Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software

Hiroshi Tsugawa, RIKEN Center for Sustainable Resource Science, Yokohama, Japan; RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

© 2019 Elsevier Inc. All rights reserved.

1 Why Do We Need Computational Mass Spectrometry?

Mass spectrometry-based metabolomics has the potential for high-throughput screening of natural products.[1–3] Nowadays, liquid chromatography coupled with high resolution tandem mass spectrometry (LC-MS/MS) is a popular technique owing to (1) the scalability of electrospray ionization (ESI) covering a large variety of chemical properties for metabolite ionization, (2) the high mass accuracy for reliably predicting molecular formulae of unknown natural products using the accurate *m/z* values of precursors, (3) the information-rich mass fragmentation (recorded as MS/MS spectrum) giving us the substructure information of metabolites, and (4) the availability of rich metabolite-MS/MS spectral libraries.[4] Therefore, decoding mass spectrometry data facilitates the dereplication of natural products as well as the discovery of novel chemical structures.[5] This chapter focuses on the annotation of natural products in LC-MS/MS based metabolomics.

The Metabolomics Standard Initiative (MSI) currently defines five different levels for reporting the confidence of metabolite annotations (Table 1)[6]: Level 0 is achieved with nuclear magnetic resonance (NMR) coupled with the guidelines for natural products, and level 1–4 assignments are achieved with MS-based metabolomics workflow. However, less than 100 natural products can be annotated as level 1 (standard confirmed) due to the lack of authentic standards in a laboratory. Since it is not realistic to measure all standard compounds for natural products to obtain their retention time and MS/MS spectra, developing the methodology for level 2 and 3 characterizations is an emerging need for screening natural products.

One of the goals of ‘computational mass spectrometry’ for metabolomics is to enhance the coverage of compound annotations defined as level 2 and 3 by using the informatics techniques in mass spectral database search, de novo formula prediction, and in silico structure elucidation using quantum chemistry, machine learning and cheminformatics.[5,6] I highlight the functions implemented in MS-DIAL[7,8] and MS-FINDER[9] software programs, although there are many alternative programs for the metabolite annotations such as MzMine,[10] XCMS family,[11] Sirius,[12] and CSI: FingerID.[13] The purpose of this chapter is to describe the practical workflow for natural product characterizations which includes (1) evaluating ion characteristics to reduce false positive identification,[14] (2) searching mass spectral databases,[15] (3) molecular formula predictions,[9,12] (4) searching structures followed by ranking the candidates,[9,12] (5) metabolite class predictions,[16] and (6) using molecular spectral networks for elucidating mass spectra.[17] The workflow is showcased with some examples of natural product annotation. MS-DIAL and MS-FINDER are freely available at RIKEN PRIMe website. (http://prime.psc.riken.jp/).

2 Evaluating the Mass Signals Before Further Annotations

Unless authentic standard compounds are available, the evaluation of precursor ion ‘characters’ is important to avoid unrewarded efforts and false positive identification. Fig. 1A shows the MS^1 spectra of 8-methylsulfinyloctyl glucosinolate (also known as glucohirsutin) in positive and negative ion modes. The glucohirsutin-derived ion is clearly detected as proton loss ion {[M-H]$^-$, *m/z* 492.1028, $C_{16}H_{30}NO_{10}S_3^-$} in negative ion mode. On the other hand, the protonated ion {[M+H]$^+$, *m/z* 494.1178,

https://doi.org/10.1016/B978-0-12-409547-2.14645-1

Table 1 Confidence levels of compound annotations defined by the working group of the Metabolomics Standard Initiative (MSI) at the 2017 annual meeting (Brisbane, Australia).

Levels	*Description*
Level 0	Confident 3D structure which follows natural product guidelines: isolated, pure compound, defined the full stereochemistry
Level 1	Confident 2D structure confirmed by the authentic standard confirming the consistency of retention time and MS/MS spectrum
Level 2	Putatively annotated structures matched to literature data, mass spectral databases, or other diagnostic evidences in MS/MS spectrum
Level 3	Putatively characterized compound class (unreached to structure level annotation) requiring at least one piece of metabolite information
Level 4	Unknown compounds of interest present in biological samples

The description is assumed as LC-MS/MS-based metabolomics in this article.

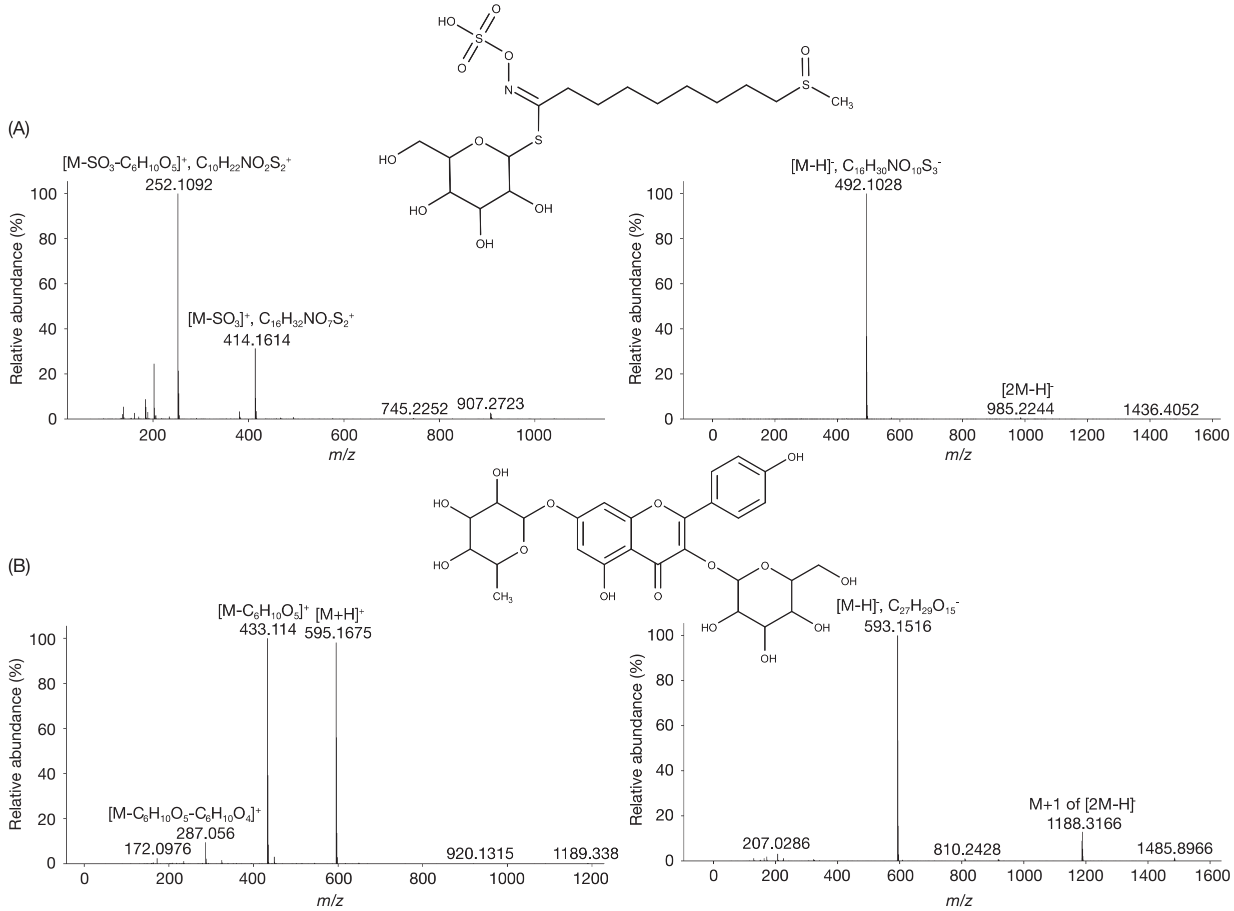

Fig. 1 The survey MS[1] spectrum, the spectrum of ions which were detected at the same retention time, to describe the importance of in-source fragment annotations. (A) The MS[1] spectra for 8-methylsulfinyloctyl glucosinolate in positive (left) and negative (right) ion modes. (B) The MS[1] spectra for kaempferol-3-*O*-glucoside-7-*O*-rhamnoside in positive (left) and negative (right) ion modes. M represents the neutralized form of compounds. These observations indicate that a metabolite is ionized into various forms such as a cluster ion [2M-H]$^-$ and in-source fragment ions.

$C_{16}H_{32}NO_{10}S_3^+$} is rarely detected in positive ion mode (<0.1% relative ion intensity in the MS^1 spectrum), while the degraded ions such as *m/z* 414.1614 {$C_{16}H_{32}NO_7S_2^+$, [M-SO$_3$]$^+$} and *m/z* 252.1092 {$C_{10}H_{22}NO_2S_2^+$, [M-SO$_3$-C$_6$H$_{10}$O$_5$]$^+$} which are fragmented in ESI ion source, also known as 'in-source fragment ions,' are strongly detected in a conventional mass spectrometer condition. Another example is shown by the MS^1 spectra of kaempferol-3-*O*-glucoside-7-*O*-rhamnoside detected at *m/z* 595.1675 in positive ion mode (Fig. 1B), where the degraded ions such as *m/z* 433.115 and *m/z* 287.0555 were also detected. These observations emphasize the importance of confirming both positive and negative ion mode data to increase the confidence in precursor ion 'characteristics' (original form or degraded form) and the adduct types: the determination of adduct forms is essential for the following annotation steps. Importantly, the degraded ions are often annotated by MS/MS spectral library searches; for example, kaempferol-3-*O*-glucoside detected at *m/z* 433.115, which is an intermediate metabolite for kaempferol-3-*O*-glucoside-7-*O*-rhamnoside in flavonoid biosynthesis.[18] It indicates that these degraded ions may be identified as 'false positive' ions by mass spectral searching processes if the precursor ion is not assigned as 'in-source fragment ion' appropriately. Therefore, the metabolite ion features should be evaluated carefully before proceeding to further annotation processes.

Below is the 'checkpoint' to evaluate the precursor ion properties.

1. Observe both MS^1 spectra of positive and negative ion modes at the same retention time area.
2. The adduct types can be estimated if the intelligible difference between ESI (+) and ESI (−) precursor ions is found. For example, the adduct types are estimated as $[M+H]^+$ and $[M\text{-}H]^-$ if 2.015 Da mass shift is observed. This is further explained in showcase section.
3. To double check the adduct forms or to determine the adduct types when either ESI (+) or ESI (−) ion cannot be detected, the mass-pairing method which checks the mass differences in MS^1 spectrum during an ionization experiment is useful. For example, the adduct forms can be determined as $[M\text{-}H]^-$ and $[M+HCOO]^-$ if the mass difference of 46 Da is observed in negative ion mode. Furthermore, the adduct type can also be determined by checking the MS/MS spectrum pattern. For example, the neutral loss of 46 Da is also observed for $[M + HCOO]^-$ type (Fig. 2A).
4. Consider the possibilities of dimers and in-source fragment ions. For example, the ion signal of *m/z* 1187.31 is interpreted as $[2M\text{-}H]^-$ for kaempferol-3-*O*-glucoside-7-*O*-rhamnoside {*m/z* 593.1516, $[M\text{-}H]^-$} as shown in Fig. 1B. To check in-source fragment ions, the most useful criterion is to confirm the MS/MS spectrum of precursor ions with higher *m/z*. If the *m/z* value of focused precursor ion is also observed in the MS/MS spectral peaks of a precursor ion with higher *m/z*, the focused ion may be in-source fragment ion derived from the precursor ion with higher *m/z*. Fig. 2B shows the MS/MS spectrum of kaempferol-3-*O*-glucoside-7-*O*-rhamnoside in positive ion mode. For example, the precursor ions with *m/z* 287.056 and 433.114 in Fig. 1B are recognized as in-source fragment ions because the same *m/z* values are also observed in the product ions of kaempferol-3-*O*-glucoside-7-*O*-rhamnoside.

Although the evaluation of mass ion signals requires the skills of mass spectrometry data analysis, many programs such as MS-DIAL,[7] XCMS,[11] and MzMine[10] can currently assist the curations. For example, the MS-DIAL software, which supports all data processing procedures for GC-MS, GC-MS/MS, LC-MS, LC-MS/MS, and data-independent acquisition data evaluates the following five characteristic features for each of the mass ion signals: (1) monoisotopic or isotopic ion, (2) adduct ion form, (3) chromatogram peak shape similarity among the peaks at the same retention time area, (4) inclusion relationship if the peak is included as a part of MS/MS spectrum of precursor ions with higher *m/z*, and (5) ion profile correlation among biological samples, where the hypothesis is that the ions having similar metabolic profiles are derived from the same metabolite origin (Fig. 3). See the details of MS-DIAL at http://prime.psc.riken.jp/. When the peak of interest is characterized as a 'true positive' metabolite, i.e., not different adduct types of known metabolites and not in-source fragment ions, the peak is further annotated by the following steps. Again, it is stated that 90% of ions detected in LC-MS are currently recognized as artificial noise or redundant ions derived from different adduct forms and in-source fragment ions of metabolites.[19] Therefore, the 'ion curation' procedure is a crucial step not only in natural product research studies but also in metabolomics studies.

3 Searching Mass Spectral Databases

Searching mass spectral databases is the most reliable method for level 2 annotations,[15] which can also be performed in MS-DIAL and MS-FINDER. Currently, more than 250,000 MS/MS spectra of 12,000 unique structures are available in publicly and commercially available databases (Table 2). Practically, 50–100 metabolites can be annotated by using a conventional scoring method that incorporates a precursor *m/z* filtering step followed by MS/MS similarity matching algorithm. Moreover, when the metabolite generating the query spectrum is not present in the databases, the state-of-the-art programs provide the mathematical algorithm for searching metabolites and generating a partially 'similar' spectrum even if the precursor *m/z* values are different.[17,20,21] The hypothesis is that the structures sharing the same substructures (motifs) would generate the same shared product ions and/or neutral losses. For instance, (1) *m/z* 78.9585 and *m/z* 96.9595 are detected in phosphate (PO_3^-)- and sulfate (HSO_4^-)-containing metabolites, respectively; (2) neutral losses of 146.0579 ($C_6H_{10}O_4$), 162.05282 ($C_6H_{10}O_5$), and 176.0321 ($C_6H_8O_6$) are monitored in metabolites containing deoxyhexose (dHex), hexose (Hex), and uronic acid (HexA) moieties, respectively; (3) aglycone ions of the flavonol class, for example, kaempferol, quercetin, and isorhamnetin, are usually detected by their specific *m/z* values of 285.0405 ($C_{15}H_9O_6^-$), 301.0354 ($C_{15}H_9O_7^-$), and 315.051 ($C_{16}H_{11}O_7^-$), respectively. The methodology facilitates the discovery of modified metabolites derived from classical metabolic pathways. In fact, discoveries of novel derivatives of natural products have been accelerated in Global Natural Products Social Molecular Networking (GNPS) which is a crowd sourced community-wide organization established for sharing of mass spectrometry data (http://gnps.ucsd.edu).[22] The network approach that utilizes mass spectral similarity has been found to be useful to discover unknown molecules, which are introduced in later section. The annotation process is further refined by the following steps that include molecular formula prediction and structure elucidation by searching structure databases.

