XIC of these fragments once correlated to the XIC of the precursor ion resulted in a correlation coefficient smaller than the set threshold (i.e. 0.9) and therefore they were excluded from the list of potential fragments of that precursor ion. The developed deconvolution algorithm was shown to be capable of successfully extracting the correct fragments of a precursor ion even with the highest level of background signal complexity. For all three levels of background complexity, the algorithm produced a negligible number of FPIs even though the artificially introduced analytical signal was at an environmentally relevant concentration level in the samples. Furthermore, our results demonstrated the capabilities of the developed deconvolution algorithm to be applied to DIA for non-target and suspect analysis of complex environmental samples.

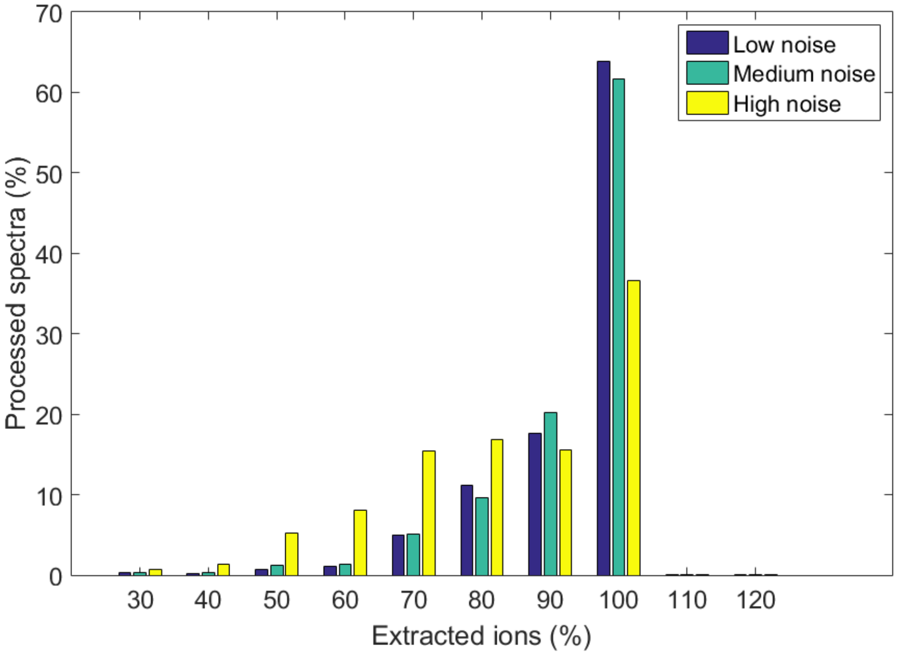


Figure 2: Depicting the percentage of extracted spectra vs the percentage of total number of processed spectra (i.e. 6000 × 3 spectra).

The validation of ULSA

All of the 3 × 6,000 extracted spectra generated by the deconvolution algorithm were processed using ULSA and a local version of MassBank. The ULSA produced a list of potential candidates ranking them from the the most similar (i.e. the highest final score) to the least similar one. During the identification process, each individual library entry was considered as an entirely different compound. This implied that there was only one true match for each spectrum, even if there were multiple spectra for that compound (e.g. morphine with 18 entries in MassBank). For example, if the third entry for morphine was originally added to the background signal, we only accepted that specific entry as a correct identification for that library inquiry even though all the other listed potential candidates belonged to morphine. This approach enabled us to truly evaluate the capabilities and limitations of ULSA in distinguishing similar spectra (i.e. spectra for the same compound recorded under different condition) from each other.

