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

Saer Samanipour[a,*], Jose A. Baz-Lomba[a], Nikiforos A. Alygizakis[b], Malcolm J. Reid[a], Nikolaos S. Thomaidis[b], Kevin V. Thomas[a,c]

[a] *Norwegian Institute for Water Research (NIVA), 0349 Oslo, Norway*
[b] *Laboratory of Analytical Chemistry, Department of Chemistry, University of Athens, Panepistimiopolis Zographou, 15771 Athens, Greece*
[c] *Queensland Alliance for Environmental Health Science (QAEHS), University of Queensland, 39 Kessels Road, Coopers Plains QLD 4108, Australia*

Abstract

LC-HR-QTOF-MS recently has become a commonly used approach for the analysis of complex samples. However, identification of small organic molecules in complex samples with the highest level of confidence is a challenging task. Here we report on the implementation of a two stage algorithm for LC-HR-QTOF-MS datasets. We compared the performances of the two stage algorithm, implemented via NIVA_MZ_Analyzer™, with two commonly used approaches (i.e. feature detection and XIC peak picking, implemented via UNIFI by Waters and TASQ by Bruker, respectively) for the suspect analysis of four influent wastewater samples. We first evaluated the cross platform compatibility of LC-HR-QTOF-MS datasets generated

*Saer Samanipour
Email address: saer.samanipour@niva.no (Saer Samanipour)
[1]NIVA, Gaustadalléen 21, 0349 Oslo, Norway
Tel: +47 98222087

via instruments from two different manufacturers (i.e. Waters and Bruker). Our data showed that with an appropriate spectral weighting function the spectra recorded by the two tested instruments are comparable for our analytes. As a consequence, we were able to perform full spectral comparison between the data generated via the two studied instruments. Four extracts of wastewater influent were analyzed for 89 analytes, thus 356 detection cases. The analytes were divided into 158 detection cases of artificial suspect analytes (i.e. verified by target analysis) and 198 true suspects. The two stage algorithm resulted in a zero rate of false positive detection, based on the artificial suspect analytes while producing a rate of false negative detection of 0.12. For the conventional approaches, the rates of false positive detection varied between 0.06 for UNIFI and 0.15 for TASQ. The rates of false negative detection for these methods ranged between 0.07 for TASQ and 0.09 for UNIFI. The effect of background signal complexity on the two stage algorithm was evaluated through the generation of a synthetic signal. We further discuss the boundaries of applicability of the two stage algorithm.The importance of background knowledge and experience in evaluating the reliability of results during the suspect screening was evaluated.

Keywords:

Two stage algorithm, LC-HR-QTOF-MS, Target and suspect screening, False detection, Confident identification

---

1. Introduction

The suspect and non-target analysis of complex environmental samples for small organic molecules is a challenging task [1, 2]. For the samples that are selected *a priori* for suspect and non-target analysis, the analyst performs a generic sample

preparation and uses a wide range instrumental analysis [2–6]. A list of suspects is then compiled including some chemical specific information. This information is used for the confident identification of suspect analytes while processing those samples [1–6]. For non-target analysis, the analyst after performing generic sample preparation and instrumental analysis investigates the data without having any prior information about the compounds of interest [2]. Recent advancements in high resolution mass spectrometers coupled to liquid and/or gas chromatography (LC-HR-MS and/or GC-HR-MS) has enabled the analysts to perform these type of analyses faster and with high levels of confidence in the identification, having access to the exact mass information [1, 7]. However, due to the large amount of data produced by these new technologies a confident identification during the suspect and non-target analysis of complex samples remains a difficult and time consuming task.

Several approaches are available for the suspect analysis of complex samples analyzed via LC-HR-MS. However, the majority of these data processing algorithms employ two different strategies in suspect analysis when dealing with the data produced via LC-HR-MS [1, 8]. The first strategy is based on peak picking in the extracted ion chromatograms (XICs) of the molecular ion and the relevant qualifier ions in the sample chromatogram. This is performed by first extracting the XICs and then performing peak picking on those XICs. That way the retention time of a peak is first defined and then matched with other fragment XICs (i.e. qualifiers). In case of a positive match between the XIC retention time of the parent ion and the fragments, those fragments are included in the spectra of that peak. The parent

ion and the included fragments, based on their retention time match, are used as diagnostic tools during the identification. Confidence in the identification then is improved by adding another orthogonal parameter to the parent ion and the fragments match. These parameters include either the XIC peak shape (employed by ChromaLynx, Waters) or isotopic fit (utilized by TASQ, Bruker). In other words, the algorithms using the XIC strategy, employ XIC retention time match and either isotopic fit or XIC peak shape as two orthogonal parameters in order to identify a suspect peak in the sample. The second data processing strategy consists of first performing peak picking or feature detection in the MS^1 (or first function having a collision energy varying from 4 eV to 6 eV) independently from the target or suspect list (e.g. UNIFI by Waters, MZmine [9], XCMS [10], and OpenMS [11]). Afterwards these algorithms group all the ions related to a specific unique feature. This operation includes grouping the fragments, isotopes, and adducts. Once the unique feature detection is completed, these algorithms use the information provided by the analyst, such as the exact masses of molecular ions and potential fragments to assign an identity to a certain unique component. In the last stage, the analyst needs to create filters based on isotopic fit, theoretical fragmentation pattern, mass accuracy of the parent/fragment ions, and/or retention time match of the standard and the sample peaks to prioritize amongst the potential candidates. The list of potential candidates then are processed using carefully designed filters in order to remove false positives, thus increasing the level of confidence in the final identification. This post processing step is necessary, independently from the detection method used (i.e. peak picking or XICs).

The commonly used data processing approaches for suspect and non-target analysis are not always as effective as needed, particularly for highly complex samples. One of the frequently observed issues with these algorithms is peak picking or feature detection (as reviewed by Zhang et al.[12] and [13]). Also it has been shown that even parameter optimization [14] may not result in a perfect outcome for all the peaks, due to the sample complexity [14, 15]. Therefore, considering that the majority of algorithms have peak picking as an essential step during data processing, they may be affected by the probability of failure during peak picking due to the high levels of noise and broad peaks. Independently from peak picking, another commonly observed issue is caused by mass calibration [16]. Improper mass calibration can create difficulties during the isotopic fit and/or the grouping process. These potential artifacts caused by the complexity of the samples may translate into the erroneous identification of a suspect peak. Overall, these algorithms require careful tuning and post processing in order to produce identifications with a high levels of confidence.

In our recent study, we developed an alternative algorithm (i.e. two stage algorithm, for detail information please see section 2.4.1) for target and suspect screening in complex samples analyzed via high resolution time of flight mass spectrometer (HR-TOF-MS) [17]. The two stage algorithm does not perform peak picking or peak deconvolution in the sample chromatogram, which reduces its probability of failure caused by the peak picking algorithm. This algorithm takes advantage of a full spec-

tral comparison between the spectra of the standard stored in a local or a public spectral library and the sample. The two stage algorithm calculates the similarity between the standard spectrum and sample spectrum at each scan number of the sample chromatogram. If a suspect analyte is present in the sample then a peak of similarity is observed in the sample. The two stage algorithm was modified in order to process the data produced via LC-HR-QTOF-MS. This algorithm is shown to be effective in the target and suspect analysis of complex samples with the highest levels of confidence and minimum post processing efforts.

Herein we report the comparison between three different data processing strategies for the suspect analysis of 89 analytes, including pharmaceuticals and illicit drugs, in four wastewater extracts (i.e. 4×89 analytes, 356 detection cases). The investigated strategies comprised of: the two stage algorithm implemented via NIVA_MZ_Analyzer^{TM}, UNIFI by Waters, and TASQ by Bruker. One hundred and fifty eight detection cases (i.e. 2×33 Oslo analytes + 7 common analytes + 2×32 Athens analyes + 7 common analytes) out of 356 total cases were artificial suspects whereas the remaining 198 detection cases were analyzed only in suspect mode. In other words, the presence and/or absence of an artificial suspect analyte in the samples was confirmed through target analysis. In this study, UNIFI was considered a representative of the methods, that perform peak picking on the whole m/z domain whereas TASQ was used as an example for peak picking on XICs. On the other hand, NIVA_MZ_Analyzer^{TM}, which does not rely on peak picking, was considered as an alternative approach. We used influent wastewater extracts as our testing matrix

due to their complexity, high level of noise, and background signal. In this study we evaluated the problems that analysts face during this type of analysis and the way that each algorithm tries to overcome these issues. The cross platform compatibility of the data produced via LC-HR-QTOF-MS as well as the applicability boundaries of the two stage algorithm were also discussed.