4 Molecular Formula Prediction

The spectrum-oriented metabolite identification has a serious drawback of database coverage. For example, the MS/MS spectra for only 12,000 structures have been registered in public and commercial databases while the dictionary of natural products (DNP) currently incorporates more than 300,000 structures of natural products.[23] Therefore, many metabolite ions are marked as

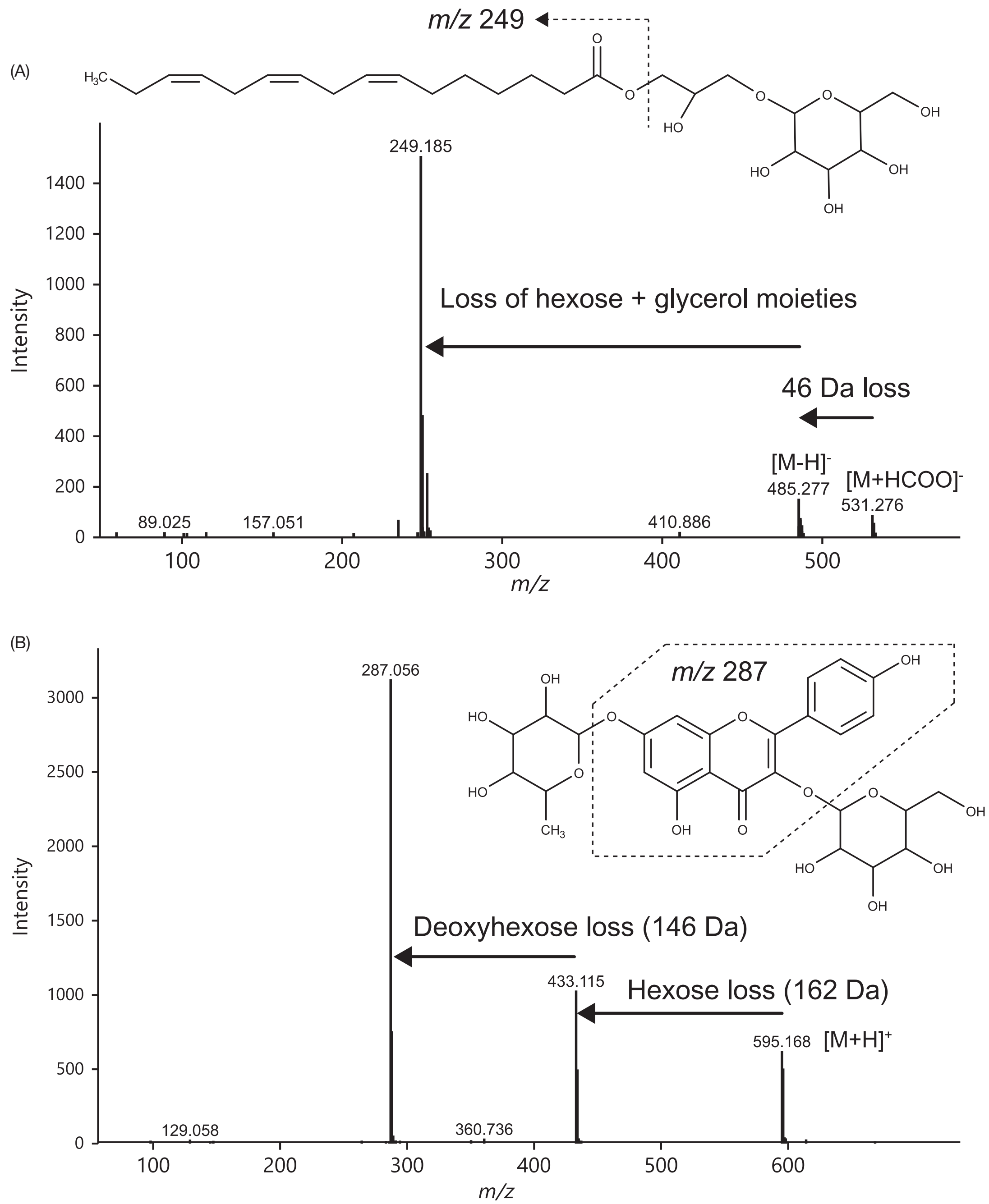

Fig. 2 Analysis of MS/MS spectra to demonstrate the peak characteristic estimations for a targeted precursor ion. (A) The MS/MS spectrum of monogalactosyl monoacylglycerol (MGMG) 16:3. The neutral loss of 46 Da was commonly observed in the metabolite ion of formate adduct $[M + HCOO]^-$ when formic acid was added to the mobile phase of LC-MS analysis. Moreover, *m/z* 249.185 represented fatty acid ion of 16:3 (16 carbons and 3 double bonds in the acyl chain), and the sugar lipid species could be detected as the adduct ion forms, such as $[M\text{-}HCOO]^-$ in negative mode and $[M + NH_4]^+$ in positive ion mode due to the low acidity of hydrogens in the molecule. These indirect observations enable us to know the adduct form and substructures of the targeted precursor ion. (B) The MS/MS spectrum of kaempferol-3-*O*-glucoside-7-*O*-rhamnoside is depicted while the survey MS^1 spectrum is depicted in Fig. 1B. Since the product ions with *m/z* 433 and *m/z* 287 were also detected in the survey MS^1 spectrum, the ions with *m/z* 433 and *m/z* 287 in the MS1 spectrum depicted in Fig. 1B can be interpreted as the in-source fragment ions of kaempferol-3-*O*-glucoside-7-*O*-rhamnoside.

'unknowns' due to the lack of mass spectrum information. Computational metabolomics attempts to fill the large 'gap' between spectrum and structure counts.

The first goal of annotating unknown natural products is to determine the molecular formula. The MS-FINDER program calculates the possible molecular elements from the precursor *m/z* and adduct type information (Table 3). The determination of adduct forms is assisted by MS-DIAL program as described above. This information is subsequently directly linked to MS-FINDER

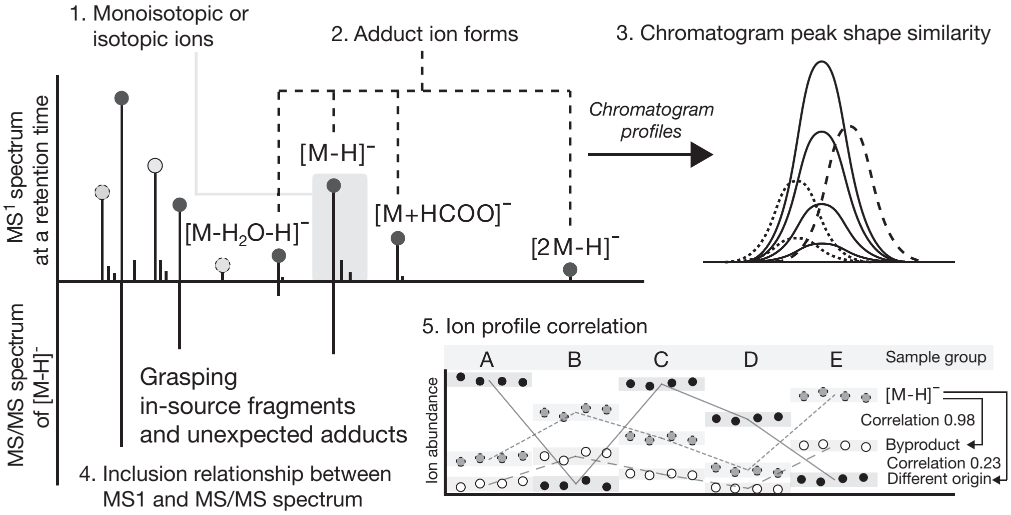
1. Monoisotopic or isotopic ions
2. Adduct ion forms
3. Chromatogram peak shape similarity
Chromatogram profiles
MS1 spectrum at a retention time
[M-H]⁻
[M-H2O-H]⁻
[M+HCOO]⁻
[2M-H]⁻
MS/MS spectrum of [M-H]⁻
Grasping in-source fragments and unexpected adducts
4. Inclusion relationship between MS1 and MS/MS spectrum
5. Ion profile correlation
A
B
C
D
E
Sample group
Ion abundance
[M-H]⁻
Correlation 0.98
Byproduct
Correlation 0.23
Different origin

Fig. 3 The summary of ion properties characterized in MS-DIAL software. (1) The ions are divided into two groups. One belongs to monoisotopic ions and another belongs to isotopic ions. (2) The adduct ion forms are evaluated by checking the mass difference between *m/z* values. The adduct ion forms are further evaluated by integrating the information of positive and negative ion mode data. (3) The ions having similar chromatographic profiles at the same retention time are linked. (4) The ion in the MS^1 spectrum is assigned as "in-source fragment candidate" if the same *m/z* peak is observed in the MS/MS spectrum of ion with higher *m/z*. (5) The ions having similar metabolic profiles among biological samples at the same retention time are linked, whereas the ions having high correlation coefficient values can be used for an additional diagnostic criterion for the in-source fragment annotations. Modified from Advances in Computational Metabolomics and Databases Deepen the Understanding of Metabolisms. *Curr. Opi. Biotechnol.*, **2018**, *54*, pp. 10–17.

Table 2 Statistics of publicly available spectral databases in 2018.

Name	*Ion mode*	*Compound*	*Spectrum*
MassBank	Positive	1316	8068
MassBank-EU	Positive	179	710
ReSpect	Positive	560	2737
GNPS	Positive	5216	8782
Fiehn HILIC	Positive	954	1701
CASMI2016	Positive	399	440
MetaboBASE	Positive	3	8
PlaSMA	Positive	356	4439
MassBank	Negative	1085	4782
MassBank-EU	Negative	51	100
ReSpect	Negative	443	1573
GNPS	Negative	1818	2351
Fiehn HILIC	Negative	776	1341
CASMI2016	Negative	163	178
MetaboBASE	Negative	258	1151
PlaSMA	Negative	294	4216
RIKEN OxPLs	Negative	386	386

The counts for recorded spectra and compounds are listed for each database, project, or repository.

program package. Practically, the molecular formula determination requires 10 ppm mass accuracy (5 mDa at *m/z* 500) and more than 30,000 mass resolution which can be achieved in a conventional time-of-flight mass spectrometer: of course, better MS accuracy and resolution increases the confidence in metabolite annotations.[24]

Candidate formulae are computationally generated from the precursor *m/z* and the adduct type for a user-defined mass tolerance (10 mDa or 20 ppm tolerances are often used for QTOF-MS data). The candidates are then filtered out by two diagnostic criteria, including the valence rule and the empirical rule.[9,25] The valence rule checks if the elemental composition meets the atom-valence principles (for example, carbon has four connectable bonds) or not, and it can simply be evaluated by two equations as (1) OV or SV is even and (2) SV is greater than or equal to 2 * (TA-1) where OV, SV, and TA mean the total number of atoms having odd valences, the sum of valences, and the total number of atoms, respectively. In addition, MS-FINDER applies the empirical elemental ratio filter, such as the hydrogen/carbon balance, which is generated using the statistics from available metabolome structure databases. According to the statistics, 99.85% of the molecular formulae in structure databases meet the equations of $H/C < 3.33$,

Table 3 Major elemental compositions to predict molecular formulae in metabolomics.

Element name	*Abbreviation*	*Monoisotopic mass*
Carbon	C	12
Hydrogen	H	1.007825032
Nitrogen	N	14.003074
Oxygen	O	15.99491462
Sulfur	S	31.972071
Phosphorus	P	30.97376163
Fluorine	F	18.99840322
Chlorine	Cl	34.96885268
Bromine	Br	78.9183371
Iodine	I	126.904473
Silicon	Si	27.97692653
Electron	e-	0.00054858

$N/C < 1.2$, $O/C < 2.2$, $P/C < 0.4$, $S/C < 1.0$, $O/P < 19.0$, and $O/S < 28.0$. This empirical filter drastically reduces the candidate formulae (Table 4).

The formula candidates are ranked by several diagnostic criteria. First, the mass error between the experimental and theoretical *m*/*z* values is evaluated. Second, the isotope ratio is also evaluated with the theoretical isotopic profile. For example, the metabolite containing Cl and/or Br halogen elements provides the unique isotopic patterns in $M + 2$ region, where M represents monoisotopic ion: the natural abundances of ^{37}Cl and ^{81}Br are 24.2% and 49.3% compared to those of ^{35}Cl and ^{79}Br, respectively. The precise observation of $M + 2$ isotopic region allows us to determine the existence or even the content details of nitrogen, oxygen, and sulfur which can realistically be achieved by using ultrahigh resolution mass spectrometer (>300,000). The MS/MS spectrum is also utilized for the diagnostics. In principle, the product ions from a metabolite can be explained by the partial molecular formulae of precursor formula, and therefore, the fulfillment of submolecular formulae to the MS/MS spectrum is more accurate in the candidate having correct molecular formula as compared to others. Finally, and most practically, the molecular formula which has already been reported in biology literatures is more plausible than others that have not yet been discovered. The above criteria are scored in MS-FINDER software which can achieve 98.0% prediction accuracy when the search space is restricted to CHNOPS elements for known structure databases.[9]

For natural products containing many unknown-unknowns which have never been reported,[26] the use of stable isotope-labeled metabolites or organisms is a powerful tool to determine the elements in molecular formula.[27] Recently, we have provided the methodology for comprehensive carbon content determinations of unknown-unknowns, where the LC-MS data set of ^{13}C-labeled and nonlabeled plants was analyzed by MS-DIAL and MS-FINDER software programs.[16] Here, the MS-DIAL program requires two LC-MS data sets, one for ^{13}C-labeled plant and another for nonlabeled plant. The carbon number is identified by using the isotope grouping function (Fig. 4). Importantly, there are many false positive isotope groups which are regarded as in-source fragment ions or unexpected adduct ions, and therefore, the procedure for ion character determination is needed with careful consideration. We demonstrated that the carbon contents of 2113 and 1491 metabolite ions in positive and negative ion mode have been identified after all isotope groups were validated with the MS-DIAL graphical user interface. Moreover, the MS-FINDER program can also import the determined elemental information which is written in the input formats (MSP and MAT) for MS-FINDER to generate molecular formula candidates: the prediction accuracy exceeded 99.8% by utilizing C-element information and restricting to CHNOSP elements.

