

Article

The Critical Assessment of Small Molecule Identification (CASMI): Challenges and Solutions

Emma L. Schymanski ^{1,*} and Steffen Neumann ^{2,*}

¹ Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, Dübendorf CH-8600, Switzerland

² IPB: Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, Halle (Saale) DE-06120, Germany

* Authors to whom correspondence should be addressed;

E-Mails: emma.schymanski@eawag.ch (E.L.S.); sneumann@ipb-halle.de (S.N.);

Tel.: +41-58-765-5537 (E.L.S.); +49-345-5582-1470 (S.N.);

Fax: +41-58-765 5826 (E.L.S.); +49-345-5582-1409 (S.N.).

Received: 1 April 2013; in revised form: 25 May 2013 / Accepted: 7 June 2013 /

Published: 25 June 2013

Abstract: The *Critical Assessment of Small Molecule Identification*, or CASMI, contest was founded in 2012 to provide scientists with a common open dataset to evaluate their identification methods. In this article, the challenges and solutions for the inaugural CASMI 2012 are presented. The contest was split into four categories corresponding with tasks to determine molecular formula and molecular structure, each from two measurement types, liquid chromatography-high resolution mass spectrometry (LC-HRMS), where preference was given to high mass accuracy data, and gas chromatography-electron impact-mass spectrometry (GC-MS), *i.e.*, unit accuracy data. These challenges were obtained from plant material, environmental samples and reference standards. It was surprisingly difficult to obtain data suitable for a contest, especially for GC-MS data where existing databases are very large. The level of difficulty of the challenges is thus quite varied. In this article, the challenges and the answers are discussed, and recommendations for challenge selection in subsequent CASMI contests are given.

Keywords: mass spectrometry; metabolite identification; small molecule identification; contest; metabolomics; non-target identification

1. Introduction

The CASMI contest, the *Critical Assessment of Small Molecule Identification*, was founded in 2012. The aim of CASMI [1] was to encourage experts to exhibit their identification methods on a common dataset and, thus, enable a better comparison of the methods available. The task was to determine the molecular formula and/or the molecular structure from the mass spectrometry data. The myriad of options available for small molecule identification (vendor software, specialized independent software, open access and open source options) makes it increasingly difficult for users and researchers alike to keep pace with the changes. Instead, offering a common dataset enables the use of expert knowledge or any chosen identification methods and provides a basis for comparison. The aim of CASMI was to include all disciplines interested in small molecule identification and, thus, enable the cross-disciplinary exchange of information and expertise. In this article, small molecule identification refers to molecules of approximately 50–1,000 Da that can be detected with mass spectrometric (MS) techniques.

Although MS identification methods are often categorized according to the chromatographic separation used (e.g., gas chromatography (GC) *versus* liquid chromatography (LC)), with relatively recent instrumental developments, such as high resolution and soft-ionization GC-MS, it is difficult to distinguish separation, detection and identification techniques and set distinct categories for a competition to allow a broad range of participants. The inaugural CASMI focused on two measurement types, liquid chromatography-high resolution mass spectrometry (LC-HRMS), where preference was given to high mass accuracy MS/MS data, and gas chromatography-electron impact-mass spectrometry (GC-MS), focusing on unit mass accuracy data. Although this excluded some participants, e.g., those with only unit mass accuracy LC-MS/MS data experience and those with high mass accuracy GC-MS data, these categories could be considered for future CASMI contests.

The data collection commenced in the early months of 2012, with the original aim of 20 challenges per category. The ‘unknowns’ could not be truly unknown for the purpose of the competition and, thus, required a confirmed identity. However, this made it difficult to obtain suitable data, especially for the GC-MS data where many of the challenges available were also in common databases. In the end, challenges were obtained from plant material, environmental samples and reference standards. As it was difficult to find GC-MS challenges that were not in the NIST database [2], challenges were provided that only had a relatively low probability ($\leq 60\%$) in the database search, although a couple with high probability ($> 90\%$) were added to give some variety. This compromise meant that the level of difficulty of the challenges was quite varied. Despite this compromise, however, the initial target of 20 suitable challenges for each category was not achieved. In the GC-MS dataset, the final 16 challenges were all confirmed with reference standards, and although other substances were available in the samples provided, they did not have matching standards. For the LC-HRMS data, it was difficult to obtain identified “unknowns”, as those already published could be linked to the names and/or institutes of the organizers, while those unpublished were often intended for a forum other than CASMI. This finally resulted in 14 LC-HRMS challenges. Six of the challenges were part of pathway elucidation efforts to determine gene function during investigations into the biochemistry of natural products and their role in the development and defenses of plants. The remaining eight environmental substances were taken

from “failed confirmations” (which were not suitable for publication alone) and one successful target identification of a rare compound, not yet published.

In retrospect, given the small number of participants for the first CASMI and the fact that not all participants contributed to all challenges within a category, the number of challenges seemed appropriate. Although a smaller number of challenges may have encouraged more participants, a larger number of challenges is needed to provide sufficient variety in the difficulty and chemical diversity of the challenges and to allow a proper evaluation. The disadvantage of providing many challenges is that it creates an advantage for fully automated entries, as methods requiring the input of human expertise are generally more time-consuming.

In this article, the challenges and the answers are discussed, along with recommendations for challenge selection in subsequent CASMI contests. Details about the participants and the outcome of the first contest can be found in [3], also in this special issue.

2. LC-HRMS Challenges and Solutions (Category 1 and 2)

The LC-HRMS challenges were sourced from plant material and standards purchased for confirmation of unknowns in environmental investigations. All challenges contained the elements C, H, N, O, P and S; no halogens were present in any compound; see Table 1 for an overview. Appendix A contains annotated spectra for each challenge, which show the composite spectrum of all available MS/MS files for each challenge and, thus, display the most intense peak where a given peak occurred in multiple spectra within the error window of 0.0001 Da plus 5 ppm. The MS spectra were also included for certain challenges. The fragments were annotated using ACD/ChemSketch [4] and Mass Frontier [5] and processed automatically using OpenBabel [6] and a script in R [7] to determine placement. Although the most realistic fragments were selected, many of these are tentative and have not been confirmed unambiguously. Appendix B provides more details on the challenge compounds, including PubChem and ChemSpider identifiers.

As it proved difficult to obtain suitable challenge compounds for this contest, there was no ‘easy vs. hard’ pre-selection. In the end, this meant some of the challenges were quite challenging, while others were too easy. Although challenges that were in reference databases were avoided as far as possible, some compounds were uploaded to MassBank [8] after the challenge data was released.

2.1. LC-HRMS Challenges 1 to 6

The first six challenges were metabolites that were encountered as part of plant metabolomics research. The compounds were measured on a Bruker micrOTOF-Q equipped with an electrospray ionization (ESI) source in positive mode, which generally achieves ≤ 5 ppm mass accuracy and 12,000 resolution during routine measurements. At this resolution, the extraction of the isotopic fine structure (which would resolve, e.g., the ^{15}N or ^{34}S isotope peaks) is not possible, but the isotope intensities are generally very accurate. The data was acquired with a 3 Hz scan frequency for both MS and MS/MS acquisitions.

Table 1. Liquid chromatography (LC) Challenges for the Critical Assessment of Small Molecule Identification (CASMI) 2012.

Challenge	Trivial Name	Formula	Exact mass
1	Kanamycin A	$C_{18}H_{36}N_4O_{11}$	484.2381
2	1,2-Bis-O-sinapoyl-beta-D-glucoside	$C_{28}H_{32}O_{14}$	592.1792
3	Glucosquerrin	$C_{14}H_{27}NO_9S_3$	449.0848
4	Escholtzine	$C_{19}H_{17}NO_4$	323.1158
5	Reticuline	$C_{19}H_{23}NO_4$	329.1627
6	Rheadine	$C_{21}H_{21}NO_6$	383.1369
10	1-Aminoanthraquinone	$C_{14}H_9NO_2$	223.0633
11	1-Pyrenemethanol	$C_{17}H_{12}O$	232.0888
12	alpha-(o-Nitro-p-tolylazo)acetoacetanilide	$C_{17}H_{16}N_4O_4$	340.1172
13	Benzylidiphenylphosphine oxide	$C_{19}H_{17}OP$	292.1017
14	1H-Benz[g]indole	$C_{12}H_9N$	167.0735
15	1-Isopropyl-5-methyl-1H-indole-2,3-dione	$C_{12}H_{13}NO_2$	203.0946
16	[1-(4-methoxyanilino)-1-oxopropan-2-yl] 6-oxo-1-propylpyridazine-3-carboxylate	$C_{18}H_{21}N_3O_5$	359.1481
17	Nitrin	$C_{13}H_{13}N_3$	211.1109

Challenge 1 was kanamycin A ($C_{18}H_{36}N_4O_{11}$), an aminoglycoside compound with antibiotic effects from bacteria. The compound was available as an authentic standard. The challenge data comprised the full-scan data, including two isotope peaks and fragment-rich MS/MS spectra at 10 eV, 20 eV and 30 eV in positive mode, shown in Figure A1. The MS/MS of the three collision energies were acquired in consecutive scans, which reduced the effective scan frequency for one collision energy to 1 Hz. The LC-HRMS/MS data was processed with the XCMS centWave feature detection [9], and the compound spectrum was extracted with CAMERA [10]. This approach is described in greater detail in [11].

