Article

PlantMAT: A Metabolomics Tool for Predicting the Specialized Metabolic Potential of a System and for Large-scale Metabolite Identifications

Feng Qiu, Dennis D. Fine, Daniel John Wherritt, Zhentian Lei, and Lloyd W. Sumner

Anal. Chem., **Just Accepted Manuscript** • DOI: 10.1021/acs.analchem.6b00906 • Publication Date (Web): 09 Nov 2016

PlantMAT: A Metabolomics Tool for Predicting the Specialized Metabolic Potential of a System and for Large-scale Metabolite Identifications

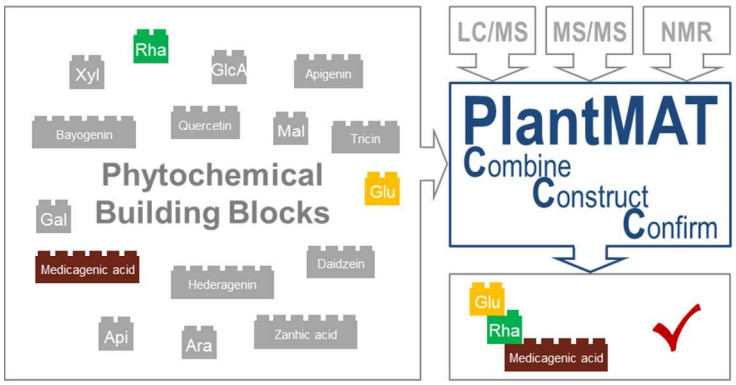
Feng Qiu,[1,2] Dennis D. Fine,[1] Daniel J. Wherritt,[1,3] Zhentian Lei,[1,2] and Lloyd W. Sumner[1,2,*]

[1]Plant Biology Division, The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401

[2]Department of Biochemistry, University of Missouri, Schweitzer Hall, Columbia, MO 65211

[3]Department of Chemistry, University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249

ABSTRACT: Custom software entitled Plant Metabolite Annotation Toolbox (PlantMAT) has been developed to address the number one grand challenge in metabolomics which is the large-scale and confident identification of metabolites. PlantMAT uses informed phytochemical knowledge for the prediction of plant natural products such as saponins and glycosylated flavonoids through combinatorial enumeration of aglycone, glycosyl, and acyl subunits. Many of the predicted structures have yet to be characterized and are absent from traditional chemical databases, but have a higher probability of being present in planta. PlantMAT allows users to operate an automated and streamlined workflow for metabolite annotation from a user-friendly interface within Microsoft Excel, a familiar, easily-accessed program for chemists and biologists. The usefulness of PlantMAT is exemplified using UHPLC-ESI-QToFMS/MS metabolite profiling data of saponins and glycosylated flavonoids from the model legume, *Medicago truncatula*. The results demonstrate PlantMAT substantially increases the chemical/metabolic space of traditional chemical databases. Ten of the PlantMAT predicted identifications were validated and confirmed through the isolation of the compounds using UHPLC-MS-SPE followed by de novo structural elucidation using 1D/2D-NMR. It is further demonstrated that PlantMAT enables the dereplication of previously identified metabolites, and is also a powerful tool for the discovery of structurally novel metabolites.



Metabolomics is a growing and maturing "omics" research field focused on the large-scale qualitative and quantitative profiling of small molecular metabolites. The number one, grand challenge of metabolomics is the large-scale, confident, chemical identification of metabolites.[1] Liquid chromatography coupled with mass spectrometry (LC-MS) is a powerful platform for metabolomics studies due to its high sensitivity, selectivity, and coverage of metabolites.[2] While LC-MS can generate extensive metabolomics data, manual interpretation of large-scale and complex metabolomics data is labor-intensive, time-consuming, and error-prone.