Table 4 Statistics for elemental compositions in molecular formulae recorded in metabolome structure databases.

	Mass	*C*	*H*	*N*	*O*	*P*	*S*	*H/C*	*N/C*	*O/C*	*P/C*	*S/C*	*O/P*	*O/S*
Max	1999.898	122.00	236.00	40.00	65.00	8.00	12.00	8.00	4.00	10.00	3.00	6.00	53.00	41.00
Min	50.01565	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mean	601.0329	29.80	42.01	2.06	9.05	0.12	0.25	1.38	0.10	0.32	0.01	0.02	0.56	1.03
Stdev	349.2922	17.83	30.32	3.04	8.36	0.49	0.67	0.43	0.15	0.27	0.04	0.09	2.34	3.40
Median	512.3229	26.00	34.00	1.00	7.00	0.00	0.00	1.39	0.04	0.27	0.00	0.00	0.00	0.00
Percentile 0.005%	52.01739	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Percentile 0.15%	72.05377	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Percentile 99.85%	1939.743	95.00	164.00	20.00	50.00	3.00	6.00	3.33	1.20	2.20	0.40	1.00	19.00	28.00
Percentile 99.995%	1998.1	116.74	220.00	32.00	60.00	7.74	8.00	6.00	4.00	6.00	1.83	3.00	49.00	35.74

A total of 90,227 molecular formulae reported in the previous research [https://pubs.acs.org/doi/full/10.1021/acs.analchem.6b00770] were used for the statistics. For example, hydrogen over carbon ratio (H/C) for 99.7% of formulae is less than 3.33.

Modified from Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal. Chem.*, **2016**, *88*(16), pp. 7946–7958.

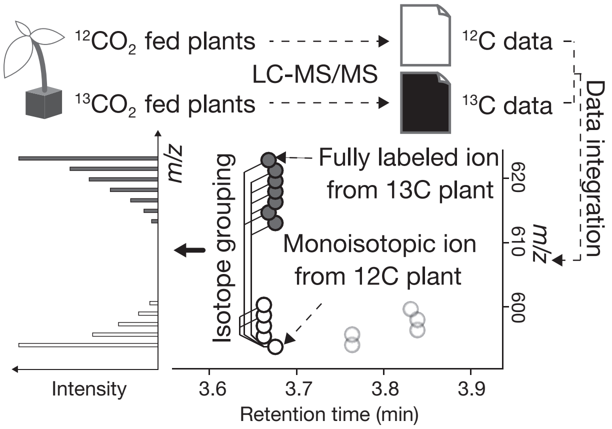

Fig. 4 The methodology to determine the carbon number of unknown molecules. The $^{12}CO_2$- and $^{13}CO_2$- fed plants were analyzed independently by LC-MS/MS. These data sets were integrated in MS-DIAL software where the ions from the same metabolite were grouped by considering the isotopic patterns. The *m/z* shift was used to determine the carbon number. Note that each spot represents an ion detected at a particular retention time and *m/z* value. Modified from A Cheminformatics Approach to Characterize Metabolomes in Stable Isotope-Labeled Organisms. *Nat. Methods*. **2019**. 10.1038/s41592-019-0358-2.

5 Introduction for Metabolite 'Class' and Structure Characterization

Once the molecular formula is reliably determined, the structure can be elucidated by various methodologies which include in silico fragmentation tools, molecular-spectrum networking, checking the characteristic ions, and metabolite classification. The first step is to retrieve candidate structures from metabolome structure databases, since most of unknown ions would be 'known' metabolites. It means that the metabolite has already been reported while the MS/MS spectrum has not been recorded. Indeed, in most of the cases, the correct structure for unknown spectra can be inferred from structure databases. On the other hand, the metabolite 'classification' and/or structure characterization of real unknowns (unknown-unknowns) is executed when no candidate is available in structure databases. In this chapter, the practical workflow for structure elucidations of natural products is showcased.

Recently, there are many well-curated structure databases for various organisms: KNApSAcK (45,852),[28] UNPD (166,995),[29] PlantCyc (3817),[30] FooDB (21,943),[31] NANPDB (3882)[32] for searching reported natural products, STOFF-IDENT (10,231)[33] for searching water-relevant compounds, LipidMAPS (34,466)[34] for searching lipid structures, HMDB[35] for searching human metabolomes, and ChEBI (53,746)[36] for searching various small molecules, where parentheses represent the count of structure records. In total, the current MS-FINDER program (version 3.12)[16] contains 321,616 metabolome structures, where the unique structure was counted by the first layer of InChIKey. Moreover, MINE (643,307)[37] in silico metabolite database which was generated by in silico enzymatic reactions of known metabolites is also encompassed. Given that MS-FINDER contains 61,048 reported molecular formulae, if this information is converted to structure information, on average 5.27 structures can be retrieved from a single formula from the structure database: the median, min, and max being 1, 1, and 1199, respectively. Therefore, the structure elucidation programs need to rank the candidate structures.

6 How to Rank Candidate Structures?

So far, several structure elucidation tools have been reported to rank the candidate structures. The tools utilize various methodologies which employ empirical rules, including the consideration of bond dissociation energies, machine learning, studying the relationship between mass spectra and structure fingerprints, and theoretical rules to interpret the mass fragmentations for the correct structures. The important thing for ranking structures is to evaluate the "validity" between the correct structure and the experimental MS/MS spectrum. Here, the MS/MS spectrum of feruloyl putrescine, also known as subaphylline, is shown in Fig. 5. In fact, a total of 14 candidate structures were retrieved from 321,616 structures for the formula $C_{14}H_{20}N_2O_3$, and one of them is phenylalanylvaline, which is a well-known dipeptide of proteinogenic amino acids. However, the MS/MS spectrum can be explained only by the correct structure, i.e., feruloyl putrescine. The product ion with *m/z* 177.054 is interpreted as C–N cleavage of feruloyl putrescine structure, and the fragment ion is observed without any hydrogen rearrangement, which can be described as $[M'-H]^+$, where M′ stands for the neutralized form of fragment ion. Namely, the positive charge is stable in the form of $R-C^+(=O)$ according to the valence and even electron rules. On the other hand, the product ion with *m/z* 177.054 can be generated from phenylalanylvaline by two sequential cleavages of C–C bond with a larger mass error (*m/z* 177.102). However, the cleavage of C–C bond with no adjacent hetero atom requires high dissociation energy, and it has rarely been observed in low energy collision-induced dissociation (low energy CID), which is used for conventional LC-MS/MS instruments. Moreover, the sequential C–C cleavages often provide one dehydrogenation in positive ion mode, where the fragment ion can be described as $[M'-3H]^+$.

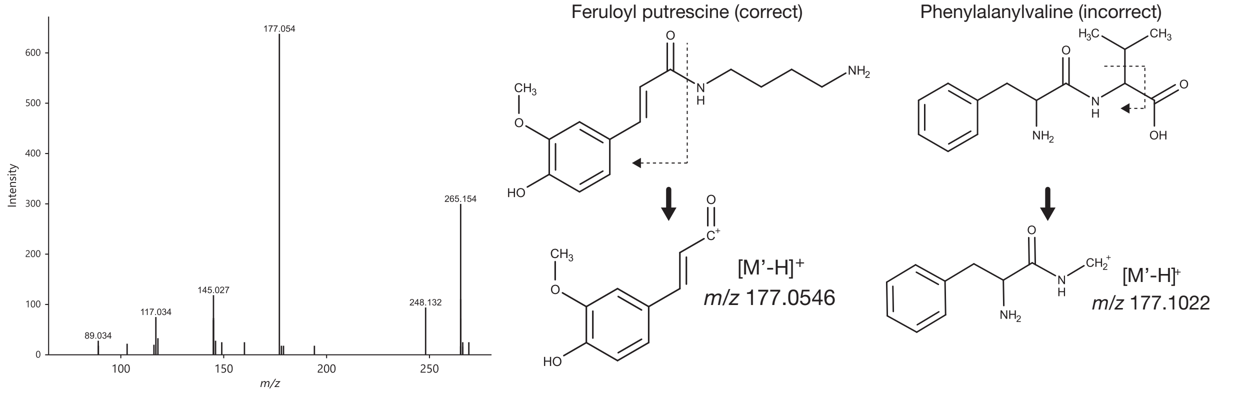

Fig. 5 The principle to rank the candidate structures. The MS/MS spectrum of feruloyl putrescine is shown. The base peak of *m/z* 177.054 can be resolved by hydrogen rearrangement (HR) rules when the correct structure, for example, feruloyl putrescine is considered. On the other hand, it cannot be resolved by HR rules when an incorrect structure, for example, phenylalanylvaline is considered. The mass of the predicted fragment ion (*m/z* 177.1022) does not match the mass of the experimental product ion (*m/z* 177.054 calculated with 272 ppm error), where the conventional accuracy of time-of-flight mass spectrometers is less than 20 ppm.

As expected, there is a possibility that a reasonable amount of fragment may be generated by an isomerization reaction in the collision cell of mass spectrometer, but the structural rearrangement is not considered in our combinatorial techniques.

In 2016, we examined the mass fragmentations of 5063 metabolites, and formulated "hydrogen rearrangement (HR) rules[9]" by refining and extending the even-electron rule to carbon, oxygen, nitrogen, phosphorus, and sulfur elements according to the statistics of assigned substructure ions. A total of nine rules have been formulated, and they enable us to interpret the mass fragmentation of metabolite in a theoretical and physicochemical manner. By using these rules, approximately 70% product ions from 5063 metabolites have been explained. For ranking structures, our hypothesis is that the number of fragment ions resolved by HR rules is larger for the correct structure than for incorrect ones. Therefore, the MS-FINDER program evaluates the relationship between a structure and a spectrum if the MS/MS spectrum can be explained reasonably using the HR rules-based mass fragmentations for the structure. In addition, several empirical rules such as mass errors, bond dissociation energies, and fragmentation linkage discrepancies are also incorporated. The above evaluations are scored for each of the candidate structures, and they are ranked accordingly. Nowadays, the accuracies for ranking structures as the top 1 candidate and any of the top 3 candidates are around 60% and 80%, respectively, when the spectral records of MassBank,[38] GNPS,[22] ReSpect,[39] and PlaSMA[16] are used for searching the available structures.

7 Checking the Unique Ions and Neutral Losses to Characterize Structure Moieties

The result of structure prediction accuracy indicated that its reliability is not ideal: 40% of suggested top candidate structures were wrong. Therefore, additional criteria are needed to elucidate molecular structures more reliably. One of the important criteria is to check the existence of unique product ions or neutral losses to characterize the molecular backbones and motifs. For instance, (1) the observation of *m/z* 78.9585 (PO_3^-) and *m/z* 96.9595 (HSO_4^-) values increases the confidence in phosphate- and sulfate-containing metabolites, respectively; (2) the observation of neutral losses of 146.0579 ($C_6H_{10}O_4$), 162.05282 ($C_6H_{10}O_5$), and 176.0321 ($C_6H_8O_6$) increases the confidence in metabolites that contain deoxyhexose (dHex), hexose (Hex), and uronic acid (HexA) moieties, respectively; (3) the observation of flavonol aglycone's ion e.g., *m/z* 285.0405 ($C_{15}H_9O_6^-$), *m/z* 301.0354 ($C_{15}H_9O_7^-$), and *m/z* 315.051 ($C_{16}H_{11}O_7^-$) is needed to characterize kaempferol, quercetin, and isorhamnetin, respectively. Practically, top 10 candidates from MS-FINDER are manually evaluated to check whether the unique ion, which can be monitored in MS from the suggested structure, exists in the experimental MS/MS spectrum.

Furthermore, the structure specificity in plants should be validated. For example, *m/z* values of aglycone for flavone and isoflavone are identical, and these isomers cannot be distinguished unless the unique ion is observed to classify their isomers. If the product ion of flavone/isoflavone aglycone is observed in *Oryza sativa*, the metabolite can be classified to 'flavone' since the biosynthetic gene for producing isoflavones has never been discovered in *O. sativa*. The characteristics including the major plant species and tissues, the adduct types detected, the unique product ions and neutral losses, and available publications on 64 metabolite classes are summarized in Table 5.

8 What If No Candidate Structure Exists in Databases?

As there are many unknown metabolites, especially in plants, developing the method for metabolite class prediction is an emerging need for structure elucidations in untargeted metabolomics. Recently, we developed an algorithm to recommend the metabolite

Table 5 Characteristic ions to define the metabolite classes.