The ULSA successfully ranked the correct spectra among the top three hits for more than 95% of the identified spectra, Figure 3. We observed similar results for all three levels of background complexity, even though at higher levels of complexity a smaller number of fragments were extracted, Figure 2. The variation in the background signal complexity did not appear to effect the ULSA in a statistically meaningful way. Therefore we observed similar results for all three levels of background complexity. There were in total 23 cases out of 18,000 where the correct spectra was ranked higher than fifth in the final hit list of the ULSA. These cases were all caused by the presence of multiple entries which were extremely similar to each other. Therefore, the ULSA had some difficulties in distinguishing one from the other. In fact for all the mentioned cases, the relative standard deviation in the final scores is $< 5\%$, which further indicates the similarity of those spectra. When looking at the distribution of the final score, for 95% of cases we observed a final score varying between 5.25 and 6 for all three levels of background complexity. The complexity level in the background signal resulted in an increase in the number of identified cases with smaller final scores when compared to the low and medium levels of complexity in the background signal. However, our results indicated that the ULSA is able to correctly annotate a spectrum even at high levels of noise/background complexity.

The developed ULSA was shown to be successful in correctly annotating the LC-HRMS spectra. This algorithm utilizes the combination of forward and reverse match factors calculated by minimizing the effect of the absolute intensity of the fragments in the spectra through the application of an optimized spectral weighting function; the number of matched fragments; mass errors for both the precursor and fragment ions; and the standard deviation of the fragment mass error to produce a reliable final score. This approach proved to be crucial in distinguishing similar compounds from each other. For example, when identifying 1-methylbenzotriazole, the spectra of 2-aminobenzimidazole showed to have a higher forward and reverse match factors compared to the correct library entry (i.e. 1-methylbenzotriazole). However, the additional parameters used in ULSA differently from other library search algorithms, increased the final score of the correct library entry. Additionally, the final hit lists produced via ULSA showed that the spectra of the same compound measured under different conditions (i.e. instrumentation and acquisition conditions) ranked higher than the spectra of different compounds, which can be considered a step forward towards the cross-platform compatibility for LC-HRMS data. However, a comparison of ULSA and other available algorithms should be done in order to further assess the cross-platform compatibility.

We also evaluated the effect of each of those parameters on the final score in ULSA. Five out of the seven parameters in the final score values produced an average score of $\sim$0.6 (i.e. from 0 to 1) whereas the two remaining resulted in an average score of $\sim$0.95 (i.e. from 0 to 1) for 100 randomly selected spectra at all three levels of noise, Figure S5. This outcome suggested that these two parameters (i.e. the number of the matched fragments in the user spectra and the number of fragments matched in the library spectra) appeared to have a higher contribution in the final scores compared to the other five parameters. Therefore, the 0.5 weight applied to these two parameters seemed appropriate when employing ULSA. In other words, by applying this weight function all seven parameters showed to have a similar effect on the final scores.

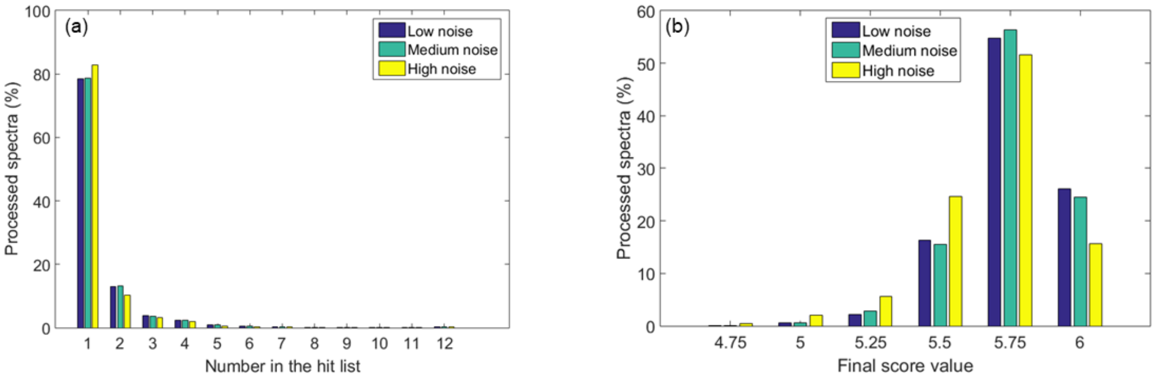


Figure 3: Depicting (a) the rank distribution of correctly identified spectra via ULSA and (b) the final score distribution for those identifications.