Challenge 2 was 1,2-bis-O-sinapoyl- β -D-glucoside ($C_{28}H_{32}O_{14}$), which was extracted from canola seeds and characterized previously [12]. The challenge data in negative mode included isotopes up to (M + 3) and a single fragment-rich MS/MS spectrum, shown in Figure A2, which was also extracted with XCMS and CAMERA, as described above. The raw data provided initially was affected by a severe calibration problem, resulting in ≈ 30 ppm mass deviation. The data, recalibrated to within 5 ppm accuracy, was provided to the participants after the contest closed, to offer them a chance to recalculate their results on more accurate data for the special issue.

Challenge 3 was glucosquerrin ($C_{14}H_{27}NO_9S_3$), found with other glucosinolates in the seeds of *Brassicaceae*. Among others, the glucosinolates 3-methylthiopropyl (3MTP, glucoibervirin), 4-methylthiobutyl (4MTB, glucoerucin), 7-methylthioheptyl (7MTH) and 8-methylthiooctyl (8MTO) are described in [13]. The challenge data was measured from a methanolic extract of *Arabidopsis*

thaliana seeds, in negative mode. Although no authentic standard was used, the confidence in the identification was quite high based on the molecular formula determined with high mass accuracy data, characteristic product ions and the consistency of the structural information (including retention time) with other glucosinolates of different chain lengths. Isotopes were present up to $(M + 4)$. The MS/MS spectra (see Figure A3) were extracted with XCMS and CAMERA (as described above) and did not contain the precursor ion for collision energies above 20 eV.

Challenges 4–6 were combined into a single sample and measured together in positive mode. As for Challenge 1, the collision energy was alternated in the raw file, but in contrast to the previous challenges, the MS/MS data was extracted from a single scan for each compound and collision energy. All peaks below an intensity of 1% of the base peak were removed. The spectra are given in Figures A4–A6. The data provided originally was not calibrated and had mass deviations up to 8 ppm. After the closing of the contest, the data was recalibrated and provided to the participants. This resulted in deviations below 5 ppm for Challenges 4 and 6, but at the same time, increased the mass error for Challenge 5 to approximately 6 ppm.

Challenge 4 was the alkaloid escholtzine ($C_{19}H_{17}NO_4$). The isotopic pattern included only peaks up to $(M + 2)$, while the 30 eV MS/MS spectrum was very noisy.

Challenge 5 was another alkaloid, reticuline ($C_{19}H_{23}NO_4$). While the 20 eV MS/MS spectrum still contains the precursor, the 30 eV spectrum contains a few additional fragments below m/z 176.

Challenge 6 was the alkaloid rheadine ($C_{21}H_{21}NO_6$). The MS/MS spectra contained more fragments than the previous challenge.

As all of these compounds were in PubChem, they could be considered “known unknowns”. Challenges 7 to 9 are absent; as discussed above, the original aim of 20 challenges was not attained, and the original numbering was kept in this article for consistency with the participant results and publications.

2.2. LC-HRMS Challenges 10 to 17

These challenges resulted from unconfirmed tentative identifications arising from the effect-directed analysis (EDA) of river water sampled from the Elbe (Czech Republic) using the passive sampler, blue rayon [14], where CASMI provided some ‘use’ for standards that otherwise had no specific purpose. As a result, some of these are quite challenging challenges, whereas others are more straightforward. All these challenges were taken from measurements of reference standards, using either ESI or atmospheric pressure chemical ionization (APCI) techniques; the fragmentation modes were either collision-induced dissociation (CID) or higher-energy collisional dissociation (HCD); the settings reported are as normalized collision energies (NCE).

Challenge 10 was 1-aminoanthraquinone, shown in Figure A7. Although amino groups are usually a distinctive loss in many compounds, here, the first losses are a water from a carbonyl group (m/z 206), resulting in a rearrangement to form a stabilizing four-membered ring with the amino-substituent, as well as the loss of the full carbonyl group itself (m/z 196). The loss of a full benzyl group results

in the fragment at m/z 146, likely also stabilized by the formation of a four-membered ring; while the remaining fragment at m/z 105.033 is likely to result from the loss of the same benzyl ring along with one of the carbonyl groups, where the charge remains with the smaller fragment. The accurate mass of the fragment confirms the formula C_7H_5O , rather than, e.g., a nitrogen adduct (m/z 105.044), such as those seen in [15].

Challenge 11 was 1-pyrenemethanol and had a difficult MS spectrum to interpret, although the very simple fragmentation pattern was also informative. Both the MS and MS/MS are plotted in Figure A8. The behavior of substances can be a lot less consistent with APCI and atmospheric pressure photo ionization (APPI) compared with ESI, and this substance undergoes an in-source loss and oxidation to an $[M - H]^+$ ion. The only losses are the hydroxy group and the complete methanol substituent. The fact that no other fragments are generated despite targeted MS/MS on the m/z 215 peak indicates that a stable aromatic backbone is likely to be present. In-source oxidation has been reported previously, for example in [16], the isobars, tonalide and galaxolide, could not be separated chromatographically, but could be identified using their different ionization behavior in positive mode. Tonalide was visible as both $[M]^{+\bullet}$ and $[M + H]^+$, whereas galaxolide was detected as $[M - H]^+$ (an in-source oxidation product) and the $[M]^{+\bullet}$ ion. The authors explained this with differing proton affinities, demonstrating that galaxolide has a lower proton affinity than the proton donors in the APPI source and, thus, competed unfavorably for the protons.

Challenge 12 was α -(o-nitro-p-tolylazo)acetoacetanilide, commonly known as “Pigment Yellow 1” and was a target compound identified only through site-specific information [14]. This challenge would be difficult for *de novo* structure elucidation, as it is quite a big molecule and has a wide variety of functional groups. The many functional groups also make it difficult to incorporate predictive selection strategies. Even with knowledge of the true structure, it was difficult to annotate all the major MS/MS peaks using either simple bond-breaking approaches or the general and library fragmentation rules in Mass Frontier [5]. The major annotations are shown in Figure A9. It is likely that a much more detailed elucidation of the fragmentation processes would be needed to annotate all peaks, which was beyond the scope of this article.

Challenge 13 was benzyldiphenylphosphine oxide and was one of the easier challenges, for database searching and structure generation alike, when taking the spectrum into account. The only “degree of freedom” was the location of the CH_2 or CH_3 group (*i.e.*, whether a benzyl or methylphenyl substituent was present). The spectrum of this compound and similar compounds were uploaded to MassBank [8,17] before the submission deadline. The major fragments are shown in Figure A10.

Challenge 14 was 1H-benz[g]indole, another stable, aromatic compound. Although the spectra (shown together in Figure A11) display more fragments than Challenge 11, the collision energy is much greater here (HCD 120 and 180 NCE, compared with CID at 35 NCE above). The fragments at m/z 167 and m/z 168 are potentially a mix of $[M]^{+\bullet}$ and $[M + H]^+$, with a H loss as the first major fragment. The remaining fragments are successive two-member losses from the aromatic system; first, CNH

followed by two C_2H_x losses. The fragments given in Figure A11 are indicative; a rearrangement may stabilize the fragments at m/z 141 and m/z 91.

Challenge 15 was 1-isopropyl-5-methyl-1H-indole-2,3-dione and has quite a small, aromatic-stabilized system with a distinctive isopropyl loss in the MS/MS spectrum, followed again by the break-up of the aromatic system (see Figure A12). The presence of m/z 91 indicates that the methyl group is attached to a benzene ring; m/z 106 indicates also that the N is attached to the same benzene ring. The carbonyl groups again display a loss of water, as well as the full substituent.

Challenge 16 was [1-(4-methoxyanilino)-1-oxopropan-2-yl] 6-oxo-1-propylpyridazine-3-carboxylate. This challenge was a candidate for an unknown identification, where the original unknown remains unidentified. This compound experiences significant fragmentation, such that neither the molecular ion nor any adducts of the molecular ion are present in the MS. The MS and MS/MS are merged in the spectrum displayed in Figure A13. Energy-based fragmentation scoring (as in, e.g., MetFrag [18]) can prioritize the wrong compounds, such as here, where the fragmentation was too favorable. Thus, the presence in a compound database does not necessarily mean that the compound is conducive to identification via MS/MS analysis.

Challenge 17 is nitrin, another unconfirmed tentative identification where the presence of a $C_6H_5(N \equiv N)^+$ fragment in the original unknown spectrum led to the (incorrect) tentative identification of nitrin. The peak instead arose from a nitrogen adduct formed during MS/MS measurements, a phenomenon observed with several aromatic compounds (e.g., [15,19]). One result of the adduct detection was the expansion of the fragment formula annotation option in RMassBank to include adducts by adding N_2 and O to the allowed elements of the subformulas [15]. The spectrum of Challenge 17 is shown in Figure A14. The fragment at m/z 105.044 corresponding to a $C_6H_5(N \equiv N)^+$ fragment is conspicuously absent in the MS/MS spectrum of this compound. Instead, fragmentation occurs between the Ns, and only a few pieces of the molecule are observed. Interestingly, the fragment at m/z 77 (characteristic for a phenyl substituent) was very small, confirming that fragmentation occurs preferably between the Ns.

3. GC-MS Challenges and Solutions (Category 3 and 4)

All GC-MS challenges are summarized in Table 2 and were sourced from real environmental samples and were confirmed with reference standards. This requirement of being certain of the identity (for the purpose of a contest), but also not being too easy to find in a database, was a big challenge for the GC-MS data, as over 200,000 compounds are now included in GC-MS databases, such as NIST [2]. As a result, challenges were selected where the probability for a database match was relatively low, *i.e.*, not a 'straightforward' identification. Many of these are quite standard compounds, but the spectra were taken from real samples (instead of the database) to add some variety. A couple of isomers were chosen to see if computational methods could match the ability of databases to distinguish isomers. A couple of challenges that did not meet this 'low probability' requirement were added to diversify the challenge

set further. There were a lot more halogens (chlorine only) present in these spectra compared with the LC-HRMS challenges.