			ESI(+)		*ESI(−)*	
Metabolite class	*Main plant*	*Adduct form*	*Product ion*	*Neutral loss*	*Product ion*	*Neutral loss*
Allicin derivatives	*Allium cepa*	[M+H]+	73.0106_C_3H_5S			
Anthraquinones	Ophiorrhiza pumila	[M-H]-			283.0612_$C_{16}H_{11}O_5$ (Anthraquinone + 1O + 1MeO + 1MeOH); 237.0557_$C_{15}H_9O_3$ (Anthraquinone + 1O + 1MeO + 1MeOH-1O-1MeO); 269.04555_$C_{15}H_9O_5$ (Anthraquinone + 2O + 1MeOH) 251.0350_$C_{15}H_7O_4$ (Anthraquinone + 2O + 1MeOH-H_2O); 253.0506_$C_{15}H_9O_4$ (Anthraquinone + 1O + 1MeOH)	H_2O, MeO, CO losses from aglycone; Major Hex losses
Benzoxazinoids	Zea may, Allium cepa, *Solanum tuberosum*	[M+H]+/[M-H]-	150.0550_$C_8H_8NO_2$ (ABOA); 164.0342_$C_8H_6NO_3$ (DIBOA); 198.0761_C_9H_{12}NO4 (HMBOA); 212.0553_$C_9H_{10}NO_5$ (DIMBOA); 226.0710_$C_{10}H_{12}$NO5 (HDMBOA)	H_2O, CO, CH_2O_2 losses from aglycone; Major Hex losses	164.0353_$C_8H_6NO_3$ (MBOA or HBOA); 180.0302_$C_8H_6NO_4$ (DIBOA); 192.0302_$C_9H_6NO_4$ (DIMBOA); 194.0459_$C_9H_8NO_4$ (HMBOA); 224.0564_$C_{10}H_{10}NO_5$ (HDMBOA)	H_2O, CO, CH_2O_2 losses from aglycone; Major Hex losses
Carbolines-camptothecin derivatives	Ophiorrhiza pumila	[M+H]+	349.1183_$C_{20}H_{17}N_2$O4 (aglycone); 303.1128_$C_{19}H_{15}N_2$O2 (aglycone-$C_3H_6O_4$); 287.1179_$C_{19}H_{15}N_2O$ (aglycone-CH_2O_2)	43.9893_CO_2; 46.0055_CH_2O2;		
Carbolines-cleaved in C ring	Ophiorrhiza pumila	[M+H]+	387.1551_$C_{20}H_{23}N_2O_6$ (aglycone)	93.0573_C_6H_7N		
Carbolines-cleaved in D ring	Ophiorrhiza pumila	[M+H]+	338.1387_$C_{20}H_{20}NO_4$ (aglycone-NH3); 352.1391_$C_{17}H_{22}$NO7 (aglycone-NH_3)	17.0260_NH_3		
Carbolines-dehydropyran in E ring	Ophiorrhiza pumila	[M+H]+/ [M-H]- & [M+HCOO]-	351.1339_$C_{20}H_{19}N_2O_4$ (aglycone); 337.1547_$C_{20}H_{21}N_2O_3$ (aglycone); 335.1390_$C_{20}H_{19}N_2O_3$ (aglycone); 299.1026_$C_{16}H_{15}N_2O_4$ (aglycone-C_4H_6O)	70.0413_C_4H_6O (by Retro Diels-Alder)		
Coumarin and derivatives	*Arabidopsis thaliana*, *Nicotiana tabacum*	[M+H]+/[M-H]-	207.0652_$C_{11}H_{11}O_4$ (Scoparone); 193.0495_$C_{10}H_9O_4$ (Scopoletin)	60.0206_$C_2H_4O_2$ & 27.9944_CO losses from Coumarin	191.0350_$C_{10}H_7O_4$ (Scopoletin); 177.0193_$C_9H_5O_4$ (Esculetin)	CH_3 & CO_2 losses from Coumarin; Major Hex losses
Flavone C,C,O-glycosides	*Glycyrrhiza glabra*, Glycyrrhiza uralensis	[M+H]+/[M-H]-			357.0980_$C_{19}H_{17}O_7$ (Apigenin + 2EtOH)	90.0322_$C_3H_6O_3$; 120.0428_$C_4H_8O_4$; Major Hex losses
Flavanone C,O-glycosides	Glycyrrhiza glabra, Glycyrrhiza uralensis	[M+H]+/[M-H]-	301.1071_$C_{17}H_{17}O_5$ (Naringenin + 2Me)	Many H_2O and CH_2O losses	355.0823_$C_{19}H_{15}O_7$ (Naringenin + 3O + 2EtOH)	90.0322_$C_3H_6O_3$; 120.0428_$C_4H_8O_4$

(*Continued*)

Table 5 (Continued)

			ESI(+)		*ESI(−)*	
Metabolite class	*Main plant*	*Adduct form*	*Product ion*	*Neutral loss*	*Product ion*	*Neutral loss*
Flavone C,C-glycosides	*Oryza sativa*, Zea may, Glycyrrhiza glabra, Glycyrrhiza uralensis	[M+H]+/[M-H]-	295.0601_$C_{17}H_{11}O_5$ (Apigenin+2Me); 311.0550_$C_{17}H_{11}O_6$ (Luteolin+2Me)	Many H_2O and CH_2O losses	353.0667_$C_{19}H_{13}O_7$; 355.0823_$C_{19}H_{15}O_7$ (Apigenin+2EtOH); 369.0616_$C_{19}H_{13}O_8$ (Luteolin+2EtOH)	
Flavone C,O-glycosides	Zea may	[M+H]+/[M-H]-	283.0601_$C_{16}H_{11}O_5$ (Apigenin+1Me)	Major Hex losses followed by H_2O and CH_2O losses	311.0561_$C_{17}H_{11}O_6$ (Apigenin+1EtOH)	90.0322_$C_3H_6O_3$; 120.0428_$C_4H_8O_4$; Major Hex losses
Flavone C-glycosides	Oryza sativa, Zea may, Glycyrrhiza glabra, Glycyrrhiza uralensis	[M+H]+/[M-H]-	283.0601_$C_{16}H_{11}O_5$ (Apigenin+1Me); 297.0757_$C_{17}H_{13}O_5$ (Acacetin+1Me); 299.0550_$C_{16}H_{11}O_6$ (Luteolin+1Me); 313.0707_$C_{17}H_{13}O_6$ (Diosmetin+1Me); 327.0863_$C_{18}H_{15}O_6$ (Dihydroxy-dimethoxyflavone+1Me)	Many H_2O and CH_2O losses	311.0561_$C_{17}H_{11}O_6$ (Apigenin+1EtOH); 325.0718_$C_{18}H_{13}O_6$ (Acacetin+1EtOH); 327.0510_$C_{17}H_{11}O_7$ (Luteolin+1EtOH); 341.0667_$C_{18}H_{13}O_7$ (Diosmetin+1EtOH)	90.0322_$C_3H_6O_3$; 120.0428_$C_4H_8O_4$; Major Hex losses
Anthocyanidin O-glycosides	a	[M]+	287.0550_$C_{15}H_{11}O_6$ (Cyanidin); 303.0499_$C_{15}H_{11}O_7$ (Delphinidin); 317.0656_$C_{16}H_{13}O_7$ (Petunidin); 331.0812_$C_{17}H_{15}O_7$ (Malvidin)	Major Hex losses[a]		
Biflavonoids-polycatechols	Glycyrrhiza glabra, Glycyrrhiza uralensis	[M+H]+/[M-H]-	289.0707_$C_{15}H_{13}O_6$ (monocatechol); 291.0863_$C_{15}H_{15}O_6$ (monocatechol); 579.1497_$C_{30}H_{27}O_{12}$ (dicatechol)		287.0561_$C_{15}H_{11}O_6$; 289.0718_$C_{15}H_{13}O_6$ (monocatechol); 577.1351_$C_{30}H_{25}O_{12}$ (dicatechol)	
Flavanol O-glycosides	Glycyrrhiza glabra, Glycyrrhiza uralensis	[M+H]+/[M-H]-	291.0863_$C_{15}H_{15}O_6$ (Catechin)	Major Hex losses	289.0718_$C_{15}H_{13}O_6$ (Catechin)	Major Hex losses
Flavanone O-glycosides	Glycyrrhiza glabra, Glycyrrhiza uralensis	[M+H]+/[M-H]-	273.0757_$C_{15}H_{13}O_5$ (Naringenin); 287.0914_$C_{16}H_{15}O_5$ (Sakuranetin); 289.0707_$C_{15}H_{13}O_6$ (Eriodictyol); 303.0863_$C_{16}H_{15}O_6$ (Hesperetin)	Major Hex losses	255.0663_$C_{15}H_{11}O_4$ (Liquiritigenin); 271.0612_$C_{15}H_{11}O_5$ (Naringenin); 285.0768_$C_{16}H_{13}O_5$ (Sakuranetin); 287.0561_$C_{15}H_{11}O_6$ (Eriodictyol); 301.0718_$C_{16}H_{13}O_6$ (Hesperetin); 319.0459_$C_{15}H_{11}O_8$ (Hexahydroxy flavanone);	Major Hex losses
Flavone O-glycosides	Oryza sativa, Zea may, Glycyrrhiza glabra, Glycyrrhiza uralensis, *Medicago truncatula*, *Glycine max*, Ophiorrhiza pumila	[M+H]+/[M-H]-	255.0652_$C_{15}H_{11}O_4$ (Dihydroxyflavone); 271.0601_$C_{15}H_{11}O_5$ (Apigenin); 287.0550_$C_{15}H_{11}O_6$ (Luteolin); 301.0707_$C_{16}H_{13}O_6$ (Diosmetin); 331.0812_$C_{17}H_{15}O_7$ (Tricin)	Major Hex losses followed by H_2O and CH_2O losses	269.0455_$C_{15}H_9O_5$ (Apigenin); 284.0326_$C_{15}H_8O_6$ (radical); 285.0405_$C_{15}H_9O_6$ (Luteolin); 299.0561_$C_{16}H_{11}O_6$ (Diosmetin); 329.0667_$C_{17}H_{13}O_7$ (Tricin)	Major Hex losses
Flavonol O-glycosides	Except for OP	[M+H]+/[M-H]-	287.0550_$C_{15}H_{11}O_6$ (Kaempferol); 299.0914_$C_{17}H_{15}O_5$ (Dimethoxyflavonol); 301.0707_$C_{16}H_{13}O_6$ (Kaempferide); 303.0499_$C_{15}H_{11}O_7$ (Quercetin);	Major Hex losses followed by H_2O and CH_2O losses	253.0506_$C_{15}H_9O_4$ (Monohydroxyflavonol); 284.0326_$C_{15}H_8O_6$ (radical) or 285.0405_$C_{15}H_9O_6$ (Kaempferol); 297.0768_$C_{17}H_{13}O_5$ (Dimethoxyflavonol); 299.0561_$C_{16}H_{11}O_6$ (Kaempferide);	Major Hex losses

			319.0448_$C_{15}H_{11}O_8$ (Quercetagetin);
			333.0605_$C_{16}H_{13}O_8$ (Patuletin);
			347.0761_$C_{17}H_{15}O_8$ (Syringetin)
					300.0276_$C_{15}H_8O_7$ (radical) or
					301.0354_$C_{15}H_9O_7$ (Quercetin);
					315.0510_$C_{16}H_{11}O_7$ (Methylquercetin);
					316.0225_$C_{15}H_8O_8$ (radical) or
					317.0303_$C_{15}H_9O_8$ (Quercetagetin);
					331.0459_$C_{16}H_{11}O_8$ (Patuletin)	

Isoflavone O-glycosides	Glycyrrhiza glabra, Glycyrrhiza uralensis, Medicago truncatula, Glycine max, Ophiorrhiza pumila	[M+H]+/ [M-H]- & [M+HCOO]-	255.0652_$C_{15}H_{11}O_4$ (Daidzein);
			269.0808_$C_{16}H_{13}O_4$ (Formononetin);
			271.0601_$C_{15}H_{11}O_5$ (Genistein);
			285.0757_$C_{16}H_{13}O_5$ (Glycitein);
			299.0914_$C_{17}H_{15}O_5$ (Afrormosin)
					253.0506_$C_{15}H_9O_4$ (Daidzein);
					267.0663_$C_{16}H_{11}O_4$ (Formononetin);
					271.0612_$C_{15}H_{11}O_5$ (Trihydroxyisoflavone)
						Major Hex losses

Fructosamine peptides	Allium cepa	[M+H]+/[M-H]-	Amino acid fragments: 144.0240_$C_6H_8O_2S$; 130.0499_$C_5H_8NO_3$; 166.0863_$C_9H_{12}NO_2$; 230.0482_$C_9H_{12}NO_4S$; 230.1387_$C_{11}H_{20}NO_4$			90.0322_$C_3H_6O_3$, 162.0534_$C_6H_{10}O_5$
Fructosamine amino acids	Allium cepa, Oryza sativa, Nicotiana tabacum, Medicago truncatula, *Solanum lycopersicum*	[M+H]+/[M-H]-	Amino acid fragments: 144.0240_$C_6H_8O_2S$; 130.0499_$C_5H_8NO_3$; 166.0863_$C_9H_{12}NO_2$; 230.0482_$C_9H_{12}NO_4S$; 230.1387_$C_{11}H_{20}NO_4$)			90.0322_$C_3H_6O_3$, 162.0534_$C_6H_{10}O_5$
Glucosinolate breakdown metabolites	Arabidopsis thaliana	[M+H]+		63.9977_CH_4OS		63.9988_CH_4OS
Glucosinolates	Arabidopsis thaliana	[M-H]-			96.9601_HSO_4 (sulfate), 195.0333_$C_6H_{11}O_5S$ (SHex)	

Iridoid glycosides	Ophiorrhiza pumila	[M-H]- & [M+HCOO]-			227.0925_$C_{11}H_{15}O_5$ (Loganin);
					225.0768_$C_{11}H_{13}O_5$ (Geniposide);
					239.0561_$C_{11}H_{11}O_6$ (Apodanthoside);
					213.0768_$C_{10}H_{13}O_5$ (Loganic acid);
					197.0819_$C_{10}H_{13}O_4$ (Deoxyloganic acid);
						CO_2 or OH loss from aglycone; Hex loss

Lignols	Arabidopsis thaliana, Oryza sativa	[M+H]+ (hard)/ [M-H]-			195.0663_$C_{10}H_{11}O_4$ (Guaiacyl); 193.0506_$C_{10}H_9O_4$ (Feruloyl)	
DGMG	a	[M-H]- or [M+HCOO]-			Fatty acid fragment ions; 397.1351_$C_{15}H_{25}O_{12}$ (DGDG-head+Glycerol-H2O)	Fatty acids loss
LPC	a	[M+H]+/ [M+HCOO]-	184.0733_$C_5H_{15}NO_4P$		Fatty acid fragment ions	60.0217_$C_2H_4O_2$ (HCOO & Me loss)
LPE	a	[M+H]+/[M-H]-			Fatty acid fragment ions	197.0459_$C_5H_{12}NO_5P$ (PE-header)
MGMG	a	[M-H]- or [M+HCOO]-			Fatty acid fragment ions	236.0902_$C_9H_{16}O_7$ (Hex & Glycerol loss)

(*Continued*)

Table 5 (Continued)