2. Experimental

2.1. Chemicals and sample collection

For our assessment we used a suite of 89 analytes consisting of pharmaceuticals (including antibiotics) and illicit drugs. These chemicals were selected based on their importance and high frequency of detection in the wastewater samples from both Oslo [18–20] and Athens [5, 21]. These chemicals were divided in Oslo analytes (33 out of 89), the Athens analytes (32 out of 89), the library analytes (17 out of 89), and the common analytes (7 out of 89). A complete list of the suspect/target analytes is included in the Supporting Information, section S1.

We examined the performances of the three selected approaches on four extracts of wastewater influent. Two out of the four samples were collected from the VEAS treatment plant in Oslo. Throughout this article we will refer to these samples as "Oslo samples". The other two samples of influent were from Athens wastewater treatment plans (hereafter referred to as "Athens samples"). These four samples were 1L flow proportional 24 h composite samples of the two studied wastewater treatment plants. These samples were part of a large sampling campaign conducted

during the spring of 2015, where we collected one sample every day. We randomly selected four samples out of the pool of samples for evaluation of data processing approaches. These four samples appeared to be representative of the pool of samples collected during the sampling campaign, based on their chromatogram similarity (TIC) to the other samples collected during the sampling campaign.

The Oslo plus common analytes, 40 compounds, and the Athens plus common analytes, 39 chemicals, were analyzed in both target and suspect analysis mode in Oslo samples and Athens samples, respectively. The Oslo analytes plus the library analytes were considered as true suspects in Athens samples whereas the Athens analytes and the library analytes were analyzed as pure suspects in Oslo samples. This experimental design enabled us to fully evaluated the capabilities and limitations of all three approaches for both suspect and retrospective analysis.

2.2. Sample preparation

The Oslo samples were extracted using Oasis HLB sorbent [18] whereas the Athens samples were extracted using a modified version [5] of the method developed by Kern et al. 2009 [22].

We considered these samples adequate, based on their complexity (as reviewed by Krauss et al. [23]), high level of background signal, large number of unique features, and finally the diversity of the compounds present in this type of sample [5, 18, 21, 24]. These characteristics of the samples enabled us to test the capabilities and also the limitations of the tested approaches.

2.3. Instrumental setups

For our analysis we employed two different instrumental setups. Xevo G2-S Q-TOF-MS (Waters, USA) was used for Oslo samples while Maxis Impact Q-TOF-MS (Bruker, Germany) was employed for Athens samples. Each instrumental setup performed the analysis under different data acquisition conditions. The diversity in the instruments and the data acquisition settings introduced an extra level of complexity to our dataset, which was desired in order to better characterize the limits in the capabilities of the investigated strategies. Detailed information regarding the chromatography and HR-MS conditions is provided in Supporting Information section 2.3.

2.4. Data processing algorithms

We evaluated three different approaches used for suspect analysis of complex samples. Among the alternative methods, the two stage algorithm implemented via NIVA_MZ_Analyzer^{TM} was selected. The two stage algorithm was previously developed for target and suspect analysis of the GC-HR-TOF-MS data [17]. The second tested algorithm was UNIFI by Waters (for more information see section 2.4.2). This algorithm was considered a representative method for software that employ feature detection or peak peaking and/or modeling as the first step of analysis (e.g. MZmine [9] and XCMS [10]). The third approach was implemented via TASQ by Bruker. This algorithm utilizes the retention time match of the XICs and the isotopic ratios as a means for identification, similar to software packages such as MassHunter by Agilent and ChromaLynx by Waters. These three software programs appeared to cover the two approaches commonly used and an alternative one for

the suspect and target analysis of complex samples analyzed via LC-HR-QTOF-MS. Moreover, this is one of the rare studies where the performance of an open access software is compared to the vendor software packages.

2.4.1. NIVA_MZ_Analyzer^{TM}

The NIVA_MZ_Analyzer^{TM} is a Graphical User Interface (GUI) via matlab [25] for implementation of the two stage algorithm [17] applicable to both GC-HR-TOF-MS and LC-HR-QTOF-MS data. This algorithm is also independent from mass spectrometer manufacturers. The two stage algorithm produces a spectra library of the standards in the first stage and then calculates the similarity score between each scan of the sample chromatogram and the standard spectrum. More information regarding the suspect analysis procedure is provided in SI, section S4.1.

During the first stage (i.e. Unique Ion Extractor) the MS^1 and MS^2 spectra for each standard peak are processed separately. These two spectra are then recorded and stored in the standard library. Therefore, for each entry in the standard library there are two spectra for the low and high collision energies. During the second stage, the MS^1 and MS^2 spectra are treated as two complementary spectra. Therefore, the similarity matrix between the library component and the sample spectra are calculated separately for the MS^1 and MS^2. The similarity matrix is the dot product of the weighted and normalized reference spectra and each scan of the sample [17]. Thus if an analyte is present in the sample, higher levels of similarity scores are observed at the same location for both the MS^1 and MS^2 spectra when compared to the background similarity values. Furthermore, five XICs (including the parent

ion and 4 qualifiers) are extracted in order to increase the level of confidence in the identification. The two stage algorithm uses a six point criterion for positive and negative identifications (for more information please see section 2.6 and S4). This algorithm showed to be effective for the suspect screening of GC-HR-TOF-MS data [17] however, this is the first implementation of the two stage algorithm for LC-HR-QTOF-MS data.

2.4.2. UNIFI

UNIFI is a commercially available software from Waters. This software first performs the peak picking/feature detection in the whole chromatogram employing the MS^1 signal, without taking into account the analytes of interest. UNIFI, utilizes the continuous wavelet transformations (CWTs) developed by Du et al., 2006 [26]. This algorithm (i.e. CWTs) performs the peak detection, peak modeling, and also the noise reduction simultaneously. Further information about the CWTs can be found elsewhere [15]. The next step is the grouping of ions in both the MS^1 and MS^2 spectra, which belong to a the same compound using the isotopic ratios, 2D feature shape (i.e. both in time and m/z domains), and the feature retention time. This process results in the final peak and/or unique feature list and the deconvoluted spectra of each individual chromatographic peak in the list. During the suspect analysis, UNIFI employs the information such as the exact mass of: the molecular ion, fragments, and the potential adducts for the identification. Finally, UNIFI employs user defined filters to priorities the features that are more likely to be positive detections. Work-flows similar to this approach have been widely utilized for suspect and non-target analysis of complex samples as reviewed elsewhere [1, 8, 27].

2.4.3. Target Analysis for Screening and Quantitation (TASQ^{TM})

TASQ, is a commercially available software package provided by Bruker for the target and suspect analysis of complex samples [28]. The algorithm implemented in this software package uses the information provided by the analyst (i.e. the exact mass of molecular ion and the fragments or the molecular formula of fragments (which contains the information of their exact mass) in order to extract the XIC of the relevant fragments for the analyzed suspect. In this case this method first performs the peak picking in the XIC of the exact mass provided by the analyst. The theoretical isotopic pattern of the molecular ion in the MS^1 is checked as the second step in the identification. Once the isotopic pattern of the molecular ion was fit, then the algorithm uses the XICs of the fragments provided by the user for increasing the confidence level in the identification. These XICs are grouped together based on the match in their retention times. During the final step, a score is generated for each peak based on the isotopic fit, mass accuracy, and XIC retention times. This methodology has been successfully used for the target and suspect analysis of environmental samples, including wastewater extracts [5].

2.5. Target and suspect analysis

The chromatograms of all four samples were analyzed both in target and suspect analysis mode (see section S3 for more information regarding the data preprocessing). We used NIVA_MZ_Analyzer^{TM} only in suspect analysis mode for all four samples (i.e. Oslo samples and Athens samples). This implied that we did not use the reten-

tion time information for any of 89 analytes investigated in this study for analysis with NIVA_MZ_Analyzer^{TM}. As a validation tool, the Oslo samples were analyzed via UNIFI both in target and suspect mode for the Oslo analytes (i.e. 33 analytes) plus the common analytes (i.e. 7 analytes). The same samples were analyzed only in suspect mode for the Athens and library analytes. In other words the Oslo samples were processed twice. First in suspect mode with all 89 analytes and a second time in target analysis mode only with the Oslo analytes plus the common analytes. A similar approach was adopted for the Athens samples where we first analyzed those samples in suspect mode for all 89 analytes and in the second round we analyzed the same samples in target mode including only the Athens analytes plus common analytes. This approach enabled the cross validation of the results of our suspect analysis for all three software.