As no external participants participated in these categories, these challenges are not described in detail. The structures and several identifiers are given in Appendix C, Figure C18.

Table 2. Gas chromatography (GC) Challenges for CASMI 2012.

Challenge	Trivial Name	Formula	Nominal mass
1	Phthalic anhydride	C ₈ H ₄ O ₃	148
2	Phthalimide	C ₈ H ₅ NO ₂	147
3	2-Chlorobenzyl alcohol	C ₇ H ₇ ClO	142
4	4-Chlorobenzyl alcohol	C ₇ H ₇ ClO	142
5	1,4-Dichlorobenzene	C ₆ H ₄ Cl ₂	146
6	Acenaphthene	C ₁₂ H ₁₀	154
7	4-Chlorobenzoic acid	C ₇ H ₅ ClO ₂	156
8	Fluorene	C ₁₃ H ₁₀	166
9	Methyl 2-chlorobenzoate	C ₈ H ₇ ClO ₂	170
10	2,4,6-Trichlorophenol	C ₆ H ₃ Cl ₃ O	196
11	Formothion	C ₆ H ₁₂ NO ₄ PS ₂	257
12	alpha-Hexachlorocyclohexane	C ₆ H ₆ Cl ₆	290
13	Dimethyl carbonotrithioate	C ₃ H ₆ S ₃	138
14	O,O,O-Trimethyl thiophosphate	C ₃ H ₉ O ₃ PS	156
15	Dibenzofuran	C ₁₂ H ₈ O	168
16	O,S,S-Trimethyl phosphorodithioate	C ₃ H ₉ PS ₂ O ₂	172

3.1. GC-MS Challenges 1 and 2

These two challenges were chosen due to the availability of standards for retention index (RI) calculation [20]. These were very closely related; only an O and NH are different.

3.2. GC-MS Challenges 3 to 16

Challenges 3–16 came from the EDA of a groundwater sample from Bitterfeld, Germany [21,22]. Fractionation using reverse-phase high performance liquid chromatography (RP-HPLC) with a C18 column and preparative GC (pcGC) was performed prior to the final GC-MS analysis (for more details, see [21]). As a result, partitioning information could be calculated for the individual fractions, and this provided additional information for the identification, which was made available to the CASMI participants. The compounds identified in the sample are quite common environmental contaminants

that could have resulted in almost trivial identification results for participants with access to a large GC-MS database.

4. Recommendation for Future CASMIs

The problem of insufficient spectra and, especially for GC, too many spectra in the databases, could be improved in future CASMIs by sourcing compounds from a synthetic laboratory, which would be able to provide rare compounds, but also confirm their identity. ‘Unknown unknowns’ are not suitable for a competition, as the identity must be known to declare the winner(s).

Due to the lack of participants in the GC-MS categories, the organizers of the next CASMI may consider adding a different category to complement the accurate mass LC-HRMS categories. Some possibilities include an accurate mass GC category, or GCxGC-TOF, or changing the focus to different MS/MS ionization techniques, rather than forming distinct GC and LC categories. It is also plausible that only two categories should be offered in the next CASMI, *i.e.*, restricting the competition to Categories 1 and 2 only. Another enhancement to the LC-HRMS categories could be the inclusion of challenges measured along with a set of standard compounds to provide reference retention times, or providing participants with candidate lists that they would need to rank.

One way to improve participation in future CASMIs could be to provide additional incentives, such as prizes. The opportunity to submit papers to a special issue does not appear to have been sufficient incentive to attract many participants in the 2012 contest. Although sponsorship would be an option, it can compromise the independence (or at least, the appearance of independence) of the competition. An alternative, more scientific incentive could be the organization of a CASMI identification workshop, which would require more participants to be successful.

CASMI could also provide the ideal exchange platform for selected ‘unknown unknowns’ in the future, where scientists could submit their unknowns and offer other experts (and expert systems) the chance to identify them. Obviously, no winners can be declared when the answer is unknown; the contributor of the ‘unknown unknown’ would be required to decide the appropriate ‘reward’.

Acknowledgments

We would like to thank all of those who provided the challenge data. Christoph Böttcher, Stephan Schmidt, Jürgen Schmidt and Jörg Ziegler from the IPB contributed LC Challenges 1–4 and their valuable time determining the structures. Toni Kutchan (formerly IPB, now at the Donald Danforth Plant Science Center, Missouri, USA) provided Challenges 5 and 6. Christine Gallampois and Martin Krauss performed the measurements for LC Challenges 10–16, while Cornelia Meinert measured the GC challenges, all at the Department of Effect Directed Analysis, Helmholtz Center for Environmental Research (UFZ) in Leipzig, Germany. Permission to use the data was kindly granted by Werner Brack. LC Challenge 17 was measured at Eawag by Matthias Ruff and Cristina Ripollés Vidal (visiting from University Jaume I, Castellón, Spain). Junho Jean from Eawag and the peer reviewers provided valuable comments on the manuscript. ES acknowledges the EU Marie Curie Postdoctoral Fellowship funding (Grant Number 299734).

Conflict of Interest

The authors have no conflicts of interest to declare and made no financial gain from organizing CASMI.

Appendix

A. Annotated spectra

This appendix contains the annotated spectra for LC-MS Challenges 1–17. The structures were determined with the help of experience, ChemSketch [4] and MassFrontier [5]. Selected fragments were added to a script in R [7], while the processing of the spectra, including placement of the fragments, was automatic. OpenBabel [6] was used to generate the images.

Figure A1. Challenge 1: annotated merged MS and MS/MS spectra of kanamycin A (electrospray ionization (ESI), positive mode).

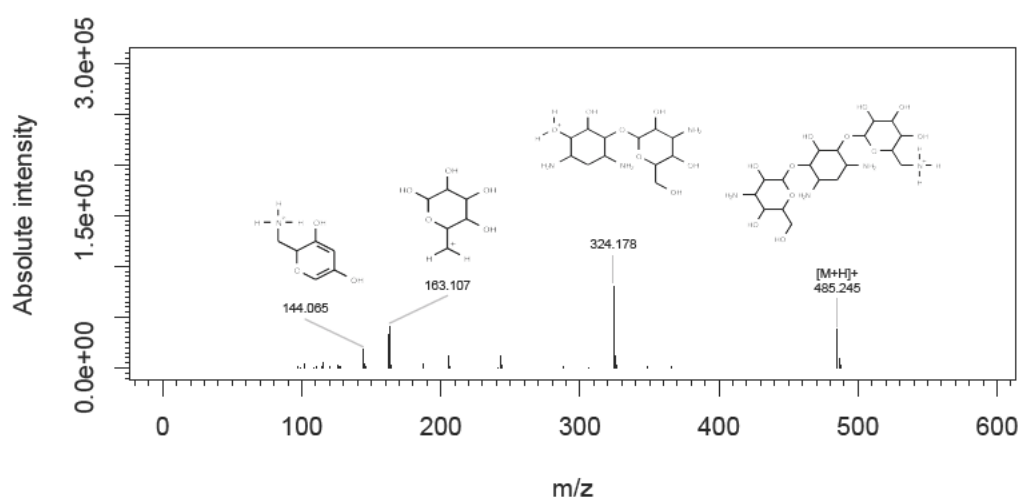


Figure A2. Challenge 2: annotated MS/MS spectrum of 1,2-bis-O-sinapoyl- β -D-glucoside (ESI, negative mode).

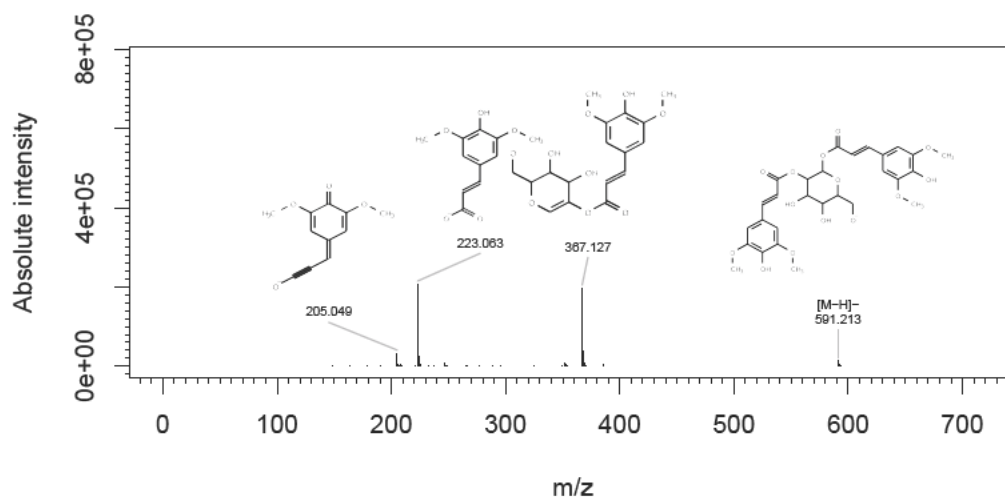


Figure A3. Challenge 3: annotated merged MS/MS spectra of glucolesquerellin (ESI, negative mode).