Application of the deconvolution algorithm and ULSA for analysis of wastewater effluent extracts

In addition to the validation of our algorithms using the semi-synthetic data we also tested the performance of both the deconvolution algorithm and the ULSA employing extracts of five wastewater effluents. We analyzed these five samples for 59 artificial unknown analytes (thus, 5 samples × 59 analytes = 295 cases) where we confirmed their presence or absence in those samples via conventional target screening. These 295 detection cases consisted of: 234 true positives (TPs) including 152 cases of positive detection with at least one high resolution (HR) spectrum entry in the library and 82 cases of positive detections with no HR spectrum entry in the library; and 61 cases of true negatives (TNs). A TP was an analyte where its presence in a sample was confirmed via target analysis whereas a TN was an analyte which had its absence confirmed via target analysis. The TPs with an HR library entry were used for both false positive and false negative identifications. On the other hand, the TPs without an HR library spectrum were specifically used to evaluate the tendency of the ULSA in falsely identify a feature even though in theory it should not have produced that identification, thus a false positive. The TNs were also used for evaluation of false positive detections. In other words, if an identification was produced for a TN, that was considered a false positive identification. This design of experiment covered all potential situations when dealing with complex environmental samples, which were: 1) An analytical signal with a related library entry (i.e. a TP with library entry); 2) An analytical signal, which does not have any HRMS entries in the library (i.e. a TP without library entry); and 3) Noise, which has been wrongly considered as a meaningful analytical signal (i.e. an NP with library entry). Therefore we were able comprehensively evaluate the capabilities and limitations of both developed algorithms.

The combination of the deconvolution algorithm and ULSA did not produce any cases of false positive identifications based on the artificial analytes. This implied that this combination of the algorithms did not produce a false identification for any of TPs with and without library entries and NPs. These algorithms, on the other hand produced 48 cases of false negative detections out of 295 detection cases. These false negative detections were caused by the low levels of these analytes in the analyzed samples and the complexity of the samples, which was directly translated into irregular peak shapes for both the fragments and precursor ions, Figure S6. Therefore, the deconvolution algorithm was not able to extract the clean spectra for these analytes and therefore these analytes were not identified. The number of fragments extracted for the successfully identified analytes varied between 3 for cocaine to 14 for amitriptyline. The number of extracted fragments for these analytes in the samples appeared to be lower than our evaluation with the semi-synthetic data. This was mainly due to the ion suppression which was caused by the complexity of the samples. We further evaluated this hypothesis by the manual inspection of the feature spectra and their comparison to the MassBank entries. The smaller number of extracted fragments showed to have a direct effect on the final score values. The final scores for the identified analytes in the effluent samples varied between 3.5 to 4.8. This decrease in the final scores was caused by the fact that the score for each fragment related parameter was adjusted for the total number of fragments either in the user spectra of the library spectra. For example, for a user spectrum with 10 fragments where only 2 out of 10 were matched a smaller final score was produced when compared to another case with 2 out of 5 extracted fragments matched. Additionally, the use of the seven complementary parameters enabled a balanced comparison between different candidates. For a certain feature in the sample from Norway for example, two different library candidates were observed, cocaine and fenoterol. The deconvolution algorithm extracted 3 fragments for that feature from the raw data. By only looking at the forward and reverse match factors or any of the seven parameters individually, we would not have been able to identify these features with a high level of confidence (i.e. level 2a). However, the combination (i.e. the summation) of these seven complementary parameters caused a final score difference of 2, which is large enough for excluding fenoterol as a potential chemical identity for that feature. This approach enabled the ULSA to successfully identify 104 analytes out of 152 TPs with library entries even with such a low number of extracted fragments.