2.6. The identification criteria during the target and suspect analysis

During both target and suspect analysis we employed conservative identification criteria, which reduced the odds of false positive detection. For the NIVA_MZ_Analyzer^{TM} a suspect analyte was considered present in the sample if it met at least 6 out of 7 criteria. These criteria consisted of peaks in the similarity matix for both the MS^1 and MS^2 spectra; the peak of the exact mass; the peak for at least 2 out 4 XICs; and finally the retention time match for all 5 to 7 peaks. These criteria were established based on our preliminary assessment of the two stage algorithm for the LC-HR-QTOF-MS data.

UNIFI used a six point criteria for positive and negative detections during the

suspect analysis whereas it used a seven point criteria for target analysis. These criteria included the isotopic fit, mass accuracy of ± 2 mDa for all the relevant ions, detection of at least 2 out of 4 fragments, and the XIC retention time match for those ions in the sample chromatogram. These XIC retention times were independent from the analyte retention time and only were used for the spectral deconvolution. During the target analysis for the Oslo analytes, the analyte retention times were compared to the retention times of the potential features in the sample. A retention window of 0.25 min was employed for the target screening, section S4 of Supporting Information. These settings were previously optimized for target and suspect analysis of wastewater samples [18].

The positive detection criteria when using TASQ consisted of: a mass accuracy of ± 2 mDa for both the exact mass and the fragments; a good isotopic fit, which was represented with the "mSigma Tolerance" parameter with a maximum of 100; a positive match for at least 2 out of 4 fragments; and fragment retention match, in order to include that fragment in the spectra of the sample peak. Also for TASQ, the retention time of the standards were compared to the retention times of the potential peaks, having a retention tolerance of 0.2 min, during the target analysis. This implied that for an analyte to have its presence confirmed in the sample, during the suspect analysis, it must have at least 6 out of 7 criteria fulfilled. For the target analysis TASQ matched the standard retention time to the sample peak retention time using a retention tolerance of 0.2 min. More information regarding the parameter settings of TASQ are provided in SI, section S4.3. A recent study of the

wastewater influent samples from Athens treatment plant showed the applicability of these parameters for our type of analysis [5].

The fragments used in both TASQ and UNIFI were extracted from the mzCloud database [29]. For analytes where more than one source was available, we gave priority to the average spectra recorded by Eawag. The mzCloud was selected as the reference database for the qualifier ions. This selection was based on the quality of the spectra stored in this database, the fact that this database was an independent database, and the ease of use [30]. We used only the annotated fragments with relatively high intensities in the average spectra of each analyte. A text file with the compound names, smiles, exact mass of the parent ion, and the exact mass of four fragments compiled and then imported into both UNIFI and TASQ. Therefore, the absolute intensity of the ions in the reference spectra (i.e. the mzCloud entry) was not considered during the identification via both UNIFI and TASQ.

For the NIVA_MZ_Analyzer^{TM} library creation, the standard mixture chromatograms of all the analytes were submitted to the UIE module. This software also has the option of importing new spectra as txt/csv files. This enables the inclusion of spectra recorded in the open databases such as MassBank [31] into the local library of NIVA_MZ_Analyzer^{TM}.

3. Results and discussions

3.1. Cross platform compatibility

The cross platform similarity values of all 7 common analytes were three order of magnitudes larger than the minimum threshold of similarity, see section S5. The largest maximum threshold of similarity at MS^1 was of 0.1627 for Carbamazepine recorded via Waters instrument whereas the smallest value for the maximum similarity threshold of 0.0072 was observed for Morphine, also recorded with Waters equipment, Figure 1. When looking at the MS^2, the highest value of 0.071 was registered with Metoprolol via Waters and the lowest value of 0.0071 was observed for Morphine recorded with Waters. The higher variability observed in the Waters instrument compared to the Bruker setup, was attributed to the applied ramp in the collision energy of the MS^2, while for the Bruker instrument a constant collision energy was applied for the MS^2 spectra. In the case of cross platform similarity, this parameter ranged from 0.0083 for Citalopram to 0.1105 for Carbamazepine. In overall, for both MS^1 and MS^2, we observed a slightly lower average cross platform similarities compared to the similarity values within each platform, Figure 1. We interpreted that the observed decrease in the cross platform similarities compared to the maximum threshold similarities were caused by the differences in the collision energies and also the hardware design of the two instrumental setups. However, the non-parametric Kruskal-Wallies test [32] at 95% confidence interval, with p values of 0.84 indicated that, in both MS^1 and MS^2, the observed differences between the cross platform similarity values and the maximum threshold similarities were not statistically significant. This implied that the average spectra, for the common analytes,

recorded by different instruments are comparable to each other. These results were consistent with the findings of Oberacher et al. [33] and Scheubert et al. [34], which suggested by using an average acquisition condition, an effective spectral weighting function, and dot product the cross platform compatibility can be achieved. This implies that these spectra can be used for suspect and/or retrospective analyses of samples acquired via different instruments. However, a more comprehensive evaluation of cross platform compatibility with a large pool of instruments and analytes is necessary for this aim and will be the subject of our near future study.

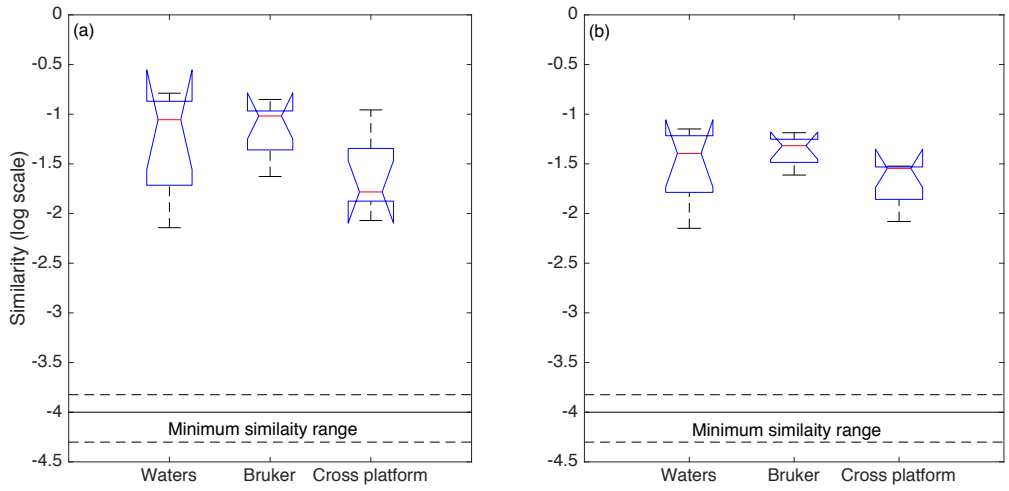

Figure 1: Calculated maximum threshold similarity and cross platform similarity of 7 common analytes at (a) MS^1 and (b) MS^2. The similarity values are available in Table S4 in Supporting Information. The solid line at the bottom of the figure depicts the averaged minimum threshold of similarity whereas the dashed lines show the variability observed in the minimum threshold of similarity.

3.2. Suspect analysis and analyte identification

Four wastewater samples (i.e. Oslo samples and Athens samples) were analyzed for 89 analytes with NIVA_MZ_Analyzer^{TM} in suspect analysis mode. The Oslo samples were analyzed in suspect analysis mode with UNIFI whereas the Athens samples were analyzed via TASQ. As a validation tool we analyzed the Oslo samples for Oslo analytes plus the common analytes (i.e. total of 40 analytes out of 89) in target analysis mode whereas for Athens samples we utilized TASQ for Athens analytes plus the common analytes (i.e. total of 39 analytes out of 89) also in target analysis mode. For the ease of discussion we are going to divide our analytes in two categories. The first category, herein referred to as artificial suspect analytes, are the analytes which their presence or absence was confirmed via target analysis in that sample. The second category or true suspects were the analytes that were not analyzed via target analysis in that sample, Tables 1 and S1. We further evaluated our results for false negative, false positive, potential false negative, and potential false positive. A false negative refers to the case where an analyte is not detected via suspect analysis however its presence in the sample is confirmed, employing target analysis. A false positive represents a case where during the suspect analysis a compound is detected while its absence in that sample is confirmed via target analysis. Potential false negative refers to cases of pure suspects, where all the evidence indicates the presence of an analyte in the sample, while the investigated software produces a negative detection. For potential false positive, we refer to a case where a software results in positive detection of an analyte even though all the evidence suggest the contrary. In cases of potential false positives and negatives, the

absence or presence of the analyte can not be confirmed due to the lack of retention time information. The presence of both artificial suspects and true suspects, enabled us to simulate a real life case and truly evaluate the limitations of each approach.