Chemical and spectral libraries have been increasingly used for compound identification and significantly increase the efficiency of large-scale metabolite annotation. Searching chemical libraries using accurate mass or molecular formula is routinely used as a first step in compound identification. A single accurate mass or molecular formula typically yields multiple matched near isobaric or isomeric compounds, and hence ambiguous results, which need refinement using other orthogonal data such as chromatographic retention time or additional structural identification methods such as tandem mass spectrometry (MS/MS).[2] Experimental MS data of analytes can be queried against MS/MS spectral libraries of authentic compounds such as those in MassBank,[3] METLIN,[4] and NIST libraries.[5] However, reference MS/MS data for natural products are often limited in these libraries and manual interpretation of MS/MS spectra requires substantial time and expertise.

Recently, specific computational tools have been developed for in silico interpretation of MS/MS spectra, such as MetFrag,[6] MAGMa,[7] and CFM-ID.[8] These tools first search molecule structure libraries, then perform in silico MS/MS fragmentation for the matches, and finally rank the candidates by measuring the similarity between the predicted and measured MS/MS spectra. These tools are effective, but ultimately fail and/or provide erroneous identifications when the

correct metabolites are not present in the searched databases; as in often the case of novel specialized metabolites. Library-based identification of plant metabolites is particularly challenging due to the limited number of structures and spectral data of authentic standards contained within the libraries.[9] It is estimated that over 200,000 different specialized metabolites are synthesized throughout the plant kingdom; however, only about 10–20% of anyplant metabolome can be identified using current technologies.[10,11] Many of the previously characterized plant metabolites are not even commercially available, and thus, no experimental data can be acquired for spectral comparison. Moreover, searches of libraries will, at the best, dereplicate known compounds, but will not identify novel compounds. Therefore, it is necessary to develop additional predictive computational tools to increase the chemical coverage space of available libraries; enabling more efficient, accurate structure dereplication and prediction. The recent development of CSI:FingerID makes it possible to identify chemical structures without searching spectral libraries. Specifically, it enables the prediction of unknown structures based on the machine learning of MS/MS data of identified structures, while it is currently limited to the prediction using positive-ionization ESI-MS/MS spectra.[12] Another example is the integration of molecular networking with an extensive in silico MS/MS spectral database which significantly expanded the chemical space searchable by MS/MS-based dereplication.[13] This approach was demonstrated with the identification of natural products including prenylated stilbenes in various *Macaranga* spp. and SAHA metabolites in the culture broth of a penicillium sp. strain.

To further enhance the identification of plant specialized metabolites, we introduce here a custom software entitled Plant Metabolite Annotation Toolbox (PlantMAT) which uses informed phytochemical knowledge for the identification of specialized metabolites through combinatorial enumeration of metabolic building blocks. Our combinatorial enumeration method originates

from the calculation of molecular formula using brute-force iteration.[14] Instead of using elements in the calculation, we use predefined substructures or "building blocks" in enumeration to predict the presence of structural components in metabolites. A similar method was previously applied in the generation of hypothetical flavonoids, where the calculations were only based on the molecular masses which could produce a large number of candidates.[15] Thus, additional steps of candidate refinement and validation are critically necessary to reduce the candidate list so that the identification accuracy can be improved. Here, we expand such methods to our identification pipeline of plant specialized metabolomics for the systematic annotation of saponins and glycosylated flavonoids.

Plant glycosides such as saponins and glycosylated flavonoids are structurally diverse compounds which consist of an aglycone conjugated with various glycosyl and acyl groups. Essentially, many of the basic building blocks associated with plant specialized metabolism biosynthesis are well known and can be used to predict combinatorial possibilities based upon informed phytochemistry, i.e., knowledge of how natural products are functionally modified and conjugated with other molecules,[16] to better define and predict the potential chemical/metabolic space and compounds potentially found within an organism. The in silico predictions include both hypothetical and previously identified structures which can be merged into a custom database searchable by accurate mass resulting in greater opportunities to identify novel metabolites. More importantly, we implement an orthogonal method, i.e., MSMS, to refine numerous candidates generated in combinatorial enumeration. The application is demonstrated in the annotation of metabolites observed in the UHPLC-ESI-QToFMS/MS data for methanol extracts of *Medicago truncatula* which is a leading model legume species for plant biology studies. As a result, PlantMAT enabled putative identification of 59 saponins and 14

glycosylated flavonoids. Ten of these compounds were purified and concentrated using UHPLC-MS-solid-phase extraction and their unambiguous structures were elucidated by 1D and 2D-NMR spectroscopy. The NMR confirmed and validated that all (100%) of PlantMAT predictions were correct.