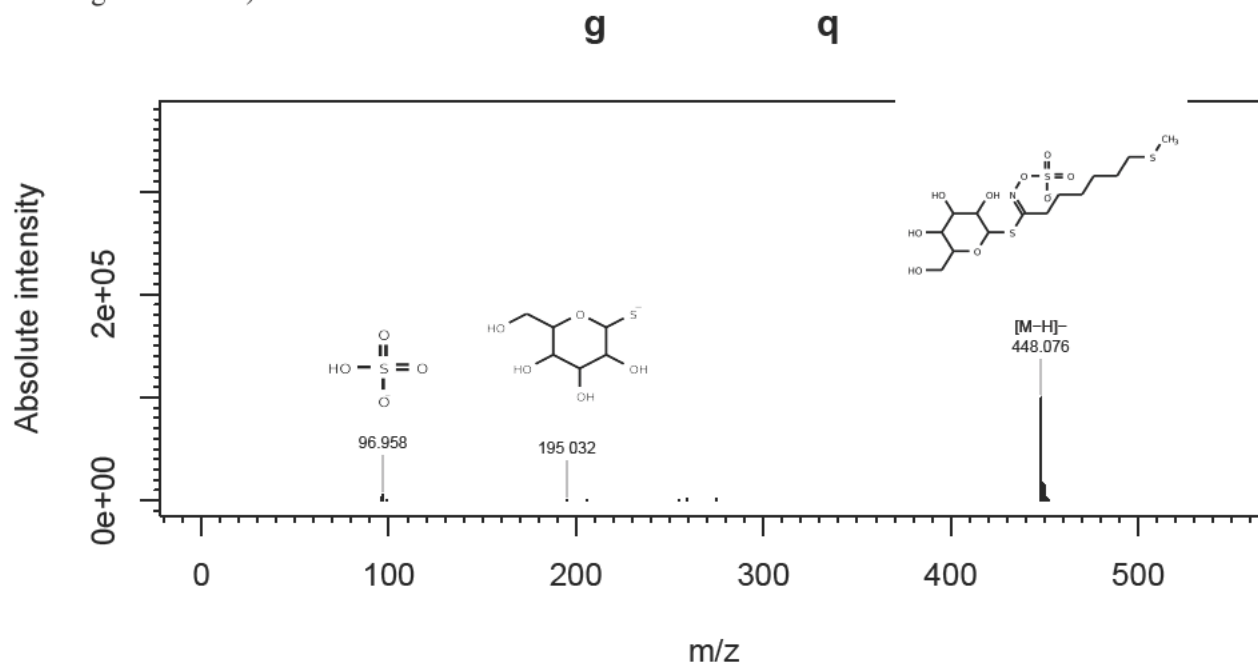


Figure A4. Challenge 4: annotated merged MS/MS spectra of escholtzine (ESI, positive mode).

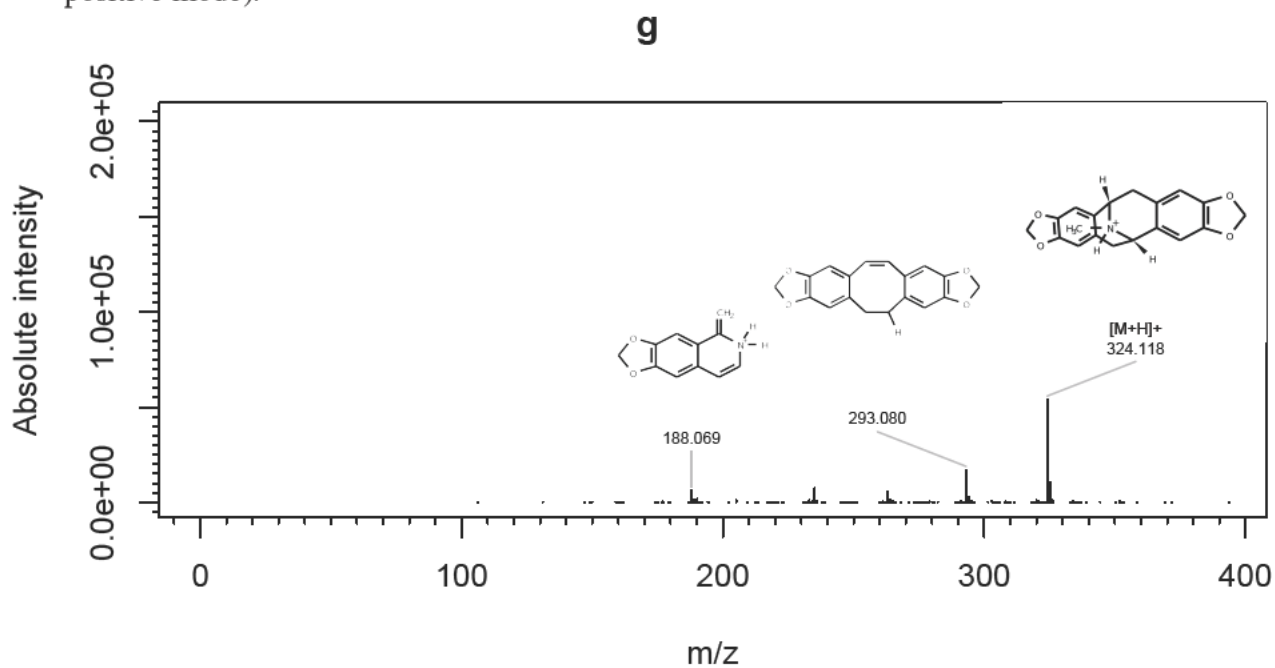


Figure A5. Challenge 5: annotated merged MS/MS spectra of reticuline (ESI, positive mode).

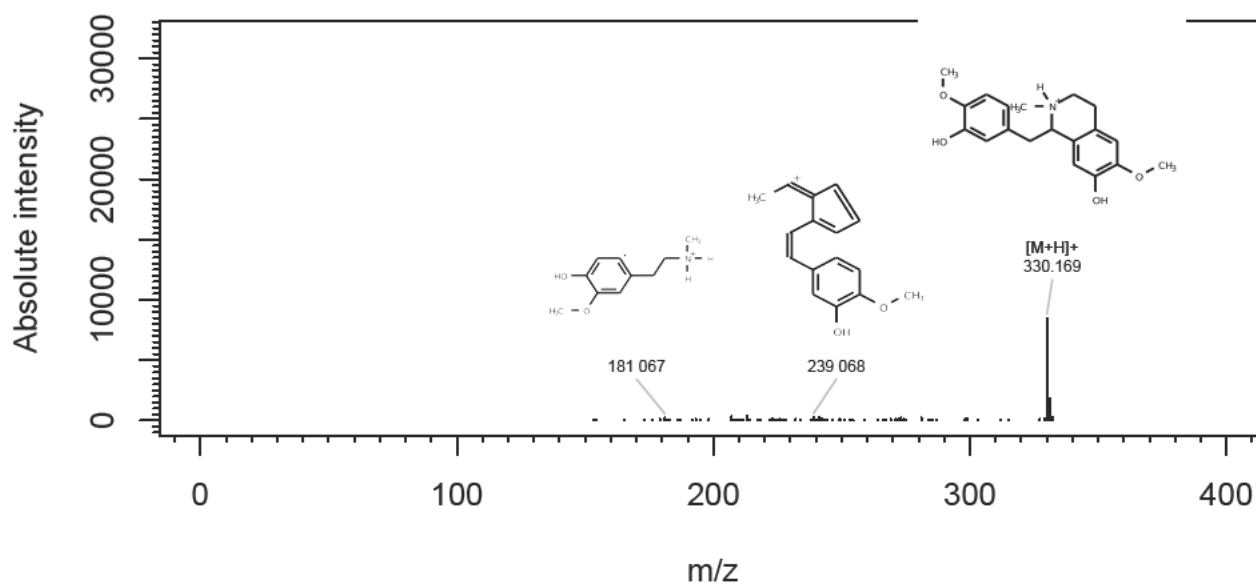


Figure A6. Challenge 6: annotated merged MS/MS spectra of rheadine (ESI, positive mode).

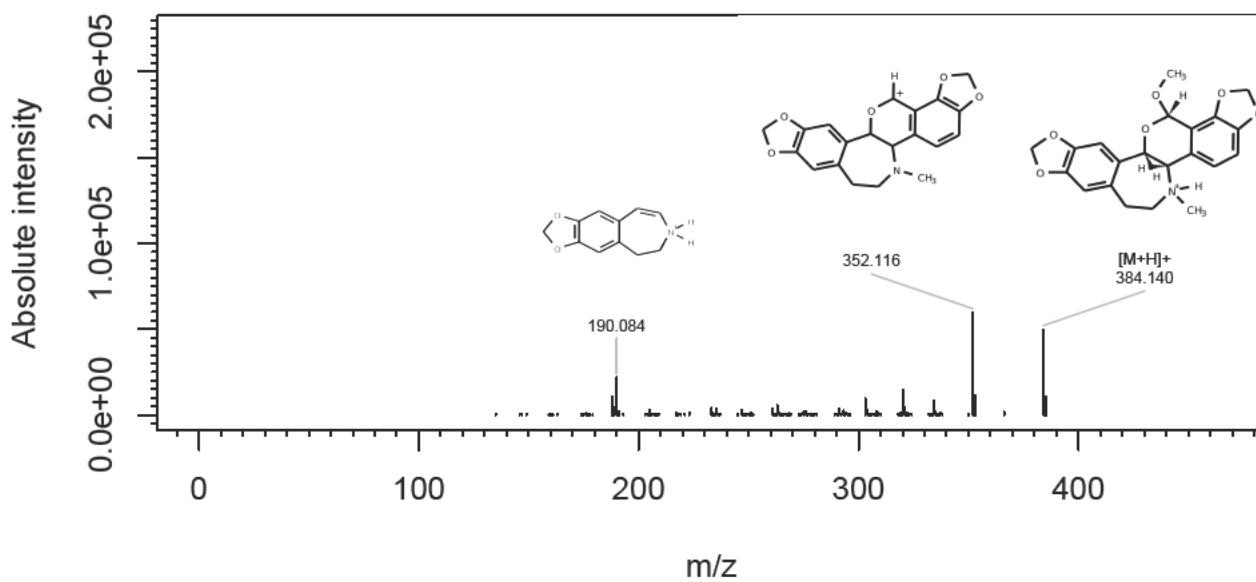


Figure A7. Challenge 10: annotated MS/MS spectrum of 1-aminoanthraquinone (ESI, positive mode).