3.2.1. Analysis of the artificial suspects

NIVA_MZ_Analyzer^{TM} confidently detected 9 out of 11 artificial suspect analytes in one of the Oslo samples, while detecting 8 out of 10 in the second sample, Table 1. The discrepancy in the number of positive detection in the Oslo samples was caused by the false negative detections of Citalopram in both samples, Morphine in the Oslo1 sample, and Pseudoephedrine in Oslo2 sample. For these compounds NIVA_MZ_Analyzer^{TM} produced a false negative, due to the fact that the detection requirements (i.e. the six point criterion) by the algorithm were not met, Figure 2. The similarity matrix in both low and high collision energies clearly indicated the positive detection of these compounds (e.g. Citalopram, Figure 2). However, the qualifiers selected for them were not present. These false negative cases were caused by high level of background signal, thus ion suppression, which affected the most the selected qualifier ions for these analytes. The effect of the background signal on NIVA_MZ_Analyzer^{TM} has been discussed in detail in section 3.3. For the Athens samples, NIVA_MZ_Analyzer^{TM} detected 22 out of 23 target validate analytes in the Athens1 sample whereas it detected 23 out of 24 analytes in the Athens2 sample. For the analytes Irbesartan and Cetirizine, in the Athens samples NIVA_MZ_Analyzer^{TM} in suspect analysis mode produced a positive detection while TASQ in target analysis mode resulted in their negative detections. We further interrogated the raw data at the expected location (i.e. the retention time)

of these analytes in Athens samples. Our investigation showed that both of these two analytes were present in the sample, Figure 3. For Irbesartan and Cetirizine, we interpret that TASQ faced some difficulties in spectral calibration, that resulted in a mass error of $\sim \pm$ 3 mDa. Consequently, TASQ reported these two analytes as negative detections. Once widening the mass error window for Irbesartan and Cetirizine all two analytes were detected both during the suspect and target analysis. Therefore, we considered these two analytes as detected and artificial suspect in the Athens samples. NIVA_MZ_Analyzer^{TM} failed to detect Citalopram, D617, and Caffeine in the Athens samples. In this case similar to Citalopram in Oslo samples, the minimum requirements of the algorithm for positive detection of these two analytes in Athens samples were not met. Therefore, NIVA_MZ_Analyzer^{TM} resulted in false negatives for Citalopram, D617 and Caffeine in the Athens samples. In overall, the NIVA_MZ_Analyzer^{TM} produced zero cases of false positives while producing few cases of false negative (i.e. 10 cases out of total analyzed cases of 158). This software showed to be a capable tool for confident detection of suspect analytes in a complex sample. Furthermore, since NIVA_MZ_Analyzer^{TM} does not need peak picking and it utilizes the whole spectrum rather than few qualifiers, during the suspect analysis, it has a lower potential of false positive detection.

UNIFI detected 13 out of 11 artificial suspect analytes in the Oslo1 sample whereas it detected 10 out of 10 in the Oslo2 sample, Table 1. We observed 2 clear cases of false positive for Hydroxycotinine and Mephedrone in Oslo1 sample. These analytes were detected during the suspect analysis, while were not detected in

target analysis mode. UNIFI did not produce any case of false negative. However, it showed to have some difficulties to distinguish analytes such as MDMA and Carbamazepine in both samples. This software produced multiple potential peaks for these two analytes. Therefore, UNIFI needed retention time information in order to distinguish these potential peaks from the actual peaks of MDMA and Carbamazepine. These difficulties were caused by the complexity of the background signal, which contained the exact mass and more than 2 qualifiers of these two analytes in multiple locations. For the sample Oslo2, UNIFI produced results which were 100% in agreement between the suspect and target analysis. Besides the two false positive cases, UNIFI showed to be a reliable tool in suspect analysis of complex samples. However, depending of the complexity of the samples UNIFI may require additional information and further post processing.

For the Athens samples, after including the two missed analytes (i.e. Irbesartan, and Cetirizine), TASQ in suspect analysis mode, detected 28 out of 25 artificial suspect analytes in sample Athens1 and 25 out of 26 in Athens2 sample. For Athens1 sample, TASQ resulted in 3 false positives for analytes D,L,N,O-Didesmethylvenlfaxine, Aliskiren, and Picaridin and 0 cases of false negative. In case of Athens2 sample TASQ produced only 1 case of false negative for Venlafaxine. The cases of false positive when performing suspect analysis with TASQ were attributed to the complexity of the samples. Therefore, other compounds, structurally similar to our analytes, were detected instead of those analytes. Similar to UNIFI, TASQ also detected multiple locations for analytes such as Gabapentin and Carbamazepine,

which indicated that this algorithm likewise to UNIFI needed retention time information for these analytes.

NIVA_MZ_Analyzer^{TM} was the only algorithm that did not result in any false positives among the tested software, while producing in total 10 cases of false negative. These false negative cases were mainly caused by the ion suppression due to the sample complexity. The other two approaches (i.e. UNIFI and TASQ) produced both cases of false positive and false negative, which was acceptable considering the complexity of the analyzed samples. These cases of false positive and false negative would have been avoided during post processing where the analysts use their knowledge to increase the level of confidence in the identifications. During this study, we took the necessary precautions to avoid the introduction of background knowledge of the analysts into the final results in order to objectively compare the performances of the investigated algorithms. Our results indicate the importance of full spectral comparison (implemented via NIVA_MZ_Analyzer^{TM}) rather than employing few qualifiers commonly used during suspect analysis in order to have the maximum level of confidence in the identification.

3.2.2. True suspect analysis

The NIVA_MZ_Analyzer^{TM} confidently detected 8 analytes out of 49 true suspect analytes (i.e. Athens plus Library analytes) in the Oslo1 sample. In the Oslo2 sample this algorithm detected 7 out of 49 true suspect analytes. We analyzed these samples for the same 49 analytes using UNIFI. Except one case (i.e. Eprosartan in Oslo2 sample) all the positive detections with NIVA_MZ_Analyzer^{TM} were

Table 1: The number of positive detection, true positives (i.e. artificial suspect analytes) in each sample processed via three tested approaches.

	Samples							
	Oslo1		Oslo2		Athens1		Athens2	
Software	Pos. detec.[a]	True pos.[b]	Pos. detec.[a]	True pos.[b]	Pos. detec.[a]	True pos.[b]	Pos. detec.[a]	True pos.[b]
NIVA_MZ_Analyzer^{TM}	9	11	8	10	22	23-25[c]	23	24-26[c]
UNIFI	13	11	10	10	-	-	-	-
TASQ	-	-	-	-	28	22-25[c]	25	23-26[c]

[a]The number of detected analytes in each sample using a specific software package; [b]The total number of artificial suspect suspect analytes in each sample; [c]After confirming the presence of Irbesartan and Cetirizine in the Athens samples these two analytes were added to the total number of true positives in Athens samples.