			ESI(+)		*ESI(−)*	
Metabolite class	*Main plant*	*Adduct form*	*Product ion*	*Neutral loss*	*Product ion*	*Neutral loss*
Saccharolipids-acylsugar	Nicotiana tabacum	[M+NH4]+/ [M-H]- & [M+HCOO]-		197.0894_$C_6H_{15}NO_6$ (AcetylFructose +NH_4)		84.0581_C_5H_8O; 98.0737_$C_6H_{10}O$
Nicotianosides	Nicotiana tabacum	[M-H]-& [M+HCOO]-				CO_2, $C_2H_4O_2$, Hex, dHex, MalonylHex losses
Nicotine and derivatives	Nicotiana tabacum	[M+H]+	132.0808_$C_9H_{10}N$	31.0417_CH_5N		
Fatty acyl glycosides	Zea may, Glycine max, Medicago truncatula, Glycyrrhiza glabra, Glycyrrhiza uralensis, Solanum lycopersicum	[M-H]-				162.0534_$C_6H_{10}O_5$ (Hex), 180.0639_$C_6H_{12}O_6$ (Hex), or 224.0538_$C_7H_{12}O_8$ (CO_2 + $C_6H_{12}O_6$) losses
Phenolic glycosides	a	[M-H]- & [M+HCOO]-			167.0350_$C_8H_7O_4$; 153.0193_$C_7H_5O_4$; 179.0714_$C_{10}H_{11}O_3$; 177.0557_$C_{10}H_9O_3$; 139.0401_$C_7H_7O_3$; 167.0350_$C_8H_7O_4$	Hex losses (sometimes Pentose, and sometimes as radical)
Caffeic acid and derivatives	Glycyrrhiza glabra, Glycyrrhiza uralensis, Solanum lycopersicum, Solanum tuberosum, Ophiorrhiza pumila	[M+H]+/[M-H]-	163.0390_$C_9H_7O_3$ (Caffeoyl); 165.0546_$C_9H_9O_3$ (Dihydrocaffeoyl)	H_2O and CO losses from Caffeoyl	179.0350_$C_9H_7O_4$ (Caffeoyl); 175.0275_$C_9H_5NO_3$ (Caffeoyl amide)	CO_2 losses from Caffeoyl. When amine is conjugated, the behavior is changed.
Cinnamic acid and derivatives	Arabidopsis thaliana, Allium cepa, Nicotiana tabacum, Solanum lycopersicum, Solanum tuberosum	[M+H]+	131.0491_C_9H_7O (Cinnamoyl)	CO losses from Cinnamoyl		
Coumaric acid and derivatives	a	[M+H]+/[M-H]-	147.0441_$C_9H_7O_2$ (Coumaroyl)	H_2O and CO losses from Coumaroyl	163.0401_$C_9H_7O_3$ (Coumaroyl); 145.0295_$C_9H_5O_2$ (Coumaroyl-H_2O when amine conjugated)	CO_2 losses from Coumaroyl
Ferulic acid and derivatives	Medicago truncatula, Oryza sativa, Zea may, Solanum lycopersicum, Solanum tuberosum, Allium cepa, Glycyrrhiza glabra, Glycyrrhiza uralensis	[M + H]+/[M-H]-	177.0546_$C_{10}H_9O_3$ (Feruloyl)	H_2O and CO losses from Feruloyl	193.0506_$C_{10}H_9O_4$ (Feruloyl); 175.0401_$C_{10}H_7O_3$ (Feruloyl-H_2O)	CH_3 & CO_2 losses from Feruloyl
Hydroxyferulic acid and derivatives	Glycyrrhiza glabra, Glycyrrhiza uralensis	[M+H]+/[M-H]-			209.0455_$C_{10}H_9O_5$ (Hydroxyferuloyl)	CH_3 & CO_2 losses from Hydroxyferuloyl
Sinapinic acid and derivatives	Arabidopsis thaliana, Oryza sativa, Zea may, Solanum lycopersicum, Solanum tuberosum	[M+H]+/[M-H]-	207.0652_$C_{11}H_{11}O_4$ (Sinapoyl)		223.0612_$C_{11}H_{11}O_5$ (Sinapoyl); 205.0506_$C_{11}H_9O_4$ (Sinapoyl-H_2O)	CH_3 & CO_2 losses from Feruloyl

Prenylated flavanones	Glycyrrhiza glabra, Glycyrrhiza uralensis	[M+H]+/[M-H]-	285.0757_$C_{16}H_{13}O_5$ (Naringenin+1Me-2H); 301.0707_$C_{16}H_{13}O_6$ (Eriodictyol+1Me-2H)	56.0621_C_4H_8 (Prenyl)	Fragment ions from retro diels-alder reaction in C ring	
Prenylated isoflavanones	Glycyrrhiza glabra, Glycyrrhiza uralensis	[M+H]+/[M-H]-	285.0757_$C_{16}H_{13}O_5$ (Trihydroxyisoflavanone+1Me-2H)	56.0621_C_4H_8 (Prenyl)		
Acyl carnitines	[a]	[M+H]+	85.0284_$C_4H_5O_2$	59.0730_C_3H_9N		
Indole derivatives	[a]	[M+H]+	144.0808_$C_{10}H_{10}N$			

Purine & pyrimidine nucleosides	[a]	[M+H]+/[M-H]-	113.0346_$C_4H_5N_2O_2$ (Uracil);
			153.0407_$C_5H_5N_4O_2$ (Xanthine);
			152.0567_$C_5H_6N_5O$ (Guanine);
			136.0618_$C_5H_6N_5$ (Adenine)
				132.0417_$C_5H_8O_4$	150.0421_$C_5H_4N_5O$ (Guranine);
					151.0261_$C_5H_3N_4O_2$ (Xanthine);
					110.0122_$C_4H_2N_2O_2$ (Uracil as radical)
						132.0428_$C_5H_8O_4$

Quinic acid and derivatives	Arabidopsis thaliana, Nicotiana tabacum, Zea may, Solanum lycopersicum, Solanum tuberosum, Ophiorrhiza pumila	[M+H]+/[M-H]-			191.0561_$C_7H_{11}O_6$ (Quinic acid); 353.0878_$C_{16}H_{17}O_9$ (Chlorogenic acid)	Major Hex losses
Steroidal saponins-dehydrofurostane	Allium cepa	[M+H]+/ [M-H]- & [M+HCOO]-	431.3156_$C_{27}H_{43}O_4$ (aglycone)	Major Hex losses		
Steroidal saponins-furostane	Allium cepa, Solanum lycopersicum, Solanum tuberosum	[M-H2O+H]+/ [M-H]- & [M+HCOO]-	415.3207_$C_{27}H_{43}O_3$ (aglycone-H_2O), 417.3363_$C_{27}H_{45}O_3$ (aglycone-H_2O), 431.3156_$C_{27}H_{43}O_4$ (aglycone-H_2O)	Major Hex losses		
Steroidal saponins-solanidine	Solanum tuberosum, Solanum lycopersicum	[M+H]+/ [M-H]- & [M+HCOO]-	(MS1) 398.3417_$C_{27}H_{44}NO$ (Solanidine), (MS1) 412.3210_$C_{27}H_{42}NO_2$ (Hydroxydehydrosolanidine)	(MS1) Major Hex losses		
Steroidal saponins-solasodine	Solanum lycopersicum, Solanum tuberosum	[M+H]+/ [M-H]- & [M+HCOO]-	(MS1) 414.3367_$C_{27}H_{44}NO_2$ (Solasodine), (MS1) 416.3523_$C_{27}H_{46}NO_2$ (Dihydrosolasodine), (MS1) 432.3472_$C_{27}H_{46}NO_3$ (Hydroxydihydrosolasodine)	(MS1) Major Hex losses		
Steroidal saponins-solasodine acetoxy	Solanum lycopersicum	[M+H]+/ [M-H]- & [M+HCOO]-	(MS1) 432.3472_$C_{27}H_{46}NO_3$ (Hydroxydihydrosolasodine-$C_2H_2O_2$), (MS1) 416.3523_$C_{27}H_{46}NO_2$ (Dihydrosolasodine-$C_2H_2O_2$)	(MS1) Major Hex losses		
Steroidal saponins-spirostane	Allium cepa, Solanum tuberosum	[M+H]+ or [M-H2O+H]+/ [M-H]- & [M+HCOO]-	413.3050_$C_{27}H_{41}O_3$ (aglycone-H_2O), 429.2999_$C_{27}H_{41}O_4$ (aglycone or aglycone-H_2O)	Major Hex losses		
Sulfate containing metabolites	[a]	[M-H]-			96.9601_HSO_4 (sulfate)	

(*Continued*)

Table 5 (Continued)

			ESI(+)		*ESI(−)*	
Metabolite class	*Main plant*	*Adduct form*	*Product ion*	*Neutral loss*	*Product ion*	*Neutral loss*
Sulfonic acids	Allium cepa	[M-H]-			79.9574_SO_3; 80.9652_HSO_3	
Triterpene saponins-aglycone A	Medicago truncatula	[M + H]+ (hard)/[M-H]- & [M + HCOO]-		Major Hex losses	485.3272_$C_{30}H_{45}O_5$ (aglycone); 467.3167_$C_{30}H_{43}O_4$ (aglycone-H_2O)	
Triterpene saponins-bayogenin	Medicago truncatula	[M + H]+/[M-H]- & [M + HCOO]-	489.3575_$C_{30}H_{49}O_5$ (Bayogenin)	Major Hex losses	487.3429_$C_{30}H_{47}O_5$ (Bayogenin)	
Triterpene saponins-hederagenin or hydroxy oleanolic acid	Medicago truncatula	[M + H]+/ [M-H]- & [M + HCOO]-	(MS1 or MS/MS) 455.3520_$C_{30}H_{47}O_3$ (Hederagenin-H_2O)	Major Hex losses		
Triterpene saponins-licoricesaponin G	Glycyrrhiza uralensis, Medicago truncatula	[M + H]+/ [M-H]- & [M + HCOO]-	(MS1 or MS/MS) 487.3418_$C_{30}H_{47}O_5$ (Hydroxyglycyrrhizin); 469.3312_$C_{30}H_{45}O_4$ (Hydroxyglycyrrhizin-H_2O); 451.3207_$C_{30}H_{43}O_3$ (Hydroxyglycyrrhizin-$2H_2O$)	Major Hex losses		
Triterpene saponins-medicagenic acid	Medicago truncatula	[M + H]+(hard)/ [M-H]- & [M + HCOO]-	(MS1 or MS/MS) 457.3312_$C_{29}H_{45}O_4$ (Medicagenic acid-CH_2O_2); 439.3207_$C_{29}H_{43}O_3$ (Medicagenic acid-CH_2O_2-H_2O)	Major Hex losses	439.3218_$C_{29}H_{43}O_3$ (Medicagenic acid-CO_2-H_2O)	
Triterpene saponins-oleanolic acid	Medicago truncatula	[M + H]+(hard)/ [M-H]- & [M + HCOO]-	(MS1 or MS/MS) 457.3676_$C_{30}H_{49}O_3$ (oleanolic acid); 439.3571_$C_{30}H_{47}O_2$ (oleanolic acid-H_2O)	Major Hex losses		
Triterpene saponins-soyasapogenol A or E	Glycine max, Medicago truncatula	[M + H]+/ [M-H]- & [M + HCOO]-	457.3676_$C_{30}H_{49}O_3$ (aglycone-H_2O); 439.3571_$C_{30}H_{47}O_2$ (aglycone-$2H_2O$)	Major Hex losses		
Triterpene saponins-soyasapogenol B	Glycine max, Medicago truncatula, Glycyrrhiza glabra, Glycyrrhiza uralensis	[M + H]+/ [M-H]- & [M + HCOO]-	441.3727_$C_{30}H_{49}O_2$ (aglycone-H_2O)	Major Hex losses		
Triterpene saponins-soyasapogenol DDMP	Glycine max, Medicago truncatula, Glycyrrhiza glabra	[M + H]+/ [M-H]- & [M + HCOO]-	441.3727_$C_{30}H_{49}O_2$ (aglycone-H_2O); 423.3621_$C_{30}H_{47}O$ (aglycone-$2H_2O$)	Major Hex losses		

(a) The product ions and neutral losses are shown for 64 metabolite classes used to characterize the plant-specific metabolites.

[a] Major Hex losses: 248.0532_$C_9H_{12}O_8$ (MalonylHex); 204.0634_$C_8H_{12}O_6$ (AcetylHex); 176.0321_$C_6H_8O_6$ (HexA); 162.0528_$C_6H_{10}O_5$ (Hex); 146.0579_$C_6H_{10}O_4$ (dHex); 132.0423_$C_5H_8O_4$ (Pen).

Modified from A Cheminformatics Approach to Characterize Metabolomes in Stable Isotope-Labeled Organisms. *Nat. Methods*, **2019.** 10.1038/s41592-019-0358-2.

classes for an unknown MS/MS spectrum, named as fragment set enrichment analysis (FSEA) (Fig. 6).[16] The FSEA concept is extrapolated from the gene set enrichment analysis (GSEA),[40] where the significant gene 'ontology' describing the function in a biological mechanism is suggested with the input of significantly increased/decreased genes set. On the other hand, in FSEA, the significant metabolite ontology (class) describing the structure backbone or motifs, which can be explained by mass spectrometers, is provided with the input of significantly detected fragment ontologies in the MS/MS spectrum. The first step in this method is to assign molecular formulae (subformulae) to each of the product ions and neutral losses in an MS/MS spectrum after the molecular formula for the precursor *m*/*z* is determined. Second, the assigned subformulae are converted to fragment ontologies, which describe substructure information, by using the dictionary of subformulae (key) and fragment ontologies (value). Finally, the set of assigned fragment ontologies is evaluated for each set of metabolite class and fragment ontologies. If the detected fragment ontologies are "enriched" in the fragment set for a metabolite class, the MS/MS spectrum is recommended as the metabolite class with the estimated *p*-value by overrepresentation analysis (ORA).[16,41]

Most important for knowledge-oriented metabolite annotation should be the creation of a reliable dictionary of metabolite ontology and fragment sets, and in fact, extensive curation has been performed to create gene sets in the course of GSEA research. The procedure for creating the fragment sets is briefly described below. First, publicly and commercially available mass spectral databases including MassBank,[38] GNPS,[22] ReSpect,[39] PlaSMA,[16] MetaboBASE,[42] Metlin,[43] and NIST were utilized. After the same metabolite records analyzed by several instruments and conditions were merged to one integrated record, MS-FINDER was executed to assign fragment substructures for all product ions and neutral losses. The assigned substructure was defined by the neutralized form with hydrogen rearrangement (HR) information, and the InChIKey (for hash-key) and SMILES code (for structure description) were also generated for the substructure. To evaluate the significance of fragment ions, the frequency of all assigned substructures was examined. For example, the product ions of PO_3^- (phosphate) and $C_7H_7^+$ (methylbenzene) were observed in the MS/MS spectra of 8.19% and 8.18% molecules, respectively, and the neutral losses of H_2O and CO_2 were observed in the MS/MS spectra of 24.7% and 17.5% molecules, respectively. From a total of 157,067 substructures assigned in the MS/MS spectral records, the product ion and neutral loss fragments observed in more than 0.2% molecules were registered as 'frequently observed fragments.' In addition, the ontology, called as 'fragment ontology,' for all significant substructures was automatically defined by ClassyFire[44] software, which provides structure ontologies from the input of structure information. By converting the substructure to fragment 'ontology,' hereafter, the ions from different aglycons such as kaempferol and quercetin can be managed as the same metabolite ontology as 'flavonol.' Finally, the parent metabolite ontology, i.e., the metabolite class, was determined by human curation that considered the resolution of mass spectrometry-derived information for structure characterization. For example, flavonol *O*-glycoside and flavanone *O*-glycoside can be distinguished by the information of MS/MS spectra while ClassyFire annotates both as flavonoid *O*-glycosides. The sets of metabolite class and fragment ontologies are termed as FSEA sets. The function is implemented in MS-FINDER currently containing 6274 FSEA sets which consist of 459 fragment ontologies in total.