Overall, the combination of the deconvolution algorithm and ULSA was shown to be effective in identifying/annotating the retention time m/z value pairs using a public library such as MassBank. This approach also demonstrated the usefulness and applicability of data independent acquisition mode as well as the public spectral libraries for non-target and suspect analysis of complex environmental samples. Despite the fact that none of the entries in the library used (i.e. MassBank) was produced by the instrumentation employed in this study, the developed method successfully identified around $\sim 70\%$ of the total library inquiries without producing any cases of false positive detections. The proposed approach minimizes the spectral differences caused by different instrumentations and acquisition conditions thus increasing the cross platform compatibility. Consequently, this approach adds to the value of the public HRMS spectral libraries such as MassBank by increasing the applicability of spectra produced via different instruments, thus cross platform compatibility. These two algorithms can be included in any type of non-target and/or suspect screening workflows for the comprehensive chemical characterization of complex environmental samples, which will be subject of our future studies.

Associated Content

Acknowledgement

The authors are grateful to the Research Council of Norway for the financial support of this project (RESOLVE, 243720). We are also grateful to the Nordic Chemicals Group (non-target screening, 2249) for their support. Saer Samanipour was partially financed by NIVA project PLATFORM (180014).

Supporting Information

The Supporting Information including details regarding the semi-synthetic data generation and score calculations is available free of charge on the ACS Publications website.

Author Information

Corresponding Author:
Saer Samanipour
E-mail: saer.samanipour@niva.no
Phone: +47 98 222 087
Address: Norwegian Institute for Water Research (NIVA)
0349 Oslo, Norway

Malcolm J. Reid
Email: malcolm.reid@niva.no
Address: Norwegian Institute for Water Research (NIVA)

0349 Oslo, Norway

Kine Bæk
Email: kine.baek@niva.no
Address: Norwegian Institute for Water Research (NIVA)
0349 Oslo, Norway

Kevin V. Thomas
Email: kevin.thomas@uq.edu.au
Address: Queensland Alliance for Environmental Health Science (QAEHS), University of Queensland, 39 Kessels Road, Coopers Plains QLD 4108, Australia

References

(1) Muir, D. C.; Howard, P. H. Are there other persistent organic pollutants? A challenge for environmental chemists. *Environmen. Sci. Technol.* **2006**, *40*, 7157–7166.

(2) Howard, P. H.; Muir, D. C. Identifying new persistent and bioaccumulative organics among chemicals in commerce. *Environmen. Sci. Technol.* **2010**, *44*, 2277–2285.

(3) Howard, P. H.; Muir, D. C. Identifying new persistent and bioaccumulative organics among chemicals in commerce II: pharmaceuticals. *Environmen. Sci. Technol.* **2011**, *45*, 6938–6946.

(4) Howard, P. H.; Muir, D. C. Identifying new persistent and bioaccumulative organics among chemicals in commerce. III: Byproducts, impurities, and transformation products. *Environmen. Sci. Technol.* **2013**, *47*, 5259–5266.

(5) Williams, E. S.; Panko, J.; Paustenbach, D. J. The European Union's REACH regulation: a review of its history and requirements. *Crit. Rev. Toxicol.* **2009**, *39*, 553–575.

(6) Wishart, David S and Tzur, Dan and Knox, Craig and Eisner, Roman and Guo, An Chi and Young, Nelson and Cheng, Dean and Jewell, Kevin and Arndt, David and Sawhney, Summit and others, HMDB: the human metabolome database. *Nucleic Acids Res.* **2007**, *35*, D521–D526.

(7) Andra, S. S.; 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.

(8) Schymanski, Emma L and Singer, Heinz P and Slobodnik, Jaroslav and Ipolyi, Ildiko M and Oswald, Peter and Krauss, Martin and Schulze, Tobias and Haglund, Peter and Letzel, Thomas and Grosse, Sylvia and others, Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. *Anal. Bioanal. Chem.* **2015**, *407*, 6237–6255.

(9) Schymanski, E. L.; Singer, H. P.; Longrée, P.; Loos, M.; Ruff, M.; Stravs, M. A.; Ripollés Vidal, C.; Hollender, J. Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. *Environ. Sci. Technol.* **2014**, *48*, 1811–1818.

(10) Gago-Ferrero, P.; Schymanski, E. L.; Bletsou, A. A.; Aalizadeh, R.; Hollender, J.; Thomaidis, N. S. Extended suspect and non-target strategies to characterize emerging polar organic contaminants in raw wastewater with LC-HRMS/MS. *Environ. Sci. Technol.* **2015**, *49*, 12333–12341.