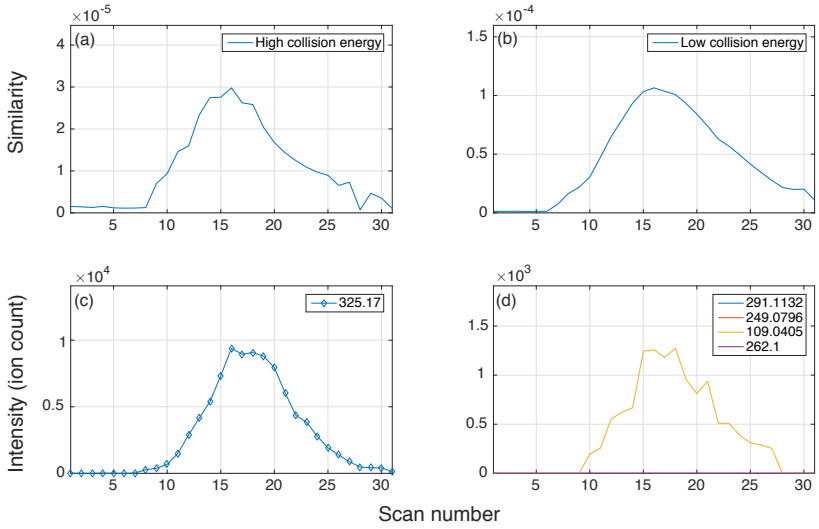

Figure 2: The similarity matrix of Citalopram (a) at MS^2, (b) at MS^1, and (c) and (d) the XICs for the exact mass and three qualifiers in Oslo1 sample.

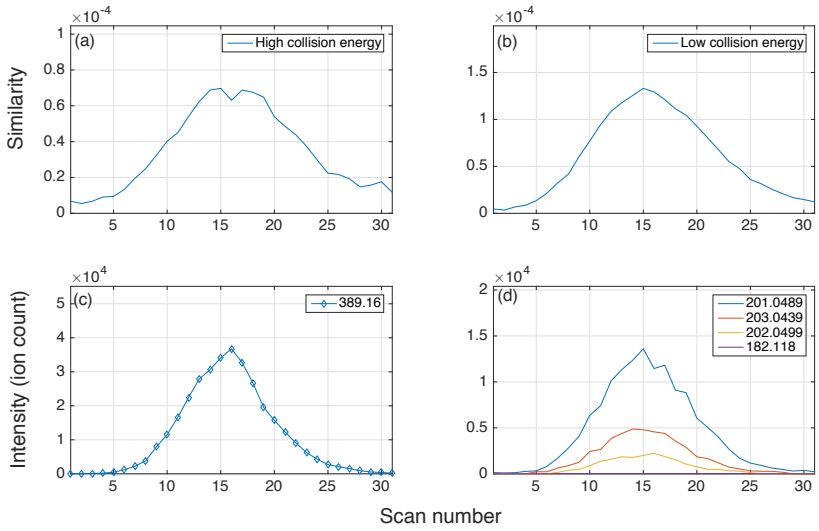

Figure 3: The similarity matrix of Cetirizine (a) at MS^2, (b) at MS^1, and (c) and (d) the XICs for the exact mass and three qualifiers in Oslo1 sample.

also detected with UNIFI. However, there were several cases of discrepancy where NIVA_MZ_Analyzer^{TM} resulted in negative detection of an analyte while UNIFI detected that analyte in the same sample. We further investigated the observed differences between NIVA_MZ_Analyzer^{TM} and UNIFI results. More detailed interrogation of the raw data revealed three distinct cases. Case one including analytes Mirtazapine, Valsartan, and Caffeine in Oslo1 sample and Valsartan, Caffeine, Ephedrine, and Atenolol acid in Oslo2 sample. For these analytes NIVA_MZ_Analyzer^{TM} was not able to confirm their presence in the Oslo samples due to lack of 6 positive detection criteria. These cases were similar to the Citalopram case in the artificial suspect analysis, Figure 2. Therefore, for these analytes the NIVA_MZ_Analyzer^{TM} produced potential false negative results. The second case included analytes Clozapine, Ecgoninemethylester and para-Methoxy-N-methylamphetamine in Oslo1 sample and Ecgoninemethylester in the Oslo2 sample. By further examining our data, we concluded that UNIFI produced potential false positive results for these analytes, Figure 4. Finally, the third case consisted of Eprosartan. This analyte was detected by NIVA_MZ_Analyzer^{TM} in Oslo2 sample whereas it was not detected via UNIFI. We scrutinized the Oslo2 dataset and observed that spectral calibration performed via UNIFI for this peak was not successful, which consequently caused its lack of detection in that sample. Therefore it was concluded that this analyte is most likely present in the sample and UNIFI produced a potential false negative result for this analyte, Figure 5.

NIVA_MZ_Analyzer^{TM} confidently detected 3 true suspect out of 50 Oslo ana-

lytes plus Library analytes in Athens1 sample. This method resulted in 5 positive detections in Athens2 sample. Analysis of the same samples using TASQ resulted in 7 positive detections in Athens1 sample and 6 detections for Athens2 sample. By interrogating the discrepancy cases between NIVA_MZ_Analyzer^{TM} and TASQ, we observed two different categories of suspect analytes. The first category included Ecgoninemethylester, Diclofenac, Hydroxycotinine, and Buprenorphine in Athens1 sample and Ecgoninemethylester, Methylone, and Methylphenidate in Athens2 sample. For these analytes TASQ resulted in positive detections while NIVA_MZ_Analyzer^{TM} did not detect them. We extracted the signal to noise ratio (S/N) reported for these analytes by TASQ and for all the mentioned analytes in both samples the S/N appeared to be smaller than 9. Furthermore, the results of NIVA_MZ_Analyzer^{TM} indicated the high level of similarity between the background signal and these analytes' signal, Figure 6. Even though the background knowledge of the wastewater samples may suggest that these analytes are likely to be present in the samples, the data interrogation did not provide enough evidence to confirm that. Further investigation is needed in order to draw any type of conclusions regarding these analytes. The second category consisted of analytes Cocaine and 4-MEC in the Athens2 sample. These analytes were not detected with TASQ whereas they were confidently detected via NIVA_MZ_Analyzer^{TM}. For Cocaine the absolute intensity of the exact mass (i.e. 850) was smaller than the threshold of 1250 set by TASQ, therefore this analyte was reported as not detected. When looking at 4-MEC the signal produced by 2 qualifiers out of 3 were large and thus they produced broad peaks. The peak picking algorithm in the TASQ failed to detect these broad qualifier peaks therefore TASQ

reported 4-MEC as not detected. Based on these evidence we concluded that TASQ produced a potential false negative for both Cocaine and 4-MEC.

NIVA_MZ_Analyzer^{TM} was able to confidently detect the suspect analytes in complex samples. This algorithm did not produce any cases of potential false positives while resulting in 7 potential false negatives out of 198 cases. The results also indicate the necessity of full spectral comparison in order to produce high level of confidence in the identification. In the case of true suspect analysis also NIVA_MZ_Analyzer^{TM} appeared to be as successful as commercially available software. In some cases this algorithm performed far better than both UNIFI and TASQ. UNIFI, resulted in 4 cases of potential false positives whereas TASQ produced 7 cases of inconclusive outcome. These inconclusive cases are plausible to be potential false positives. In overall, algorithms such as NIVA_MZ_Analyzer^{TM}, which perform full spectral analysis rather than few qualifier inspection are more likely to produce reliable results with a high level of confidence.

Table 2: The number of positive detections using the three algorithm investigated in this study.

	Pure suspects			
	Samples			
Softwares	Oslo1	Oslo2	Athens1	Athens2
NIVA_MZ_Analyzer^{TM}	8	7	3	5
UNIFI	15	13	-	-
TASQ	-	-	7	6

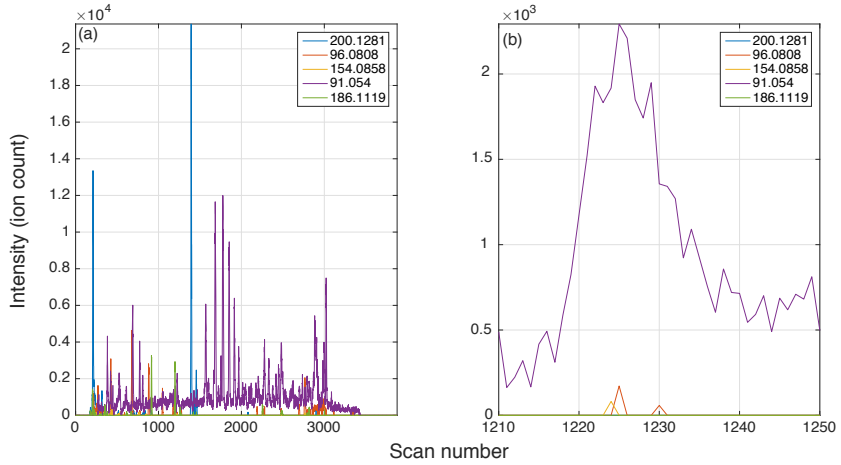

Figure 4: The XICs of the exact mass and 4 relevant ions of Ecgoninemethylester (a) over the whole chromatogram and (b) the zoomed in region based on reported retention time by UNIFI in the Oslo2 sample.