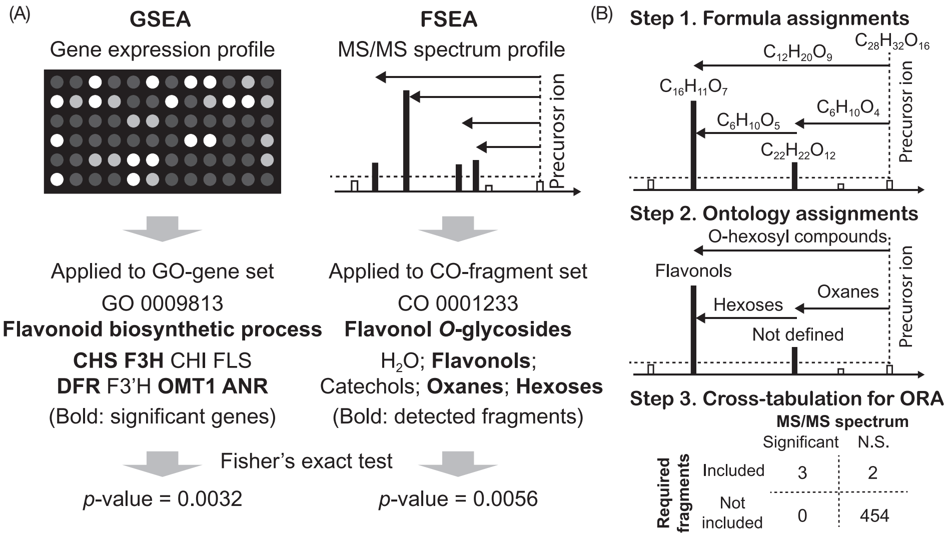

Fig. 6 The methodology for metabolite classification of unknown molecules by using fragment set enrichment analysis (FSEA). (A) The concept of FSEA was compared with the workflow of gene set enrichment analysis (GSEA). The genes showing significant change in expression in a biological study are evaluated for whether the significant genes are enriched in the set of a functional ontology, for example, flavonoid biosynthetic process. On the other hand, the significant peaks which are greater than a threshold are evaluated for whether the peaks annotated by the fragment ontology term are enriched in the set of a metabolite class, for example, flavonol *O*-glycosides. (B) The procedure for FSEA classification is described. First, the formula assignment was performed for each product ion and neutral loss. Second, the subformula information was applied to the fragment ontology candidates by using the fragment database. Finally, cross-tabulation table was prepared to evaluate the reliability of the estimated *p*-value by overrepresentation analysis (ORA). Modified from A Cheminformatics Approach to Characterize Metabolomes in Stable Isotope-Labeled Organisms, *Nat. Methods*. **2019**. 10.1038/s41592-019-0358-2.

Finally, the methodology to calculate the *p*-value in metabolite class recommendations is described. The important thing to utilize ORA is to define the 2 × 2 cross-tabulation table (Fig. 6). First, fragment ions with more than 5% relative abundance are defined as significant peaks; the abundance for neutral loss is defined by the intensity of the lower *m/z* value. Second, fragment ontology assignment to the experimental MS/MS spectrum is performed. Third, the left-top panel of the cross-tabulation table is incremented when the fragment ontology for an FSEA set is monitored in significant peaks; the remaining ontology candidates assigned to the matched peak are excluded. The left-bottom panel is incremented when the assigned ontology is not registered in the FSEA set. As a default, the right-top panel is automatically calculated by undetected ontologies in the FSEA set. The right-bottom panel is counted by the current knowledge-space of fragment ontologies ($n = 459$). There are two other options for determining the right panels: (1) nonsignificant peaks defined by less than 5% relative abundance ions can be utilized and (2) the right panel is counted by the significant peaks, i.e., those with more than 5% relative abundance ions in the reversed spectrum. The reversed spectrum is created by mirroring the *m/z* nominal places with a 5-mass shift to change the value of minimum and maximum *m/z* ions as well. For example, the *m/z* sequence [101.023, 305.035, 421.098, 634.201, 754.235] is changed to [106.023, 227.035, 440.098, 555.201, 759.235]. Here, 305.035 was converted to 555.201 as $428 + (428–305) + 5 + 0.201$ in which 428 was the rounded value of the mean (427.629) of 101.023 and 754.235. In fact, the definition of nonsignificant peaks is an important part for the Fisher-exact test, however, there is no consensus regarding the value of this methodology. Currently, false-positive and false-negative rates are different among three methods, and the investigation is ongoing. Finally, the Fisher-exact test is performed for *p*-value calculations using the 2 × 2 cross-tabulation table.

The FSEA algorithm provides information on metabolite class candidates without the need of any structure databases, i.e., a complete de novo recommendation. On the other hand, an MS/MS spectrum can belong to several metabolite classes. All of the recommended classes can be utilized totally with the characteristic ions, and therefore, the top three to five candidates would partially become the correct classes in the current FSEA program. Although the FSEA approach was not implemented for a valid metabolite class prediction, it yielded accuracies of 60.2% and 75.3% in positive and negative ion mode, respectively, for assigned candidates (top three predictions) without requiring any structure search.

9 Network Analysis Using Metabolite Ion Features to Propagate Known Mass Spectra to Unknowns

The 'networking' of metabolite ion features enables us to obtain intuitive insights into unknown metabolites. For example, MetaMapR[45] connects most of EI-MS features detected in GC-MS based metabolomics by using the mass spectral similarity between metabolite ions and using the ion abundance correlation among biological samples. For example, since the EI-MS patterns in trimethylsilyl (TMS)-derivative hexoses are highly conserved among the stereoisomers, an unknown metabolite linked to glucose, for example, can be annotated as 'hexose.' In addition, the global natural product social molecular networking (GNPS) provides the networking tool,[22] called as molecular networking, where the metabolite ions having similar tandem (MS/MS) mass spectra are linked. Since the similar structures often share the same product ions and neutral losses, an unknown metabolite can be elucidated if the MS/MS spectrum is linked to those of known metabolites. The current MS-DIAL and MS-FINDER software programs[16] provide the functions to perform the molecular-spectrum networking for GC-MS and LC-MS/MS data sets.

In contrast to a conventional mass spectral matching algorithm, such as dot product, the MS/MS similarity between precursor ions with different *m/z* values is calculated with the consideration of the precursor mass shift. First, the spectrum of larger precursor ion (spectrum A) is used as a template and spectrum B is fitted to spectrum A. The peak is defined as a "match" when the product ions or neutral losses of spectrum A and B are within a mass tolerance. The ion abundances of matched peaks are also considered for spectral similarity calculation.

The linkage of metabolite ions can be created with different ideas. In our previous study, the linkage was created by the term "similarity of metabolite ontologies," since our platform provides the metabolite ontology as well as metabolite structure information. Similarity was calculated by the Levenshtein distance, which was standardized from 0 to 100: the term similarity between 'flavonol *O*-glycosides' and 'flavone *O*-glycosides' is 91%. Another linkage was created using the dictionary of biotransformation (Table 6) with retention time restrictions[46]; 0.5 min is defined as a "near" elution profile. For example, two metabolite ions determined as $C_{12}H_{23}NO_7$ and $C_6H_{13}NO_2$ were correlated with the relationship of substrate and product in hexosylation, since the subtraction of these formulae yields $C_6H_{10}O_5$. When using a reverse phase column, the empirical retention time rule is also applied to decrease false positive linkages. For example, the hexosylated metabolite is eluted earlier. In the previous study, the integrated networking method which involved mass spectral similarity, ontology term similarity, and bioreaction linkage was used to map all metabolite ion features detected in 12 plant samples.

9.1 *Showcase 1*: PlaSMA ID-3265 (Positive) [M + H]+, PlaSMA ID-1978 (Negative) [M-H]$^-$, $C_{39}H_{58}O_{14}$, medicagenic Acid *O*-Malonyl Hexose

In the root extract of *Medicago truncatula*, we found an unknown metabolite which was detected with *m/z* 751.3985 and *m/z* 749.3729 in positive and negative ion modes, respectively (Fig. 7). According to the mass difference, the adduct type could be determined as $[M + H]^+$ and $[M\text{-}H]^-$, respectively. Moreover, MS-FINDER provided $C_{39}H_{58}O_{14}$ as the top candidate of the molecular formula, and we also confirmed the 39-mass shift in the set of the ^{12}C- and ^{13}C-labeled plant data. As triterpene saponins consisting of $C_xH_yO_z$ are one of the main natural products of *M. truncatula*, the unknown molecule can be considered as

Table 6 The major bioreaction found in biosynthetic pathway in plant-specific metabolites.

ID	*Name*	*Formula difference*	*Mass difference*	*Elution behavior*
1	Quinate	$+C_7H_{10}O_5$	174.0528234	−
2	Shikimate	$+C_7H_8O_4$	156.0422587	−
3	Tartarate	$+C_4H_4O_5$	132.0058732	−
4	Coumaryl alcohol	$+C_9H_8O$	132.0575149	+
5	Malate	$+C_4H_4O_4$	116.0109586	−
6	Deoxyhexose	$+C_6H_{10}O_4$	146.0579088	−
7	Coniferyl alcohol	$+C_{10}H_{10}O_2$	162.0680796	+
8	Catechol	$+C_6H_4O$	92.02621475	+
9	Vanillate	$+C_8H_6O_3$	150.0316941	+
10	Syringate	$+C_9H_8O_4$	180.0422587	+
11	Hydroxybenzoate	$+C_7H_4O_2$	120.0211294	+
12	Caffeate	$+C_9H_6O_3$	162.0316941	+
13	Dimethoxyquinol	$+C_8H_8O_3$	152.0473441	+
14	Hydroxyquinol	$+C_6H_4O_2$	108.0211294	+
15	Coumarate	$+C_9H_6O_2$	146.0367794	+
16	Sinapyl alcohol	$+C_{11}H_{12}O_3$	192.0786442	+
17	Isoprenylation	$+C_5H_8$	68.06260026	+
18	Vanillyl alcohol	$+C_8H_8O_2$	136.0524295	+
19	Ferulate	$+C_{10}H_8O_3$	176.0473441	+
20	Pentose	$+C_5H_8O_4$	132.0422587	−
21	Protocatechus alcohol	$+C_7H_4O_3$	136.016044	+
22	Hydroxybenzyl alcohol	$+C_7H_6O$	106.0418648	+
23	Caffeyl alcohol	$+C_9H_8O_2$	148.0524295	+
24	Syringyl alcohol	$+C_9H_{10}O_3$	166.0629942	+
25	Sinapate	$+C_{11}H_{10}O_4$	206.0579088	+
26	Quinol	$+C_6H_4O$	92.02621475	+
27	Syringyl	$+C_{11}H_{12}O_3$	192.0786442	+
28	Guaiacyl	$+C_{10}H_{10}O_2$	162.0680796	+
29	Glycerol	$+C_3H_6O_2$	74.03677943	−
30	Hexose	$+C_6H_{10}O_5$	162.0528234	−
31	Uronate	$+C_6H_8O_6$	176.032088	−

In the elution-behavior column, +/− characters indicate that the product metabolite was eluted later and earlier than the time of elution of the precursor form.
Modified from A Cheminformatics Approach to Characterize Metabolomes in Stable Isotope-Labeled Organisms. *Nat. Methods.* **2019**. 10.1038/s41592-019-0358-2.

belonging to the class of triterpene saponins. In fact, the fragment set enrichment analysis (FSEA) recommended triterpene saponins with *p*-values of 1.98×10^{-10} and 2.49×10^{-6} from the MS/MS spectra of positive and negative ion mode, respectively. A total of 14 aglycons for triterpene saponins have been proposed in *M. truncatula*, the unique product ion with *m/z* 439.3217 in negative ion mode has been recognized as the aglycone of medicagenic acid detected as $[\text{aglycone-}CO_2\text{-}H_2O\text{-}H]^-$.[47] Moreover, the product ions with *m/z* 457.3312 and 439.4571 in positive ion mode are also recognized as the characteristic ions, $[\text{aglycone-}CH_2O_2+H]^+$ and $[\text{aglycone-}CH_2O_2\text{-}H_2O+H]^+$ for medicagenic acid. In addition, when considering the neutral loss between the precursor ion and aglycone fragments, the glycoside can be annotated as malonyl hexose. These accumulated evidences enabled us to annotate the metabolite as medicagenic acid *O*-malonyl hexose.

9.2 *Showcase 2*: PlaSMA ID-3312 (Positive) [M+H]+, PlaSMA ID-1956 (Negative) $[M\text{-}CO_2\text{-}H]^-$, $C_{36}H_{38}O_{19}$ tricine *O*-guaiacylglyceyl *O*-Malonyl Hexose

In the leaf and stem of *Oryza sativa*, an unknown metabolite was found at *m/z* 775.2085 and 729.2036 in positive and negative ion mode, respectively (Fig. 8). According to the mass difference of 46.00, the adduct types were determined as $[M + H]^+$ and $[M\text{-}CO_2\text{-}H]^-$, respectively, with careful considerations. Since the mass shifts in the sets of ^{12}C and ^{13}C labeling data were 36 and 35 (owing to CO_2 loss) in positive and negative ion modes, respectively, MS-FINDER provided $C_{36}H_{38}O_{19}$ as the top candidate for molecular formula. FSEA suggested the metabolite belonging to flavonoid classes. We concluded that the unknown contains the tricine aglycone, a major flavone type found in *O. sativa,* since the product ions with *m/z* 331.081 and 329.667, which are the characteristic fragment ions for the tricine moiety were observed in positive and negative ion mode, respectively. When the original *m/z* value of $[M\text{-}H]^-$ (773.19) is considered in negative ion spectrum, the product ions with *m/z* 729.2032 and 525.1422 are considered as the neutral losses of CO_2 (43.99) and malonyl hexose (248.05). In addition, the product ion with m/z 195.0651 in negative ion mode and the neutral loss of 198.094 in positive ion mode are known as the characteristic features of *O*-guaiacylglycerol. These evidences enabled us to annotate the metabolite as tricine *O*-guaiacylglyceyl *O*-malonyl hexose.