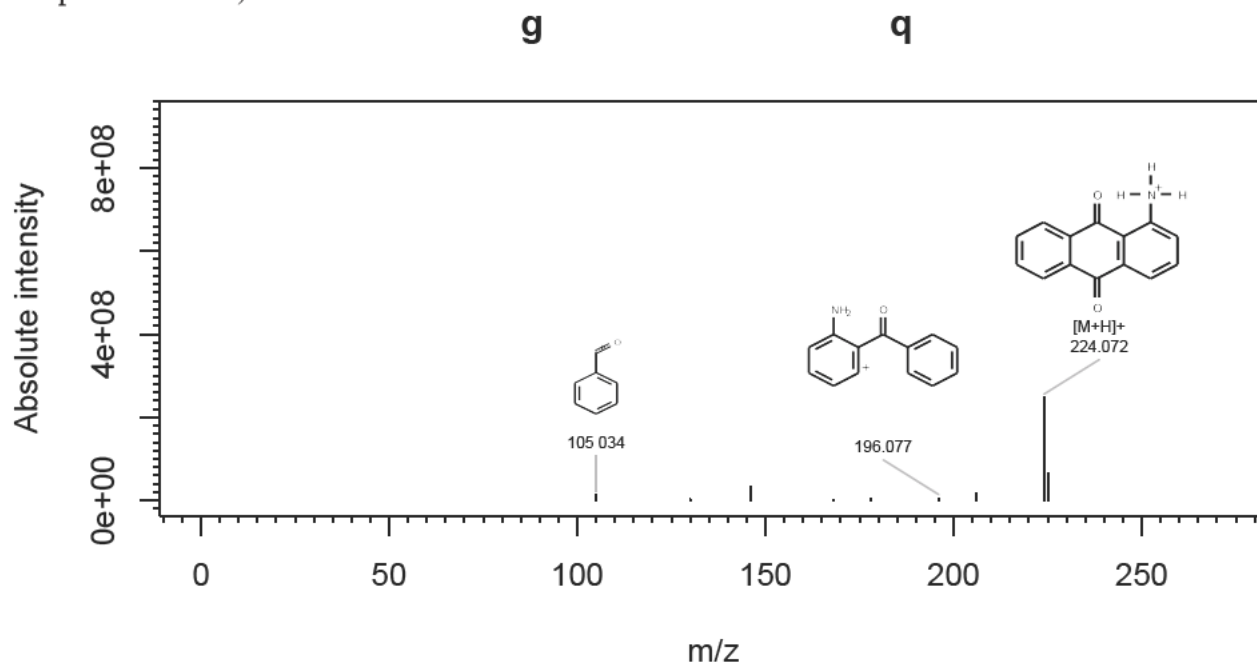


Figure A8. Challenge 11: annotated merged MS and MS/MS spectra of 1-pyrenemethanol (APCI, positive mode).

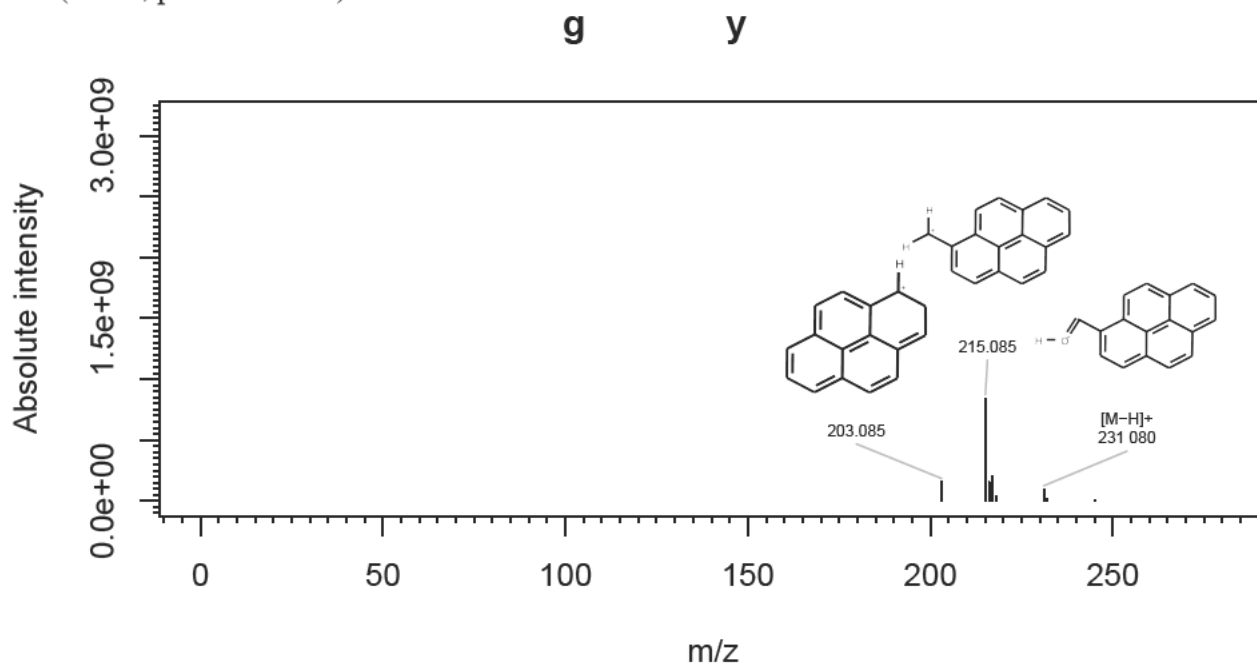


Figure A9. Challenge 12: annotated merged MS and MS/MS spectra of “Pigment Yellow 1” (APCI, positive mode).

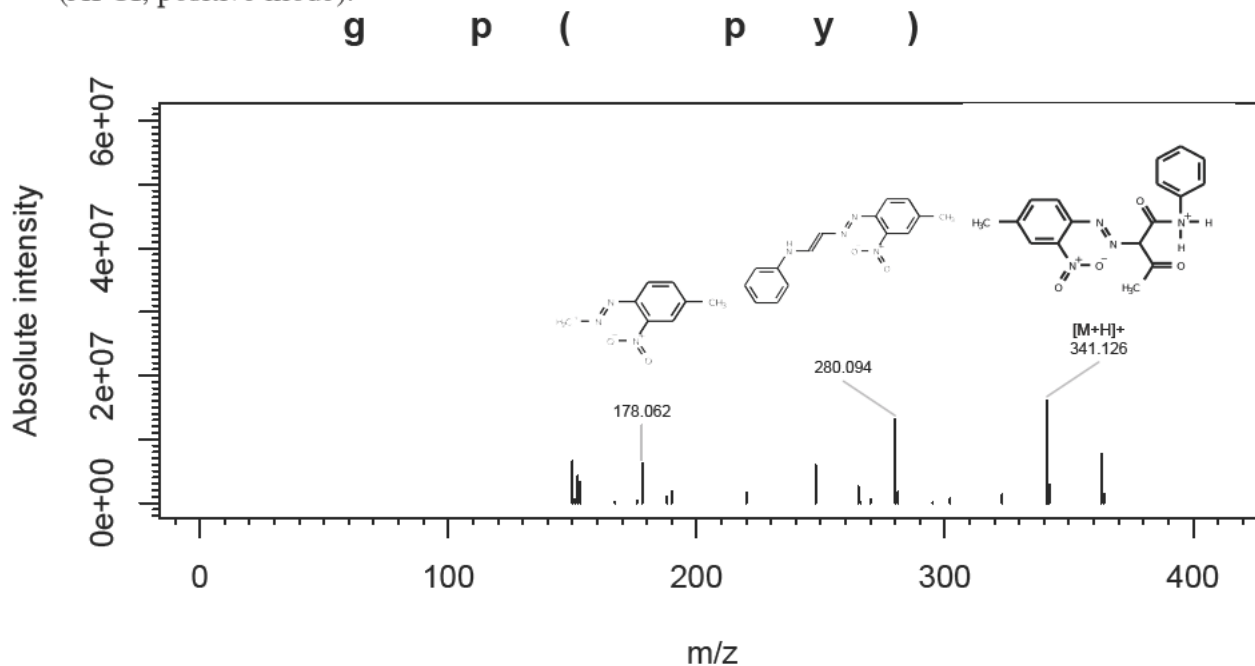


Figure A10. Challenge 13: annotated merged MS/MS spectra of benzyldiphenylphosphine oxide (ESI, positive mode).