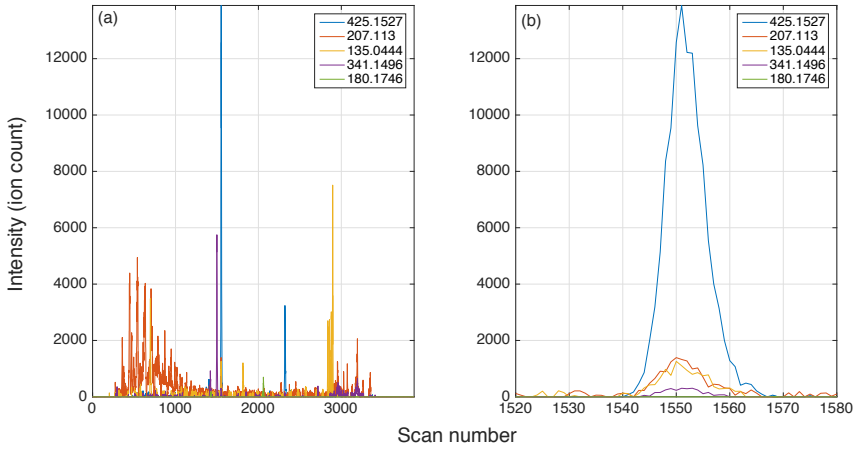

Figure 5: The XICs of the exact mass and 4 relevant ions of Eprosartan (a) over the whole chromatogram and (b) the zoomed in region based on reported retention time by NIVA_MZ_Analyzer^{TM} in the Oslo1 sample.

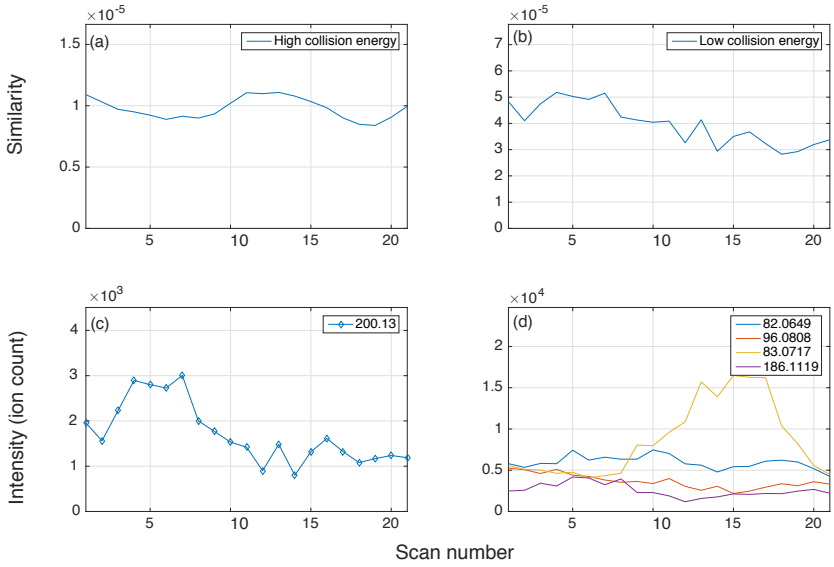

Figure 6: The similarity matrix of Ecgoninemethylester (a) at MS^2, (b) at MS^1, and (c) and (d) the XICs for the exact mass and three qualifiers in Athens1 sample.

3.2.3. Rates of false positive and false negative

We compared the 3 algorithms investigated in this study in terms of rates of false positive and negative, Table 3. The rate of false positives was defined as the ratio of false positives and false positives plus true negatives whereas the rate of false negative was defined as the ratio of false negatives and the false negatives plus true positives [35]. For these calculations we used only the artificial suspect analytes. The rates of false positive and false negative enabled a direct comparison of these 3 algorithms and also provided us with indicative levels of confidence in the final identifications produced by each algorithm.

NIVA_MZ_Analyzer^{TM} resulted in the smallest rate of false positive compared to

the other two methods. This algorithm produced the highest rate of false negatives compared to the other algorithms, Table 3. The higher rate of false negative was attributed to the highly conservative detection criteria used by this algorithm, which guarantees the highest level of confidence in the identifications. Both UNIFI and TASQ produced between 0.06 to 0.09 rates of false positives. This again indicated the importance of full spectral comparison rather than the qualifier approach. The performances of UNIFI and TASQ were comparable, considering the limited number of evaluated cases. In overall, the NIVA_MZ_Analyzer^{TM} appeared to be more robust compared to the other algorithms when looking at the levels of confidence in the identifications. Moreover, both UNIFI and TASQ require post processing in order to avoid cases of false positive, thus a high level of confidence in the identification.

Table 3: The number of true positive, true negative, false positive, false negative, rate of false positive, and rate of false negative calculated for all 3 algorithm investigated in this study.

Algorithm	True +	True -	False +	False -	Rate false +	Rate false -
NIVA_MZ_Analyzer^{TM}	72	88	0	10	0	0.12
UNIFI	21	59	4[a]	2	0.06	0.09
TASQ	51	29	5[b]	1-4[c]	0.15	0.02-0.07[d]

[a]The number of false positives includes analytes MDMA and Carbamazepine, where the retention time information is necessary for correct detection; [b]The number of false positives includes analytes Gabapentin and Carbamazepine, where the retention time information is necessary for correct detection; [c]This number of false negatives includes analytes Irbesartan and Cetirizine, which were not detected due to the calibration issues; [d]This rate of false negatives includes analytes Irbesartan and Cetirizine, which were not detected due to the calibration issues.

3.3. The effect of the sample complexity and concentration on the two stage algorithm

We examined the effect of both background complexity and analyte concentration on the performance of the two stage algorithm. We generated synthetic background

signal using real data plus noise. The procedure of the signal generation is explained in detail in Supporting Information, section S6. In our evaluation, we looked at three different levels of complexity in the background signal. These levels consisted of: low complexity, where there were less than ∼10% common features between the background signal and the analyte signals; medium complexity, where we observed ∼ 50% similarity between the background signal and the analyte signals; and finally, the high complexity, in which there were ∼ 80% similarity between the background signal and the analyte signals, Figure S3. We carried out the identification employing two stage algorithm for 7 common analytes at 7 signal dilution factors ranging from 1 to 0.01. By applying different signal dilution factors, we observed the signal to background ratio (S/B) varying from 158 to 0.1. The signal was defined as the sum of all intensities in the analyte spectra whereas the background was defined as the averaged sum of all the intensities in the background spectra (See section S6 in SI).

The two stage algorithm confidently distinguished all 7 analytes from the background signal for all three background levels of complexity at signal dilution factors ≥ 0.25. The background complexity had a negative effect on the performances of two stage algorithm. For simplicity we are going to focus on two extreme cases Paracetamol and Morphine. These two analytes appeared to be representative for all 7 common analytes. Paracetamol due to its size and structure produced a limited number of ions in both low and high collision energies. Therefore, the two stage algorithm was less affected by complexity of the background signal, Figure 7. For

Morphine, the number of low intensity ions (i.e. small fragments) generated, particularly at MS^2, were larger than Paracetamol. Therefore, the two stage algorithm faced more difficulties in separating the analyte signal from the background. However also in the case of Morphine the two stage algorithm was able to separate the two signals.

The concentration showed to have a positive effect on the performances of two stage algorithm. At larger concentrations the two stage algorithm was able to separate the analyte signal from the background signal at all three complexity levels and for all 7 common analytes. At high signal dilution factors (i.e. lower concentrations) with high background complexity the two stage algorithm was not able to confidently detect the analyte signal in the background, Figure 8. The level of dissimilarity for the lowest concentrations of successfully detected analytes via two stage algorithm was reevaluated. The results of this experiment indicate that the two stage algorithm requires at least 10% dissimilarity between the analyte and the background signal in order to be able to separate the two signals from each other. We also evaluated the S/B of 7 common analytes at different concentration levels and background complexities. We observed a minimum S/B values of ~ 1 required for two stage algorithm in order to distinguish the analyte signal from the background. For Paracetamol the S/B values of ~ 1, for both low and high collision energy spectra, were reached at signal dilution factor of 0.1 at the highest background complexity whereas for the low background complexity this S/B ratios were obtained with signal dilution factors 0.01. On the other hand, for Morphine, the S/B values of ~ 1 were reached at higher

concentrations (i.e. signal dilution factor of 0.25) compared to Paracetamol for all three background complexity levels. Therefore, for an analyte such as Morphine with highly complex background higher concentrations were needed in order for two stage algorithm to separate the analyte signal from the background.