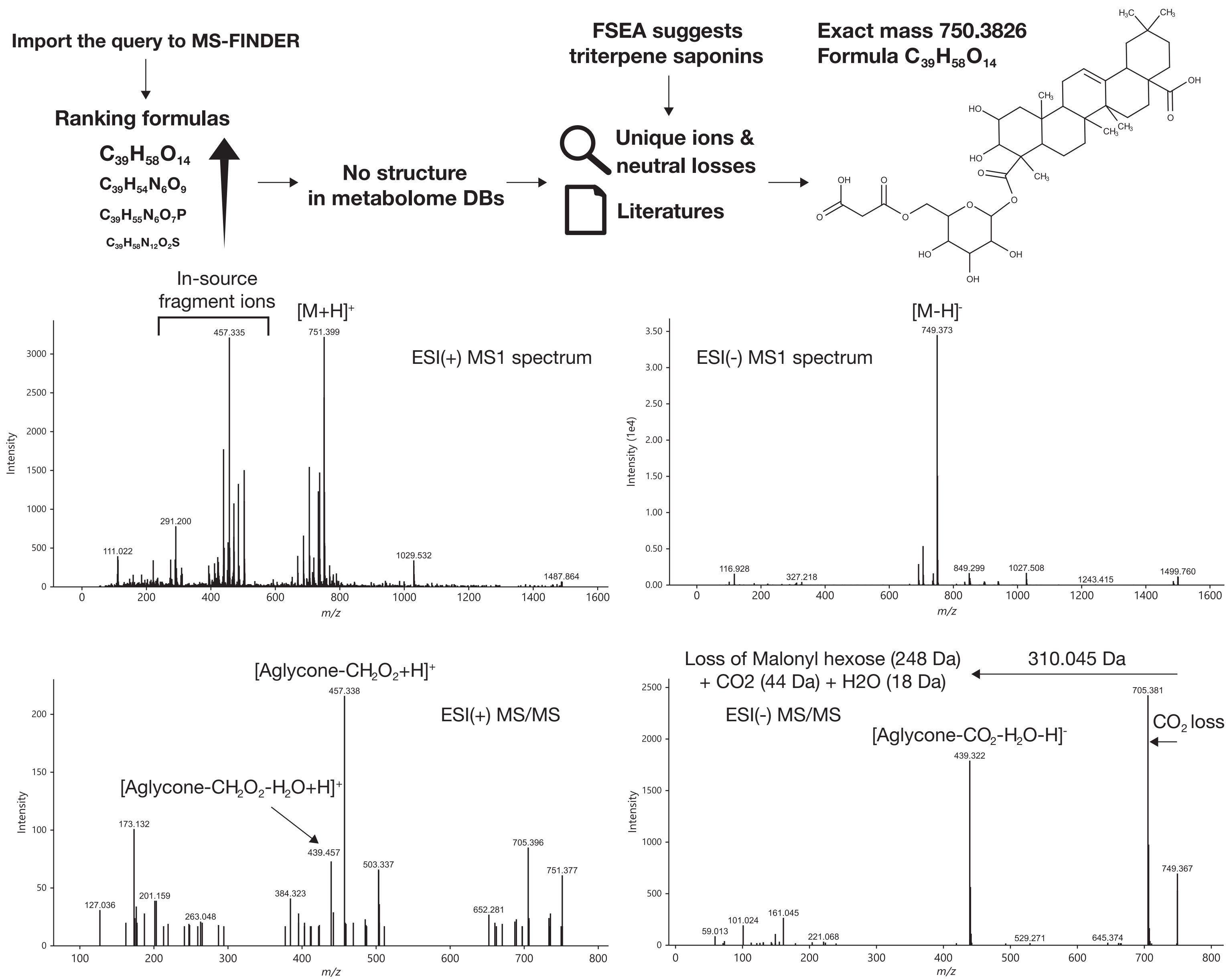

Fig. 7 Characterization of medicagenic acid *O*-malonyl hexose by computational mass spectrometry. The precursor ion and the adduct ion forms were determined as $[M+H]^+$ with *m/z* 751.399 and $[M\text{-}H]^-$ with *m/z* 749.373 in positive and negative ion mode, respectively. In addition, the carbon number for this unknown molecule was determined as 39 from the set of ^{12}C and ^{13}C plant data in MS-DIAL software. After the query of the MS^1 and MS/MS spectra was imported into MS-FINDER software, the molecular formula $C_{39}H_{58}O_{14}$ was determined as the top candidate in both positive and negative ion mode. Unfortunately, the suitable candidate structure matching the experimental MS/MS spectrum was not recorded in the metabolome structure databases. By integrating indirect evidences which contain the FSEA recommendation proposing triterpene saponins, the observation of characteristic ions (*m/z* 439 in negative ion mode) and the biosynthetic pathway in *Medicago truncatula*, the MS/MS spectrum was decoded as medicagenic acid *O*-malonyl hexose, where more accurate definition for structures can be achieved by synthetic chemistry and nuclear magnetic resonance (NMR).

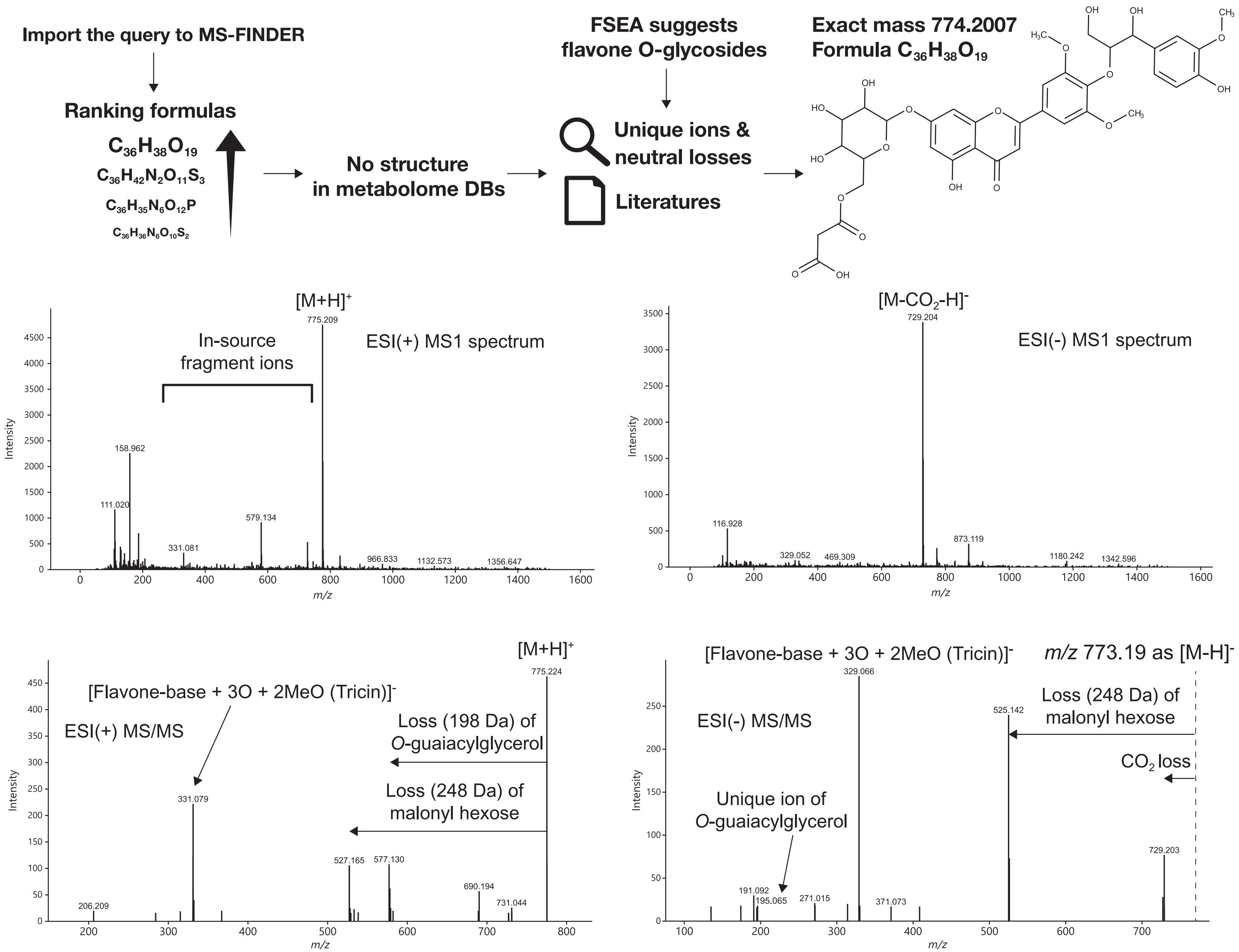

Fig. 8 Characterization of tricine *O*-guaiacylglyceyl *O*-malonyl hexose by computational mass spectrometry. The precursor ion and the adduct ion forms were determined as $[M+H]^+$ for *m/z* 775.209 and $[M\text{-}CO_2\text{-}H]^-$ for *m/z* 729.204 in positive and negative ion mode, respectively. In addition, the carbon number for this unknown molecule was determined as 36 from the set of ^{12}C and ^{13}C plant positive ion mode data in MS-DIAL software. After the query of the MS^1 and MS/MS spectra was imported into MS-FINDER software, the molecular formula $C_{36}H_{38}O_{19}$ was determined as the top candidate in both positive and negative ion mode. Unfortunately, the suitable candidate structure matching the experimental MS/MS spectrum was not recorded in the metabolome structure databases. By integrating indirect evidences which contain the FSEA recommendation proposing flavone *O*-glycosides, the observation of characteristic ions (*m/z* 331 and *m/z* 329 in positive and negative ion modes) and the biosynthetic pathway in *Oryza sativa*, the MS/MS spectrum was decoded as tricine *O*-guaiacylglyceyl *O*-malonyl hexose, where more accurate definition for structures can be achieved by synthetic chemistry and nuclear magnetic resonance (NMR).

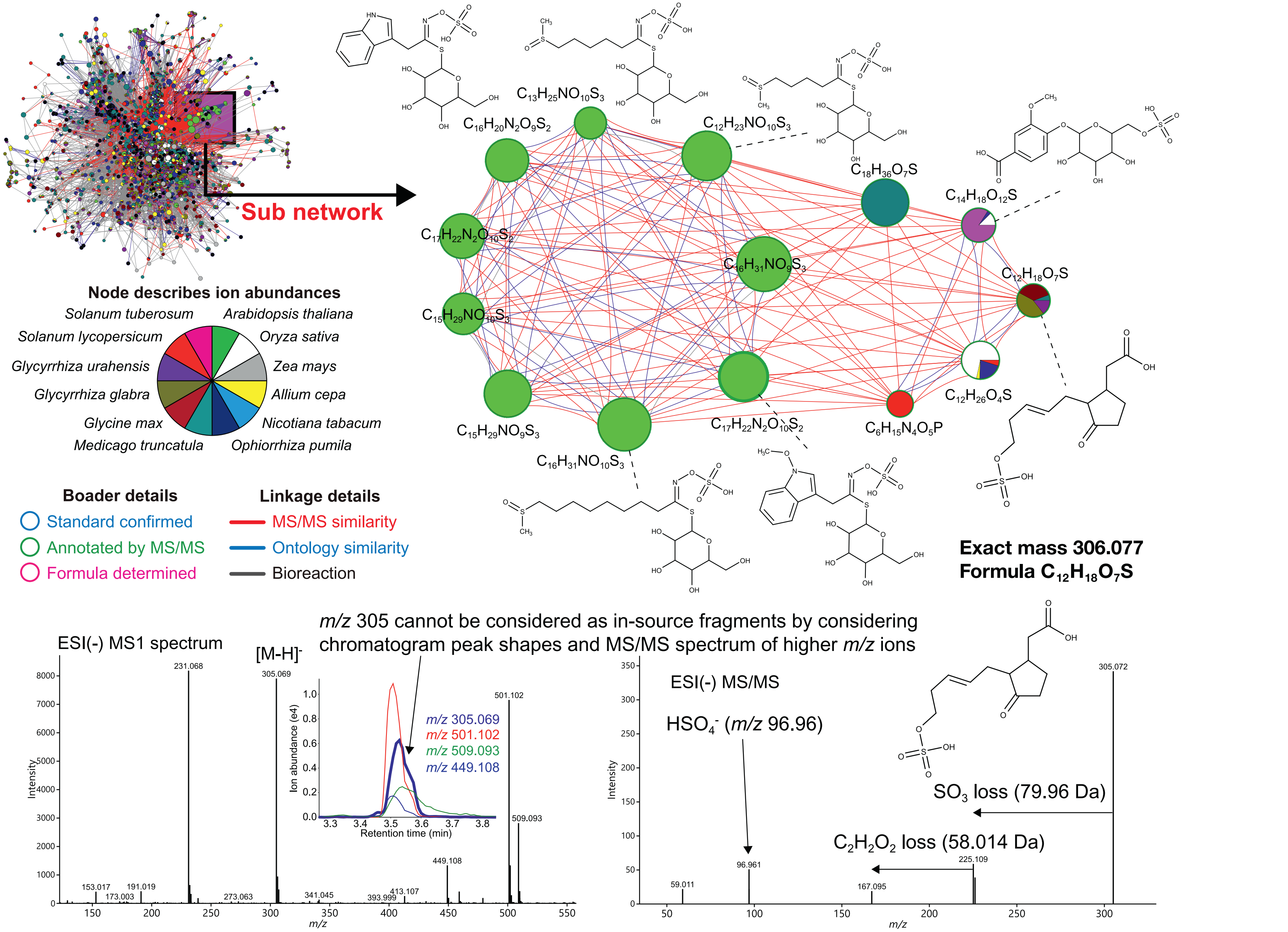

Fig. 9 Characterization of 12-hydroxyjasmonic acid sulfate by computational mass spectrometry. A unique cluster was found in the plant metabolome network using negative ion mode data. The cluster contained glucosinolates which were linked by the high similarity of MS/MS spectra based on *m/z* 96.96 (HSO_4^-). An unknown molecule, *m/z* 305.069, was also connected to the glucosinolate group suggesting the existence of sulfate motif in the structure. Although the ion was not detected in positive ion mode data, the feature of *m/z* 305.069 was determined as the undegraded ion (not in-source fragment) due to its unique chromatographic peak shape in the retention time range. Furthermore, the carbon number was determined as 12 from the integrated analysis of ^{12}C and ^{13}C plant data. After the molecular formula was predicted as $C_{12}H_{18}O_7S$, 12-hydroxyjasmonic acid sulfate was assigned as the top candidate from the metabolome structure database in MS-FINDER software. Modified from A Cheminformatics Approach to Characterize Metabolomes in Stable Isotope-Labeled Organisms. *Nat. Methods*. **2019**. 10.1038/s41592-019-0358-2.