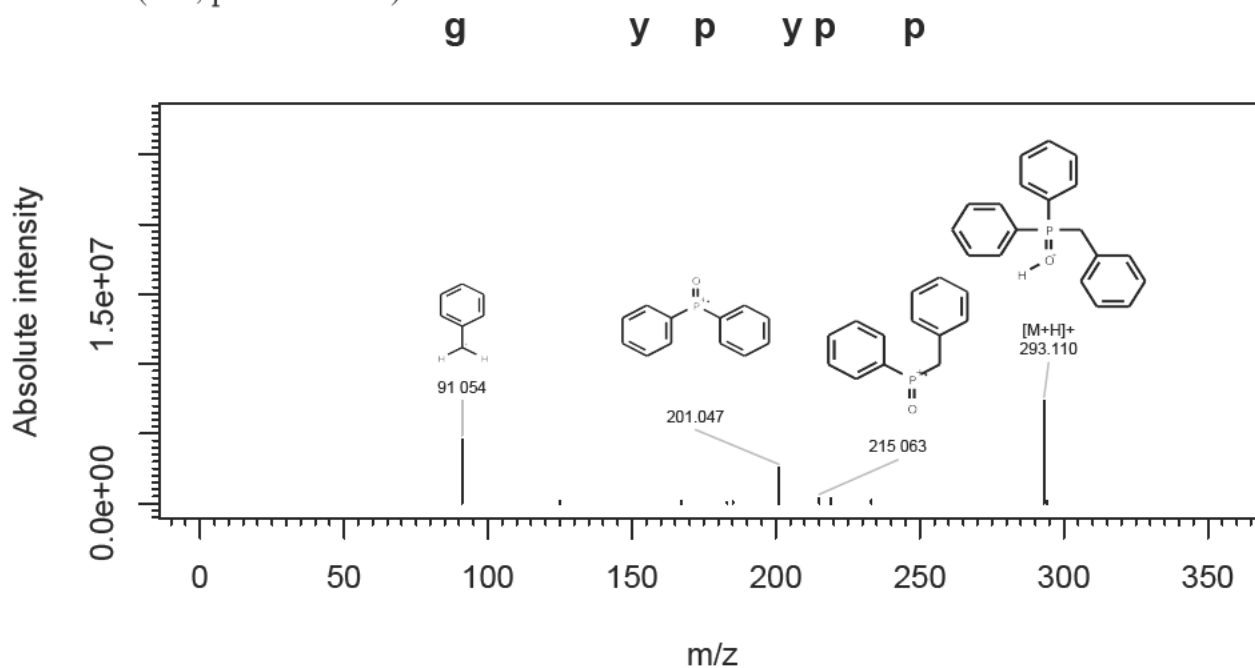


Figure A11. Challenge 14: annotated merged MS/MS spectra of 1H-benz[g]indole (APCI, positive mode).

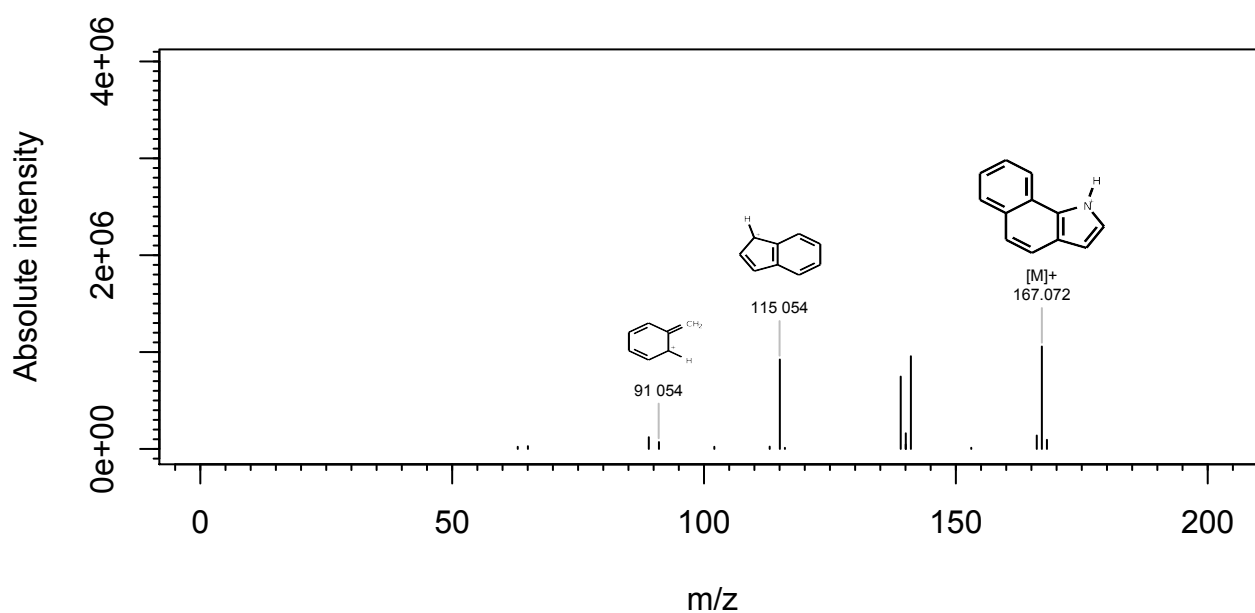


Figure A12. Challenge 15: annotated merged MS and MS/MS spectra of 1-isopropyl-5-methyl-1H-indole-2,3-dione (APCI, positive mode).

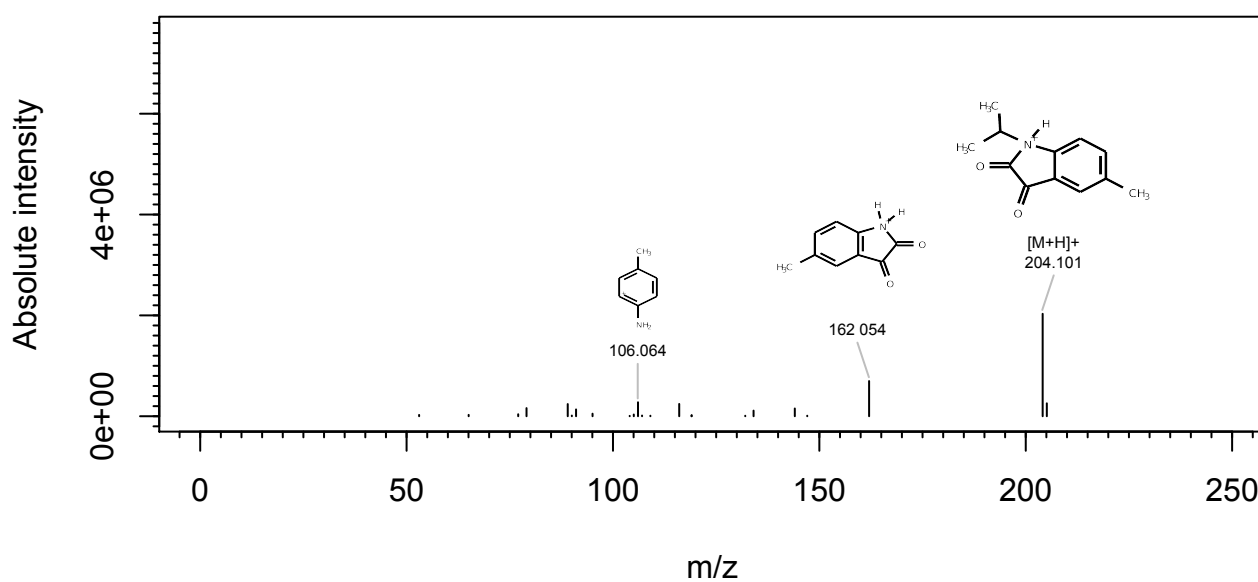


Figure A13. Challenge 16: annotated merged MS and MS/MS spectra of [1-(4-methoxyanilino)-1-oxopropan-2-yl] 6-oxo-1-propylpyridazine-3-carboxylate (APCI, positive mode). The $[M+H]^+$ ion was not observed in the measured spectra.