The two stage algorithm was shown to be able to distinguish the analyte signal from the complex background signal. This algorithm requires S/B values of ≥ 1 for both low and high energy channels in order to be able to distinguish the analyte signal from the background. Furthermore, a minimum of 10% dissimilarity between the analyte signal and the background is needed for the confident identification of an analyte in the complex background.

4. Conclusions

The comparison between an alternative algorithm (i.e. two stage algorithm implemented via NIVA_MZ_Analyzer^{TM}) and two commonly used approaches (i.e. UNIFI and TASQ) was performed via the suspect analysis of 89 analytes in four wastewater influent samples. We observed cross platform similarities larger than 3 orders of magnitude from the minimum threshold of similarity. A slight decrease in the level of similarity was observed in cross platform similarities vs within platform similarity. However, the Kruskal-Wallies test results indicated that the differences observed were not statistically significant. This was indicative of the cross platform applicability of the data produced via different instrument for suspect/retrospective analysis for the instrumental setups evaluated in this study. Further research on this subject is

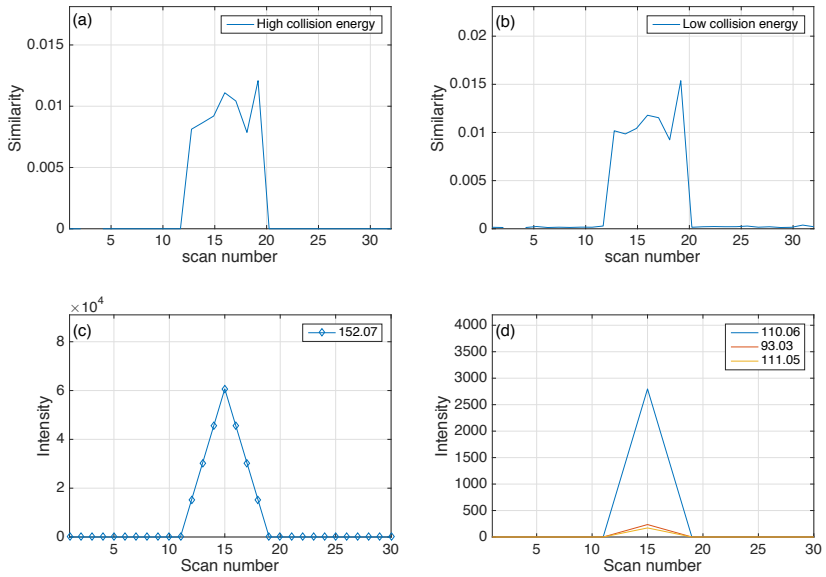

Figure 7: The similarity matrix of Paracetamol (a) at MS^2, (b) at MS^1, and (c) and (d) the XICs for the exact mass and three qualifiers. These data are generated at low background complexity and the lowest signal dilution factor (i.e. 1).

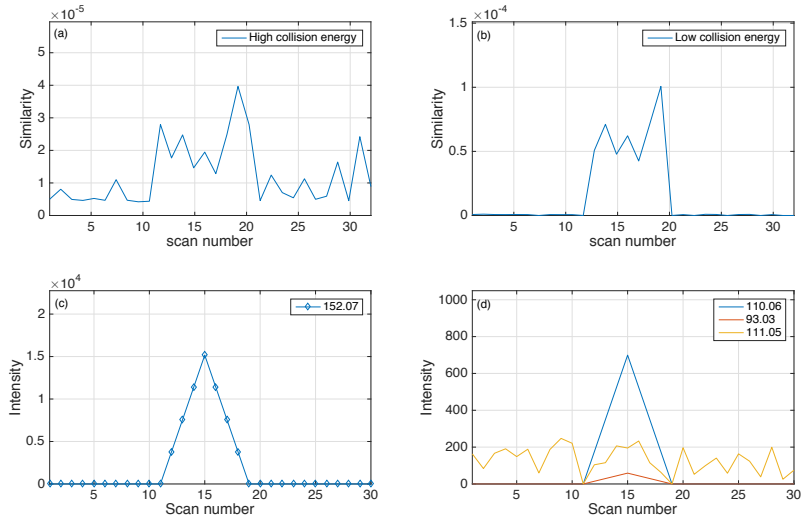

Figure 8: The similarity matrix of Paracetamol (a) at MS^2, (b) at MS^1, and (c) and (d) the XICs for the exact mass and three qualifiers. These data are generated at high background complexity and the highest signal dilution factor (i.e. 0.25).

needed in order to validate a universal data acquisition procedure for LC-HR-QTOF-MS data.

For the artificial suspect analytes, the two stage algorithm did not produced any cases of false positive while resulting in 10 false negatives out of total 158 artificial suspect cases. UNIFI and TASQ both produced $\sim$ 5 cases of false positives and between 0 to 4 cases of false negatives. For true suspects, the two stage algorithm again did not produced any cases of potential false positives while resulting in 7 cases of potential false negatives. UNIFI for the pure suspects produced 4 cases of potential false positives and 1 case of potential false negative. TASQ for the pure suspects produced 7 cases of inconclusive/potential false positive and 2 cases of false

negative. Our direct comparison of the three approaches with the means of rate of false positive indicated that the two stage algorithm with a rate of false positive of 0 resulted in the most robust method with the highest level of confidence in the identifications.

The evaluation of the background complexity effect on the two stage algorithm showed that this algorithm requires at least 10% of dissimilarity between the background and the analyte signal. Three levels of complexity ranging from 20% dissimilarity (i.e. the highest level of complexity) to 90% dissimilarity between the analyte and background signal were evaluated. We also evaluated the effect of S/B ratio on the performances of the two stage algorithm. We observed that this algorithm needs ≥ 1 S/B ratios for both MS^1 and MS^2 in order for the algorithm to produce the highest level of confidence in the identification. For the conventional approaches, the signal to noise ratio was considered an indication of the background complexity. Overall, the two stage algorithm showed to be more affected by the signal suppression compared to the conventional methods, due to its use of the full spectra. These evaluation enabled us to define the boundaries in which the two stage algorithm is capable to produce reliable results.

The two stage algorithm as well as UNIFI and TASQ were capable of performing suspect analysis in the investigated complex samples. However, contrary to the two stage algorithm both UNIFI and TASQ needed fine tuning of the filters and post processing in order to minimize the odds of false positive detections. With UNIFI

and TASQ the background knowledge and the experience of the analyst plays an important role in increasing the level of confidence in the final identification.

This study showed the cross platform compatibility of the data produced via two tested HR-TOF-MS instruments. This implies that generation of a universal HR spectral database similar to the NIST library is also possible for LC-HR-QTOF-MS data. However, more efforts and future studies are needed in order to establish a standard operational procedure (SOP) for LC-HR-QTOF-MS data acquisition.

5. Acknowledgements

The authors are thankful to Prof. Bert Van Bavel and Dr. Katherine Langford for their help during this project. We are also thankful to the Research Council of Norway for the financial support of this project (RESOLVE, 243720).

6. References

[1] Eva Gorrochategui, Joaquim Jaumot, Sílvia Lacorte, and Romà Tauler. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. *Trends Anal. Chem.*, 82:425–442, 2016.

[2] Emma L Schymanski, Heinz P Singer, Jaroslav Slobodnik, Ildiko M Ipolyi, Peter Oswald, Martin Krauss, Tobias Schulze, Peter Haglund, Thomas Letzel, Sylvia Grosse, et al. Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. *Anal. Bioanal. Chem.*, 407(21):6237–6255, 2015.

[3] Emma L Schymanski, Heinz P Singer, Philipp Longree, Martin Loos, Matthias Ruff, Michael A Stravs, Cristina Ripolles Vidal, and Juliane Hollender. Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. *Environ. Sci. Technol.*, 48(3):1811–1818, 2014.

[4] Christoph Moschet, Alessandro Piazzoli, Heinz Singer, and Juliane Hollender. Alleviating the reference standard dilemma using a systematic exact mass suspect screening approach with liquid chromatography-high resolution mass spectrometry. *Anal. Chem.*, 85(21):10312–10320, 2013.