9.3 *Showcase 3*: PlaSMA ID-611 (neg), $C_{12}H_{18}O_7S$ 12-Hydroxyjasmonic Acid Sulfate $[M-H]^-$

In the plants of the family Fabaceae, including *M. truncatula*, *Glycine* max, *Glycyrrhiza uralensis*, and *Glycyrrhiza glabra*, we found an unknown metabolite with *m/z* 305.069 in negative ion mode, linked to glucosinolates with high MS/MS similarities in the plant metabolome network (Fig. 9). Glucosinolates are known as the metabolites generating *m/z* 96.96 (HSO_4^-) as the characteristic ion, and the product ion was also detected in the MS/MS spectrum of unknown metabolite. In addition, the carbon content has been determined as 12 from the set of ^{12}C and ^{13}C labeling data. It indicates that the unknown metabolite is described as $C_{12}H_aN_bO_{>3}P_cS_{>0}$, which was finally determined as $C_{12}H_{18}O_7S$ by MS-FINDER program. Only one structure, 12-hydroxyjasmonic acid sulfate, could be retrieved from the structure databases. All MS/MS spectral peaks can be resolved by HR rules, and the spectrum pattern is similar to that of the spectrum shown in the previous report (the slight difference is explained by the difference between the machine and conditions).[48] This result showed that the integrated plant metabolome network gives us the opportunity to elucidate the structures of unknown natural products as GNPS does.

10 Conclusion

Metabolite structure is an essential information to interpret its physiological functions. Although the metabolomics for primary and lipid metabolites has been well-developed with a rich source of standard compounds and theoretical MS/MS libraries, the metabolic profiling of plants is still not complete owing to their chemical diversity. As it is important to decode the MS/MS spectrum for the structure elucidations, the database containing the spectra of various metabolites is essential to understand the relationship between the structure and its spectrum. Many great efforts to accumulate spectral records have been performed so far in MassBank, MassBank-EU (https://massbank.eu/MassBank/), GNPS, MoNA (http://mona.fiehnlab.ucdavis.edu/), ReSpect, PlaSMA, and CASMI[49] projects in addition to commercialized databases such as NIST and Metlin. Nowadays, a total of 321,616 experimental MS/MS spectra from 11,969 compounds are available for spectral searches, machine learnings, and seeds for molecular networking. Computational mass spectrometry, also known as mass spectrometry informatics, is the emerging technology in metabolomics for the global identification of metabolites in living organisms and human exposome. The advances in this field would facilitate the dereplication for natural products as well as the discovery of novel metabolites from plants.

Reference

1. Goering, A. W.; Haines, R. R.; Labeda, D. P.; Tchalukov, K. A.; Albright, J. C.; Doroghazi, J. R.; Metcalf, W. W.; Kelleher, N. L.; Ju, K.-S. A Roadmap for Natural Product Discovery Based on Large-Scale Genomics and Metabolomics; *Nat. Chem. Biol.* **2014**, *10*, 963–968.
2. Wishart, D. S. Emerging Applications of Metabolomics in Drug Discovery and Precision Medicine; *Nat. Rev. Drug Discov.* **2016**, *15*, 473–484.
3. Cao, L.; Pevzner, P. A.; Mohimani, H.; Gurevich, A.; Shcherbin, E.; Shlemov, A.; Korobeynikov, A.; Dorrestein, P. C.; Mikheenko, A.; Nothias, L.-F. Dereplication of Microbial Metabolites through Database Search of Mass Spectra; *Nat. Commun.* **2018**, *9*, 1–12.
4. Domingo-Almenara, X.; Montenegro-Burke, J. R.; Benton, H. P.; Siuzdak, G. Annotation: A Computational Solution for Streamlining Metabolomics Analysis; *Anal. Chem.* **2018**, *90*, 480–489.
5. Tsugawa, H. Advances in Computational Metabolomics and Databases Deepen the Understanding of Metabolisms; *Curr. Opin. Biotechnol.* **2018**, *54*, 10–17.
6. Blaženović, I.; Kind, T.; Ji, J.; Fiehn, O. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics; *Metabolites* **2018**, *8*, 31.
7. Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: Data-Independent MS/MS Deconvolution for Comprehensive Metabolome Analysis; *Nat. Methods* **2015**, *12*, 523–526.
8. Lai, Z.; Tsugawa, H.; Wohlgemuth, G.; Mehta, S.; Mueller, M.; Zheng, Y.; Ogiwara, A.; Meissen, J.; Showalter, M.; Takeuchi, K.; et al. Identifying Metabolites by Integrating Metabolome Databases with Mass Spectrometry Cheminformatics; *Nat. Methods* **2018**, *15*, 53–56.
9. Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihira, D.; Tanaka, W.; Cajka, T.; Saito, K.; Fiehn, O.; Arita, M. Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software; *Anal. Chem.* **2016**, *88*, 7946–7958.
10. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data; *BMC Bioinformatics* **2010**, *11*, 395.
11. Mahieu, N. G.; Genenbacher, J. L.; Patti, G. J. A Roadmap for the XCMS Family of Software Solutions in Metabolomics; *Curr. Opin. Chem. Biol.* **2016**, *30*, 87–93.
12. Böcker, S.; Dührkop, K. Fragmentation Trees Reloaded; *Aust. J. Chem.* **2016**, *8*, 1–26.
13. Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Searching Molecular Structure Databases With Tandem Mass Spectra Using CSI:FingerID; *Proc. Natl. Acad. Sci.* **2015**, *112*, 12580–12585.
14. Majumder, E. L.-W.; Guijas, C.; Siuzdak, G.; Domingo-Almenara, X.; Montenegro-Burke, J. R.; Benton, H. P. Autonomous METLIN-Guided in-Source Fragment Detection Increases Annotation Confidence in Untargeted Metabolomics; *Anal. Chem.* **2019**, *91*, 3246–3253.
15. Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M.; et al. Identification of Small Molecules Using Accurate Mass MS/MS Search; *Mass Spectrom. Rev.* **2017**, *9999*, 1–20.
16. Tsugawa, H.; Nakabayashi, R.; Mori, T.; Yamada, Y.; Takahashi, M.; Rai, A.; Sugiyama, R.; Yamamoto, H.; Nakaya, T.; Yamazaki, M.; Kooke, R.; Bac-Molenaar, J.; Oztolan-Erol, N.; Keurentjes, J.; Arita, M.; Saito, K. A Cheminformatics Approach to Characterize Metabolomes in Stable Isotope-Labeled Organisms; *Nat. Methods* **2019**, in press. https://doi.org/10.1038/s41592-019-0358-2.
17. Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J. M.; et al. Mass Spectral Molecular Networking of Living Microbial Colonies; *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E1743–E1752.
18. Saito, K.; Yonekura-Sakakibara, K.; Nakabayashi, R.; Higashi, Y.; Yamazaki, M.; Tohge, T.; Fernie, A. R. The Flavonoid Biosynthetic Pathway in Arabidopsis: Structural and Genetic Diversity; *Plant Physiol. Biochem.* **2013**, *72*, 21–34.
19. Mahieu, N. G.; Patti, G. J. Systems-Level Annotation of a Metabolomics Data Set Reduces 25000 Features to Fewer than 1000 Unique Metabolites; *Anal. Chem.* **2017**, *89* (19), 10397–10406.

20. Moorthy, A. S.; Wallace, W. E.; Kearsley, A. J.; Tchekhovskoi, D. V.; Stein, S. E. Combining Fragment-Ion and Neutral-Loss Matching during Mass Spectral Library Searching: A New General Purpose Algorithm Applicable to Illicit Drug Identification; *Anal. Chem.* **2017**, *89* (24), 13261–13268.
21. van der Hooft, J. J. J.; Wandy, J.; Barrett, M. P.; Burgess, K. E. V.; Rogers, S. Topic Modeling for Untargeted Substructure Exploration in Metabolomics; *Proc. Natl. Acad. Sci.* **2016**, *113* (48), 13738–13743.
22. Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; et al. Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking; *Nat. Biotechnol.* **2016**, *34* (8), 828–837.
23. Banerjee, P.; Erehman, J.; Gohlke, B. O.; Wilhelm, T.; Preissner, R.; Dunkel, M. Super Natural II-a Database of Natural Products; *Nucleic Acids Res.* **2015**, *43* (D1), D935–D939.
24. Nakabayashi, R.; Saito, K. Ultrahigh Resolution Metabolomics for S-Containing Metabolites; *Curr. Opin. Biotechnol.* **2017**, *43*, 8–16.
25. Kind, T.; Fiehn, O. Seven Golden Rules for Heuristic Filtering of Molecular Formulas Obtained by Accurate Mass Spectrometry; *BMC Bioinformatics* **2007**, *8*, 105.
26. da Silva, R. R.; Dorrestein, P. C.; Quinn, R. A. Illuminating the Dark Matter in Metabolomics; *Proc. Natl. Acad. Sci.* **2015**, *112*, 12549–12550.
27. Creek, D. J. Stable Isotope Labeled Metabolomics Improves Identification of Novel Metabolites and Pathways; *Bioanalysis* **2013**, *5*, 1807–1810.
28. Afendi, F. M.; Okada, T.; Yamazaki, M.; Hirai-Morita, A.; Nakamura, Y.; Nakamura, K.; Ikeda, S.; Takahashi, H.; Altaf-Ul-Amin, M.; Darusman, L. K.; et al. KNApSAcK Family Databases: Integrated Metabolite-Plant Species Databases for Multifaceted Plant Research; *Plant Cell Physiol.* **2012**, *53*, 1–12.
29. Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H. Z.; Xu, X. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology; *PLoS One* **2013**, *8* (4), 1–10.
30. Schläpfer, P.; Zhang, P.; Wang, C.; Kim, T.; Banf, M.; Chae, L.; Dreher, K.; Chavali, A.; Nilo-Poyanco, R.; Bernard, T.; Kahn, D.; Rhee, S. Y. K. Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants; *Plant Physiol.* **2017**, *173*, 2041–2059.
31. Scalbert, A.; Andres-Lacueva, C.; Arita, M.; Kroon, P.; Manach, C.; Urpi-Sarda, M.; Wishart, D. Databases on Food Phytochemicals and Their Health-Promoting Effects; *J. Agric. Food Chem.* **2011**, *59*, 4331–4348.
32. Ntie-Kang, F.; Telukunta, K. K.; Döring, K.; Simoben, C. V.; Moumbock, A. F. A.; Malange, Y. I.; Njume, L. E.; Yong, J. N.; Sippl, W.; Günther, S. NANPDB: A Resource for Natural Products from Northern African Sources; *J. Nat. Prod.* **2017**, *80*, 2067–2076.
33. Letzel, T.; Bayer, A.; Schulz, W.; Heermann, A.; Lucke, T.; Greco, G.; Grosse, S.; Schüssler, W.; Sengl, M.; Letzel, M. LC–MS Screening Techniques for Wastewater Analysis and Analytical Data Handling Strategies: Sartans and their Transformation Products as an Example; *Chemosphere* **2015**, *137*, 198–206.
34. Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E. A.; Glass, C. K.; Merrill, A. H.; Murphy, R. C.; Raetz, C. R. H.; Russell, D. W.; Subramaniam, S. LMSD: LIPID MAPS Structure Database; *Nucleic Acids Res.* **2006**, *35*, D527–D532.
35. Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The Human Metabolome Database for 2018; *Nucleic Acids Res.* **2017**, *1*, 1–10.
36. Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved Services and an Expanding Collection of Metabolites; *Nucleic Acids Res.* **2016**, *44*, D1214–D1219.
37. Jeffryes, J. G.; Colastani, R. L.; Elbadawi-Sidhu, M.; Kind, T.; Niehaus, T. D.; Broadbelt, L. J.; Hanson, A. D.; Fiehn, O.; Tyo, K. E. J.; Henry, C. S. MINEs: Open Access Databases of Computationally Predicted Enzyme Promiscuity Products for Untargeted Metabolomics; *Aust. J. Chem.* **2015**, *7*, 44.
38. 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.
39. Sawada, Y.; Nakabayashi, R.; Yamada, Y.; Suzuki, M.; Sato, M.; Sakata, A.; Akiyama, K.; Sakurai, T.; Matsuda, F.; Aoki, T.; et al. RIKEN Tandem Mass Spectral Database (ReSpect) for Phytochemicals: A Plant-Specific MS/MS-Based Data Resource and Database; *Phytochemistry* **2012**, *82*, 38–45.
40. Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; et al. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide; *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 15545–15550.
41. Kim, S.-Y.; Volsky, D. J. PAGE: Parametric Analysis of Gene Set Enrichment; *BMC Bioinformatics* **2005**, *6*, 144.
42. Qiu, F.; Fine, D. D.; Wherritt, D. J.; Lei, Z.; Sumner, L. W. Plant MAT: A Metabolomics Tool for Predicting the Specialized Metabolic Potential of a System and for Large-Scale Metabolite Identifications; *Anal. Chem.* **2016**, *88*, 11373–11383.
43. Koellensperger, G.; Guijas, C.; Benton, H. P.; Huan, T.; Wolan, D. W.; Warth, B.; Aisporna, A. E.; Hermann, G.; Domingo-Almenara, X.; Spilker, M. E.; et al. METLIN: A Technology Platform for Identifying Knowns and Unknowns; *Anal. Chem.* **2018**, *90*, 3156–3164.
44. Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; et al. ClassyFire: Automated Chemical Classification With a Comprehensive, Computable Taxonomy; *Aust. J. Chem.* **2016**, *8*, 1–20.
45. Grapov, D.; Wanichthanarak, K.; Fiehn, O. MetaMapR: Pathway Independent Metabolomic Network Analysis Incorporating Unknowns; *Bioinformatics* **2015**, *31*, 2757–2760.
46. Morreel, K.; Saeys, Y.; Dima, O.; Lu, F.; Van de Peer, Y.; Vanholme, R.; Ralph, J.; Vanholme, B.; Boerjan, W. Systematic Structural Characterization of Metabolites in Arabidopsis via Candidate Substrate-Product Pair Networks; *Plant Cell* **2014**, *26*, 929–945.
47. Pollier, J.; Morreel, K.; Geelen, D.; Goossens, A. Metabolite Profiling of Triterpene Saponins in Medicago Truncatula Hairy Roots by Liquid Chromatography Fourier Transform Ion Cyclotron Resonance Mass Spectrometry; *J. Nat. Prod.* **2011**, *74* (6), 1462–1476.
48. Gidda, S. K.; Miersch, O.; Levitin, A.; Schmidt, J.; Wasternack, C.; Varin, L. Biochemical and Molecular Characterization of a Hydroxyjasmonate Sulfotransferase from Arabidopsis Thaliana; *J. Biol. Chem.* **2003**, *278* (20), 17895–17900.
49. Schymanski, E. L.; Ruttkies, C.; Krauss, M.; Brouard, C.; Kind, T.; Dührkop, K.; Allen, F.; Vaniya, A.; Verdegem, D.; Böcker, S.; et al. Critical Assessment of Small Molecule Identification 2016: Automated Methods; *Aust. J. Chem.* **2017**, *9*, 22.