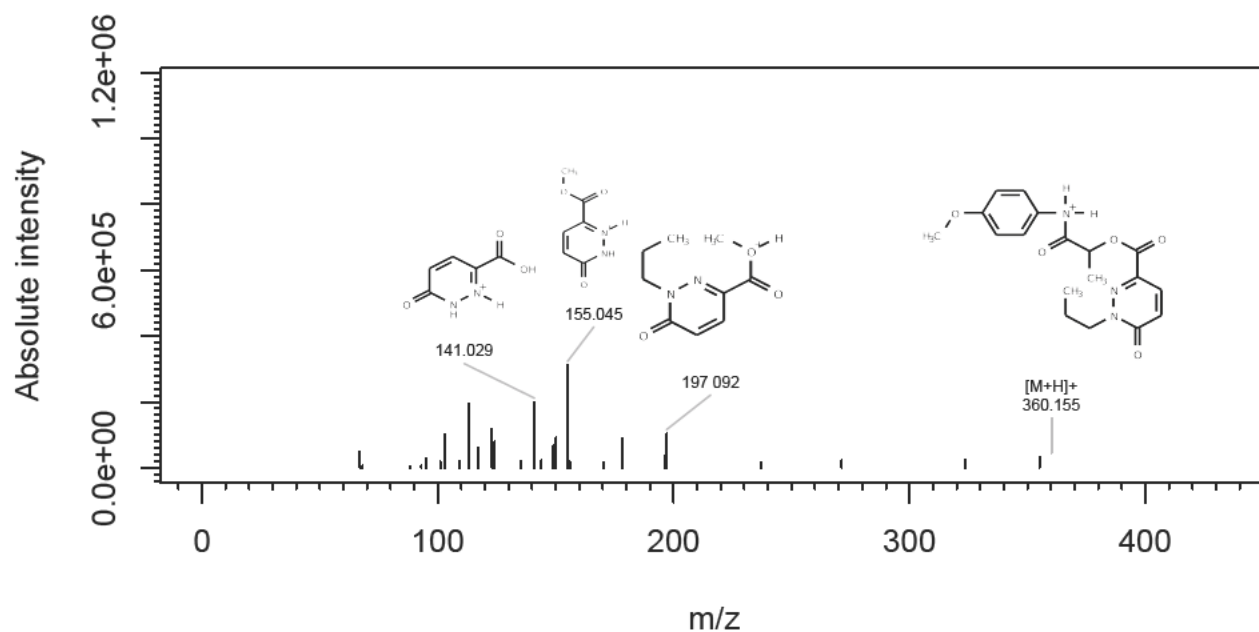
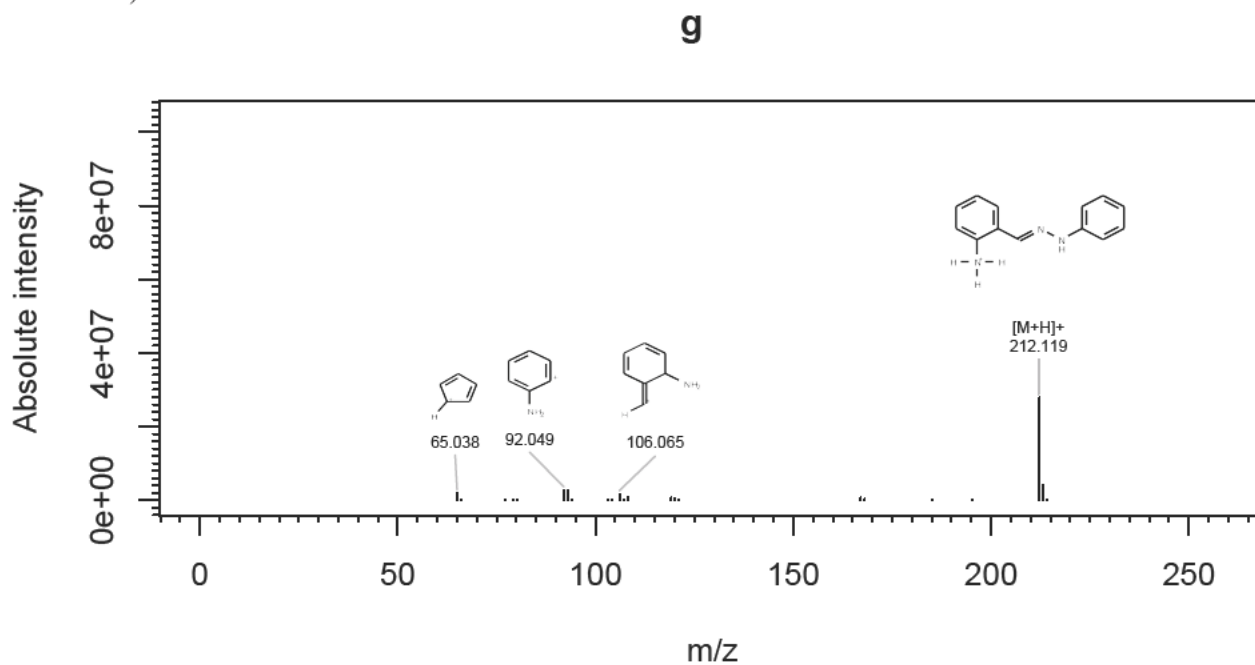


Figure A14. Challenge 17: annotated merged MS and MS/MS spectra of nitrin (ESI, positive mode).



B. Structures for the LC-HRMS challenges (Categories 1 and 2)

Figures B15 to B17, contain the structures and identifiers for the LC-HRMS challenges, Categories 1 and 2.

Figure B15. Structures and identifiers for LC-HRMS Challenges 1–4.

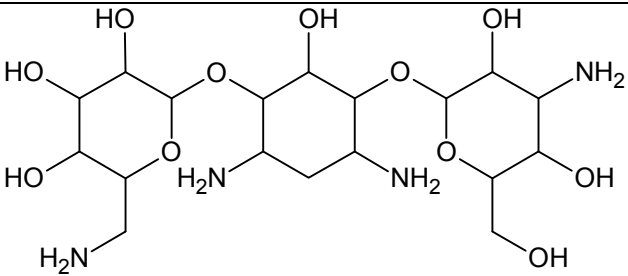
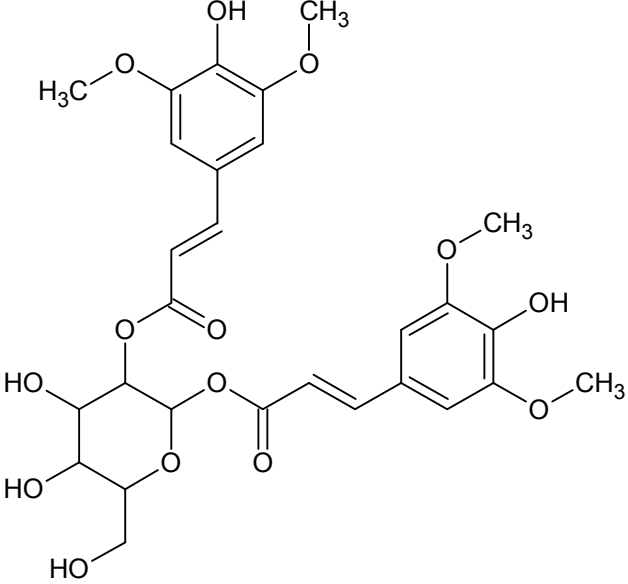
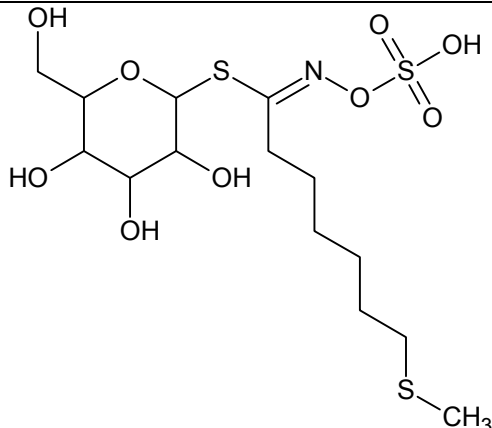
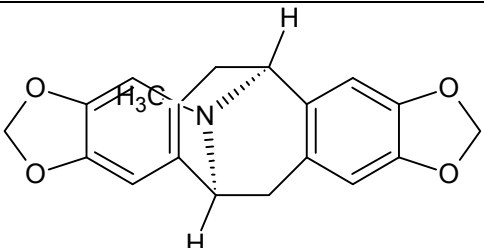
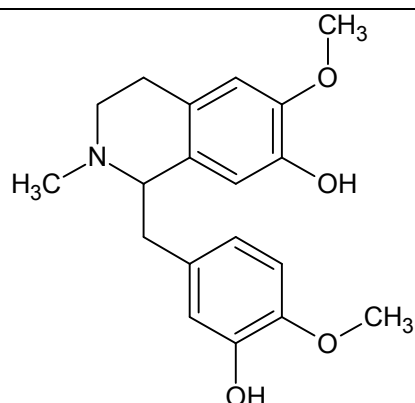
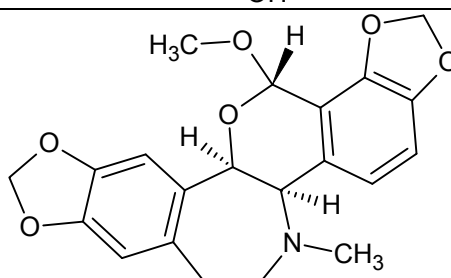
<p>Challenge 1 Kanamycin A $C_{18}H_{36}N_4O_{11}$ PubChem: 6032 ChemSpider: 5810</p>	
<p>Challenge 2 1,2-Bis-O-sinapoyl-beta-D-glucoside $C_{28}H_{32}O_{14}$ PubChem: 5280665 ChemSpider: 4444262</p>	
<p>Challenge 3 Glucosquerellin $C_{14}H_{27}NO_9S_3$ PubChem: 46173875 ChemSpider: NA</p>	
<p>Challenge 4 Escholtzine $C_{19}H_{17}NO_4$ PubChem: 12304178 ChemSpider: 16740500</p>	

Figure B16. Structures and identifiers for LC-HRMS Challenges 5–13.