(11) Aceña, J.; Stampachiacchiere, S.; Pérez, S.; Barceló, D. Advances in liquid chromatography–high-resolution mass spectrometry for quantitative and qualitative environmental analysis. *Anal. Bioanal. Chem.* **2015**, *407*, 6289–6299.

(12) Kind, T.; Fiehn, O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal. Rev.* **2010**, *2*, 23–60.

(13) Krauss, M.; Singer, H.; Hollender, J. LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns. *Anal. Bioanal. Chem.* **2010**, *397*, 943–951.

(14) Chiaia-Hernandez, A. C.; Schymanski, E. L.; Kumar, P.; Singer, H. P.; Hollender, J. Suspect and nontarget screening approaches to identify organic contaminant records in lake sediments. *Anal. Bioanal. Chem.* **2014**, *406*, 7323–7335.

(15) Arnhard, K.; Gottschall, A.; Pitterl, F.; Oberacher, H. Applying 'Sequential Windowed Acquisition of All Theoretical Fragment Ion Mass Spectra'(SWATH) for systematic toxicological analysis with liquid chromatography-high-resolution tandem mass spectrometry. *Anal. Bioanal. Chem.* **2015**, *407*, 405–414.

(16) Li, G.-Z.; Vissers, J. P.; Silva, J. C.; Golick, D.; Gorenstein, M. V.; Geromanos, S. J. Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. *Proteomics* **2009**, *9*, 1696–1719.

(17) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature methods* **2015**, *12*, 523–526.

(18) Tautenhahn, R.; Böttcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics* **2008**, *9*, 504.

(19) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, *78*, 779–787.

(20) Li, H.; Cai, Y.; Guo, Y.; Chen, F.; Zhu, Z.-J. MetDIA: Targeted Metabolite Extraction of Multiplexed MS/MS Spectra Generated by Data-Independent Acquisition. *Anal. Chem.* **2016**, *88*, 8757–8764.

(21) Schulze, Tobias and Schymanski, E and Stravs, M and Neumann, S and Krauss, M and Singer, H and others, NORMAN MassBank. *Towards a community-driven, open-access accurate mass spectral database for the identification of emerging pollutants. NORMAN Network Bulletin* **2012**, *3*, 9–10.

(22) Yi, L.; Dong, N.; Yun, Y.; Deng, B.; Ren, D.; Liu, S.; Liang, Y. Chemometric methods in data processing of mass spectrometry-based metabolomics: A review. *Anal. Chem. acta* **2016**, *914*, 17–34.

(23) Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. *Trends Anal. Chem.* **2016**, *82*, 425–442.

(24) Oberacher, H.; Arnhard, K. Current status of non-targeted liquid chromatography-tandem mass spectrometry in forensic toxicology. *TrAC Trends Anal. Chem.* **2016**, *84*, 94–105.

(25) Pavlic, M.; Libiseller, K.; Oberacher, H. Combined use of ESI–QqTOF-MS and ESI–QqTOF-MS/MS with mass-spectral library search for qualitative analysis of drugs. *Anal. Bioanal. Chem.* **2006**, *386*, 69–82.

(26) Hernández, F.; Sancho, J.; Ibáñez, M.; Abad, E.; Portolés, T.; Mattioli, L. Current use of high-resolution mass spectrometry in the environmental sciences. *Anal. Bioanal. Chem.* **2012**, *403*, 1251–1264.

(27) Katajamaa, M.; Miettinen, J.; Orešič, M. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **2006**, *22*, 634–636.

(28) Horai, Hisayuki and Arita, Masanori and Kanaya, Shigehiko and Nihei, Yoshito and Ikeda, Tasuku and Suwa, Kazuhiro and Ojima, Yuya and Tanaka, Kenichi and Tanaka, Satoshi and Aoshima, Ken and others, MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, *45*, 703–714.

(29) Mistrik, R.; Lutisan, J.; Huang, Y.; Suchy, M.; Wang, J.; Raab, M. mzCloud: A Key Conceptual Shift to Understand ‘Who’s Who’in Untargeted Metabolomics. Metabolomics Society 2013 Conference, Glasgow, July. 2013; pp 1–4.