[5] Pablo Gago-Ferrero, Emma L. Schymanski, Anna A. Bletsou, Reza Aalizadeh, Juliane Hollender, and Nikolaos S. Thomaidis. Extended suspect and non-target strategies to characterize emerging polar organic contaminants in raw wastewater with LC-HRMS/MS. *Environ. Sci. Technol.*, 49(20):12333–12341, 2015.

667 [6] Aurea C Chiaia-Hernandez, Emma L Schymanski, Praveen Kumar, Heinz P
668 Singer, and Juliane Hollender. Suspect and nontarget screening approaches to
669 identify organic contaminant records in lake sediments. *Anal. Bioanal. Chem.*,
670 406(28):7323–7335, 2014.

671 [7] Bin Zhou, Jun Feng Xiao, Leepika Tuli, and Habtom W Ressom. LC-MS-based
672 metabolomics. *Molecular BioSystems*, 8(2):470–481, 2012.

673 [8] Lunzhao Yi, Naiping Dong, Yonghuan Yun, Baichuan Deng, Dabing Ren, Shao
674 Liu, and Yizeng Liang. Chemometric methods in data processing of mass
675 spectrometry-based metabolomics: A review. *Anal. Chem. acta*, 914:17–34,
676 2016.

677 [9] Mikko Katajamaa, Jarkko Miettinen, and Matej Orešič. Mzmine: toolbox for
678 processing and visualization of mass spectrometry based molecular profile data.
679 *Bioinformatics*, 22(5):634–636, 2006.

680 [10] Colin A Smith, Elizabeth J Want, Grace O'Maille, Ruben Abagyan, and Gary
681 Siuzdak. XCMS: processing mass spectrometry data for metabolite profiling
682 using nonlinear peak alignment, matching, and identification. *Anal. chem.*,
683 78(3):779–787, 2006.

684 [11] Marc Sturm, Andreas Bertsch, Clemens Gröpl, Andreas Hildebrandt, Rene Hus-
685 song, Eva Lange, Nico Pfeifer, Ole Schulz-Trieglaff, Alexandra Zerck, Knut Rein-
686 ert, et al. Openms–an open-source software framework for mass spectrometry.
687 *BMC bioinformatics*, 9(1):163, 2008.

[12] Jianqiu Zhang, Elias Gonzalez, Travis Hestilow, William Haskins, and Yufei Huang. Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Current genomics*, 10(6):388–401, 2009.

[13] Jamie B Coble and Carlos G Fraga. Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery. *J. Chromatogr. A*, 1358:155–164, 2014.

[14] Gunnar Libiseller, Michaela Dvorzak, Ulrike Kleb, Edgar Gander, Tobias Eisenberg, Frank Madeo, Steffen Neumann, Gert Trausinger, Frank Sinner, Thomas Pieber, et al. Ipo: a tool for automated optimization of xcms parameters. *BMC bioinformatics*, 16(1):1, 2015.

[15] Ralf Tautenhahn, Christoph Böttcher, and Steffen Neumann. Highly sensitive feature detection for high resolution LC/MS. *Bioinformatics*, 9(1):1, 2008.

[16] Frans M Van Der Kloet, Ivana Bobeldijk, Elwin R Verheij, and Renger H Jellema. Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping. *J. Proteome. Res.*, 8(11):5132–5141, 2009.

[17] Saer Samanipour, Katherine Langford, Malcolm J Reid, and Kevin V Thomas. 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. Chromatogr. A*, 1463:153–161, September 2016.

[18] Jose A. Baz-Lomba, Malcolm J. Reid, and Kevin V. Thomas. Target and sus-

pect screening of psychoactive substances in sewage-based samples by UHPLC-QTOF. *Anal. Chem. acta*, 914:81–90, 2016.

[19] Jose Antonio Baz-Lomba, Stefania Salvatore, Emma Gracia-Lor, Richard Bade, Sara Castiglioni, Erika Castrignanò, Ana Causanilles, Felix Hernandez, Barbara Kasprzyk-Hordern, Juliet Kinyua, et al. Comparison of pharmaceutical, illicit drug, alcohol, nicotine and caffeine levels in wastewater with sale, seizure and consumption data for 8 european cities. *BMC public health*, 16(1):1035, 2016.

[20] Grung Merete, S. Heimstad Eldbjorg, Moe Morten, Schlabach Martin, Svenson Anders, Thomas Kevin, and Woldegiorgis Andreas. Human and veterinary pharmaceuticals, narcotics, and personal care products in the environment. Technical report, IVL, NILU and NIVA, 2007.

[21] Nikiforos A Alygizakis, Pablo Gago-Ferrero, Viola L Borova, Alexandra Pavlidou, Ioannis Hatzianestis, and Nikolaos S Thomaidis. Occurrence and spatial distribution of 158 pharmaceuticals, drugs of abuse and related metabolites in offshore seawater. *Sci. Total Environ.*, 541:1097–1105, 2016.

[22] Susanne Kern, Kathrin Fenner, Heinz P Singer, René P Schwarzenbach, and Juliane Hollender. Identification of transformation products of organic contaminants in natural waters by computer-aided prediction and high-resolution mass spectrometry. *Environ. Sci. Technol.*, 43(18):7039–7046, 2009.

[23] Martin Krauss, Heinz Singer, and Juliane Hollender. LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns. *Anal. Bioanal. Chem.*, 397(3):943–951, 2010.

[24] Petros Dimitriou-Christidis, Alex Bonvin, Saer Samanipour, Juliane Hollender, Rebecca Rutler, Jimmy Westphale, Jonas Gros, and J. Samuel Arey. GC × GC quantification of priority and emerging nonpolar organic micropollutants in all types of wastewater streams: Analysis methodology, chemical occurrence, and partitioning. *Environ. Sci. Technol.*, 49:7914–7925, 2015.

[25] MATLAB User's Guide. The mathworks. *Inc., Natick, MA*, 5, 1998.

[26] Pan Du, Warren A Kibbe, and Simon M Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065, 2006.

[27] Anna A Bletsou, Junho Jeon, Juliane Hollender, Eleni Archontaki, and Nikolaos S Thomaidis. Targeted and non-targeted liquid chromatography-mass spectrometric workflows for identification of transformation products of emerging pollutants in the aquatic environment. *Trends Anal. Chem.*, 66:32–44, 2015.

[28] Bruker corporation. TASQ^{TM} software ver1.1 - Target Analysis for Screening and Quantitation (TASQ). `www.bruker.com/products/mass-spectrometry-and-separations/ms-software/tasq/tasq-software.html`, August 2016.

[29] R Mistrik, J Lutisan, Y Huang, M Suchy, J Wang, and M Raab. mzCloud: A key conceptual shift to understand 'who's who'in untargeted metabolomics. In *Metabolomics Society 2013 Conference, Glasgow, July*, pages 1–4, 2013.

[30] Maria Vinaixa, Emma L Schymanski, Steffen Neumann, Miriam Navarro,

752 Reza M Salek, and Oscar Yanes. Mass spectral databases for LC/MS-and
753 GC/MS-based metabolomics: State of the field and future prospects. *TrAC*
754 *Trends Anal. Chem.*, 78:23–35, 2016.

755 [31] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku
756 Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken
757 Aoshima, et al. Massbank: a public repository for sharing mass spectral data
758 for life sciences. *J. Mass. Spectrom.*, 45(7):703–714, 2010.

759 [32] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance
760 analysis. *J. Amer. Statist. Assoc.*, 47(260):583–621, 1952.

761 [33] Herbert Oberacher, Marion Pavlic, Kathrin Libiseller, Birthe Schubert, Michael
762 Sulyok, Rainer Schuhmacher, Edina Csaszar, and Harald C Köfeler. On the
763 inter-instrument and inter-laboratory transferability of a tandem mass spectral
764 reference library: 1. results of an austrian multicenter study. *J. Mass. Spectrom.*,
765 44(4):485–493, 2009.

766 [34] Kerstin Scheubert, Franziska Hufsky, and Sebastian Böcker. Computational
767 mass spectrometry for small molecules. *J. Cheminform.*, 5(1):1, 2013.

768 [35] Neil A Macmillan and C Douglas Creelman. *Detection theory: A user's guide*.
769 Psychology press, 2004.