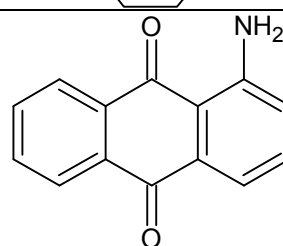
Challenge 5
Reticuline
 $C_{19}H_{23}NO_4$
PubChem: 10233
ChemSpider: 9816



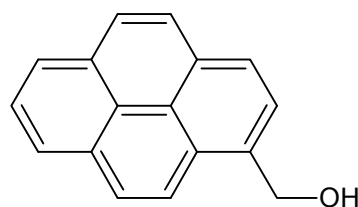
Challenge 6
Rheadine
 $C_{21}H_{21}NO_6$
PubChem: 197775
ChemSpider: 171184



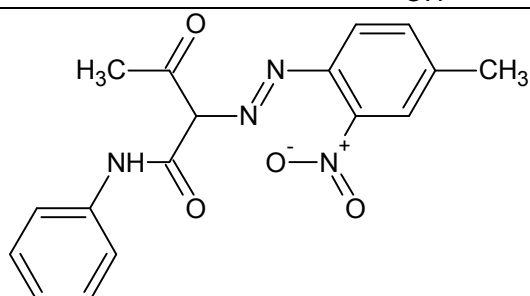
Challenge 10
1-Aminoanthraquinone
 $C_{14}H_9NO_2$
PubChem: 6710
ChemSpider: 6454



Challenge 11
1-Pyrenemethanol
 $C_{17}H_{12}O$
PubChem: 104977
ChemSpider: 94729



Challenge 12
alpha-(o-Nitro-p-tolyl
azo)acetoacetanilide
 $C_{17}H_{16}N_4O_4$
PubChem: 221491
ChemSpider: 192174



Challenge 13
Benzyl-diphenyl phosphine
oxide
 $C_{19}H_{17}OP$
PubChem: 76293
ChemSpider: 68772

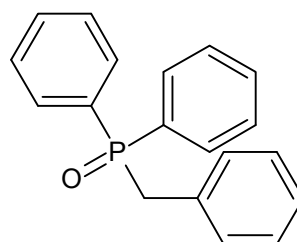
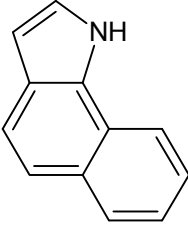
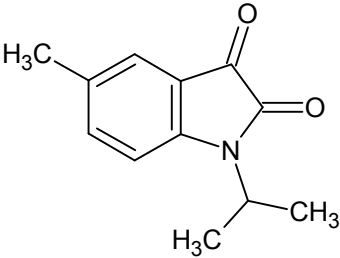
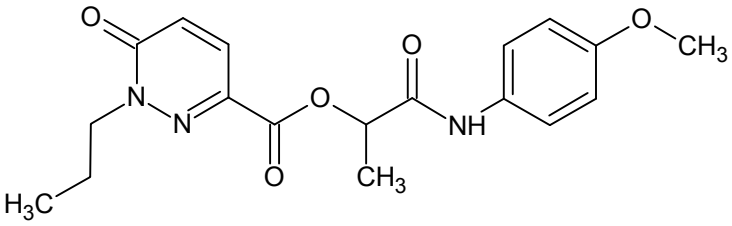
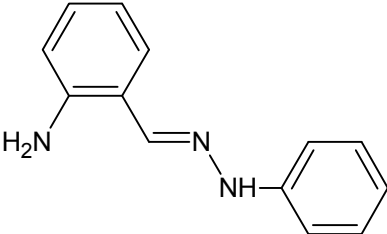


Figure B17. Structures and identifiers for LC-HRMS Challenges 14–17.

Challenge 14 1H-Benz[g]indole $C_{12}H_9N$ PubChem: 98617 ChemSpider: 89061	
Challenge 15 1-Isopropyl-5-methyl-1H-indole-2,3-dione $C_{12}H_{13}NO_2$ PubChem: 2145522 ChemSpider: 1606080	
Challenge 16 1-[(4-Methoxyphenyl) amino]- 1-oxo-2-propanyl 6-oxo-1- propyl-1,6-dihydro-3- pyridazinecarboxylate $C_{18}H_{21}N_3O_5$ PubChem: 18091616 ChemSpider: 16896706	
Challenge 17 Nitrin $C_{13}H_{13}N_3$ PubChem: 68380 ChemSpider: 61666	

C. Structures for GC-MS Challenges

Figure C18 contains the structures and identifiers for the GC-MS challenges, Categories 3 and 4.

Figure C18. Structures and identifiers for GC-MS Challenges 1–16.

Challenge 1 Phthalic anhydride $C_8H_4O_3$ PubChem: 6811 ChemSpider: 6552		Challenge 2 Phthalimide $C_8H_5NO_2$ PubChem: 6809 ChemSpider: 6550	
Challenge 3 2-Chlorobenzyl alcohol C_7H_7ClO PubChem: 28810 ChemSpider: 26799		Challenge 4 4-Chlorobenzyl alcohol C_7H_7ClO PubChem: 13397 ChemSpider: 12823	
Challenge 5 1,4-Dichlorobenzene $C_6H_4Cl_2$ PubChem: 4685 ChemSpider: 13866817		Challenge 6 Acenaphthene $C_{12}H_{10}$ PubChem: 6734 ChemSpider: 6478	
Challenge 7 4-Chlorobenzoic acid $C_7H_5ClO_2$ PubChem: 6318 ChemSpider: 6079		Challenge 8 Fluorene $C_{13}H_{10}$ PubChem: 6853 ChemSpider: 6592	
Challenge 9 Methyl 2-chlorobenzoate $C_8H_7ClO_2$ PubChem: 11895 ChemSpider: 11402		Challenge 10 2,4,6-Trichlorophenol $C_6H_3Cl_3O$ PubChem: 6914 ChemSpider: 21106172	
Challenge 11 Formothion $C_6H_{12}NO_4PS_2$ PubChem: 17345 ChemSpider: 16412		Challenge 12 alpha-Hexachloro-cyclohexane $C_6H_6Cl_6$ PubChem: 727 ChemSpider: 10468511	
Challenge 13 Dimethyl carbonotrithioate $C_3H_6S_3$ PubChem: 16840 ChemSpider: 15959		Challenge 14 O,O,O-Trimethyl thiophosphate $C_3H_9O_3PS$ PubChem: 9038 ChemSpider: 8686	
Challenge 15 Dibenzofuran $C_{12}H_8O$ PubChem: 568 ChemSpider: 551		Challenge 16 O,S,S-Trimethyl phosphorodithioate $C_3H_9PS_2O_2$ PubChem: 31435 ChemSpider: 29165	

References

1. Schymanski, E.L.; Neumann, S. CASMI website. Available online: <http://www.casmi-contest.org/> (accessed on 28 February 2013).
2. NIST/EPA/NIH. *NIST 2011 Mass Spectral Library*; National Institute of Standards and Technology, US Secretary of Commerce: Gaithersburg, Maryland, USA, 2011.
3. Schymanski, E.L.; Neumann, S. CASMI: And The Winner is *Metabolites* **2013**, *3*, 412–439.
4. ACD. *ACD/ChemSketch (Freeware) 12.00 (Version 12.01)*; Advanced Chemistry Development, Inc.: Toronto, Canada, 2010.
5. HighChem. *Mass Frontier Version 6.0*; HighChem/Thermo Scientific: Bratislava, Slovakia, 2013.
6. O’Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, doi:10.1186/1758-2946-3-33 .
7. R Development Core Team. The R Project. Available online: <http://www.r-project.org/> (accessed on 12 March 2013).
8. Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; *et al.* MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, *45*, 703–714.
9. Tautenhahn, R.; Böttcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinform.* **2008**, *9*, doi:10.1186/1471-2105-9-504.
10. Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T.R.; Neumann, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* **2012**, *84*, 283–289.
11. Neumann, S.; Thum, A.; Böttcher, C. Nearline acquisition and processing of liquid chromatography-tandem mass spectrometry data. *Metabolomics* **2013**, *9*, 84–91.
12. Wolfram, K.; Schmidt, J.; Wray, V.; Nimtz, M.; Milkowski, C.; Schliemann, W.; Strack, D. Profiling of phenylpropanoids in transgenic low-sinapine oilseed rape (*Brassica napus*). *Phytochemistry* **2010**, *71*, 1076–1084.
13. Brown, P.D.; Tokuhisa, J.G.; Reichelt, M.; Gershenzon, J. Variation of glucosinolate accumulation among different organs and developmental stages of *Arabidopsis thaliana*. *Phytochemistry* **2003**, *62*, 471–481.
14. Gallampois, C.M.G. Integrated Biological-Chemical Approach for the Identification of Polyaromatic Mutagens in Surface Waters. PhD thesis, Faculty of Mathematics, Informatics and Natural Sciences, RWTH Aachen, Germany, 2012.
15. Stravs, M.A.; Schymanski, E.L.; Singer, H.P.; Hollender, J. Automatic recalibration and processing of tandem mass spectra using formula annotation. *J. Mass Spectrom.* **2013**, *48*, 89–99.
16. Chiaia-Hernandez, A.C.; Krauss, M.; Hollender, J. Screening of lake sediments for emerging contaminants by liquid chromatography atmospheric pressure photoionization and electrospray ionization coupled to high resolution mass spectrometry. *Environ. Sci. Technol.* **2013**, *47*, 976–986.
17. The MassBank Consortium. MassBank Mass Spectral Database. Available online: <http://www.massbank.jp/> (accessed on 09 January 2013).

18. Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinform.* **2010**, *11*, doi:10.1186/1471-2105-11-148.
19. Shaffer, C.J.; Schröder, D.; Alcaraz, C.; Žabka, J.; Zins, E.L. Reactions of doubly ionized benzene with nitrogen and water: A nitrogen-mediated entry into superacid chemistry. *Chem. Phys. Chem.* **2012**, *13*, 2688–2698.
20. Schymanski, E.; Gallampois, C.; Krauss, M.; Meringer, M.; Neumann, S.; Schulze, T.; Wolf, S.; Brack, W. Consensus structure elucidation combining GC/EI-MS, structure generation, and calculated properties. *Anal. Chem.* **2012**, *84*, 3287–3295.
21. Meinert, C.; Schymanski, E.; Kuster, E.; Kuhne, R.; Schuurmann, G.; Brack, W. Application of preparative capillary gas chromatography (pcGC), automated structure generation and mutagenicity prediction to improve effect-directed analysis of genotoxicants in a contaminated groundwater. *Environ. Sci. Pollut. Res.* **2010**, *17*, 885–897.
22. Schymanski, E.L.; Meinert, C.; Meringer, M.; Brack, W. The use of MS classifiers and structure generation to assist in the identification of unknowns in effect-directed analysis. *Anal. Chem. Acta* **2008**, *615*, 136–147.

© 2013 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).