(30) Oberacher, H.; Pavlic, M.; Libiseller, K.; Schubert, B.; Sulyok, M.; Schuhmacher, R.; Csaszar, E.; Köfeler, H. C. On the inter-instrument and the inter-laboratory transferability of a tandem mass spectral reference library: 2. Optimization and characterization of the search algorithm. *J. Mass Spectrom.* **2009**, *44*, 494–502.

(31) Samokhin, A.; Sotnezova, K.; Lashin, V.; Revelsky, I. Evaluation of mass spectral library search algorithms implemented in commercial software. *J. Mass Spectrom.* **2015**, *50*, 820–825.

(32) Lam, H. Building and searching tandem mass spectral libraries for peptide identification. *Mol. Cell. Proteomics* **2011**, *10*, R111–008565.

(33) Huan, T.; Tang, C.; Li, R.; Shi, Y.; Lin, G.; Li, L. MyCompoundID MS/MS Search: Metabolite identification using a library of predicted fragment-ion-spectra of 383,830 possible human metabolites. *Anal. Chem.* **2015**, *87*, 10619–10626.

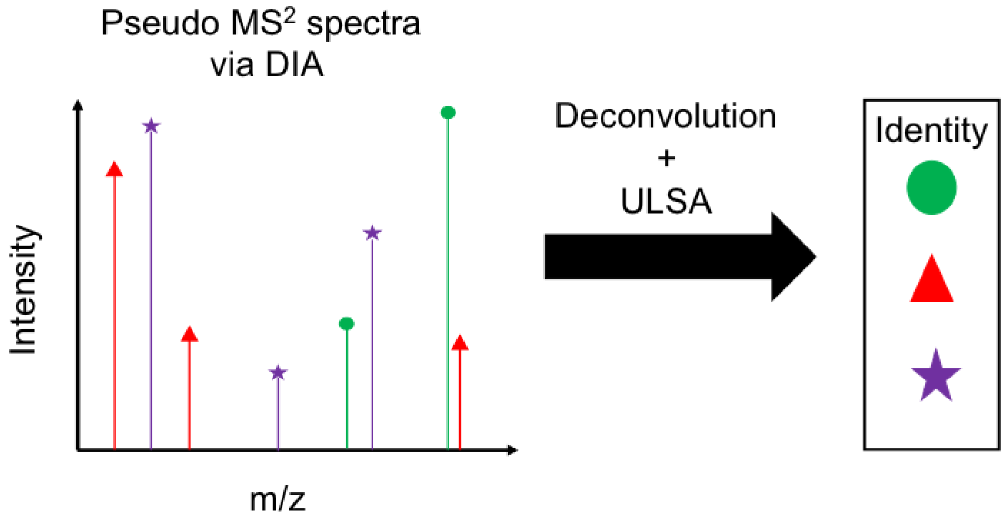
(34) Samanipour, S.; Reid, M. J.; Thomas, K. V. Statistical variable selection: An alternative prioritization strategy during the non-target analysis of LC-HR-MS data. *Anal. Chem.* **2017**, *89 (10)*, 5585–5591.

(35) Samanipour, S.; Baz-Lomba, J. A.; Alygizakis, N. A.; Reid, M. J.; Thomaidis, N. S.; Thomas, K. V. Two stage algorithm vs commonly used approaches for the suspect

screening of complex environmental samples analyzed via liquid chromatography high resolution time of flight mass spectroscopy: A test study. *J. Chromatogr. A* **2017**, *1501 (2017)*, 68–78.

(36) MATLAB version 9.1 Natick, Massachusetts: The MathWorks Inc., **2016**.

(37) Samanipour, S.; Langford, K.; Reid, M. J.; Thomas, K. V. A two stage algorithm for target and suspect analysis of produced water via gas chromatography coupled with high resolution time of flight mass spectrometry. *J. Chromatogra. A* **2016**, *1463*, 153–161.



TOC only for review.