METHODS AND TOOLS

Software Development. PlantMAT is an XLSM extension developed with Visual Basic for Applications 7.0 (VBA) in Microsoft Excel 2010. PlantMAT can be executed within Microsoft Excel 2010 and later versions running in Windows environment; including Excel 2013 and 2016. Opening the application provides an introductory page containing a descriptive overview of the functions as well as a graphical representation of the workflow. Next to this page is a spreadsheet-based aglycone library which stores the chemical information for each plant metabolite, including common name, chemical class, molecular formula, exact mass, plant origin (genus/species name) and textual identifier of the chemical structure using universal simplified molecular-input line-entry system (SMILES strings). Compounds can be manually added to this library or imported from other chemical libraries/databases. PlantMAT also includes a spreadsheet-based user interface (see Figure S1) where a single query, including the exact mass and MS/MS spectrum (*m/z* and intensity values), can be manually entered in the query form. In situations where multiple queries are needed, a batch job can be submitted by importing the peak list in either a comma delimited text file (CSV) or an Excel spreadsheet file (XLSX) as well as the MS/MS data in text file in which each line represents a fragment ion in the form of *m/z* value and intensity delimited by a space. Finally, the annotation results are provided in a spreadsheet which can be printed and/or exported as a CSV file for further data manipulation. The calculation

procedures for data analysis (see details in Results and Discussion) were programmed in a VBA-code module. All macros are activated by buttons, and pop-up dialog boxes are used to specify important parameters for individual processing steps. PlantMAT is available free of charge under a Creative Commons CC-BY-SA-NC license at https://sourceforge.net/projects/plantmat/.

Instrumentation and Analytical Methods. Metabolite profiling of 80% methanol / 20% water (v/v) extracts of *Medicago truncatula* (Jemalong A17 genotype) aerial and root materials was performed using a Waters ACQUITY UPLC (Waters Co., Milford, MA) coupled to a Bruker maXis Impact ESI-QToFMS system having a mass resolution of ~40,000 (Bruker Daltonics, Billerica, MA). The acquired data were first processed using Bruker's Compass DataAnalysis 4.1 software as follows. Peak deconvolution was initially performed using the 'Dissect' function and then molecular formulae were predicted using the 'SmartFormula' function which uses accurate mass and isotope ratios to refine the molecular formula prediction. The resultant peak list was exported as either a CSV or an XLSX file containing five columns including: peak number, retention time, peak area, exact mass, and molecular formula. The MS/MS spectral data of individual peaks were also exported as a XLSX file using a VB-code script provided within the DataAnalysis software and all the files were saved in the same folder. The Bruker pre-processed data, including peak list and MS/MS spectra, were imported into PlantMAT for a batch prediction of structures. Putative identities of the selected metabolites were then confirmed by de novo structural elucidation using 1D- and 2D-NMR, following automated, UHPLC-solid-phase extraction (SPE) of the selected metabolites using a Bruker/Spark Holland Prospekt II SPE system. Details and protocols of above procedures are provided in Supporting Information.

RESULTS AND DISCUSSION

PlantMAT involves three calculation steps as illustrated in Figure 1. Initially, a combinatorial enumeration process is performed to determine all possible structural possibilities of glycosides using components including aglycones, glycosyl, and acyl groups, as well as all possible glycosyl sequences. In the second step, all combinatorial possibilities are scrutinized and refined by comparing the experimentally measured MS/MS fragment ions to the predicted fragment ions resulting from the combinatorial neutral losses of glycosyl and acyl moieties. The final step involves the scoring and ranking of all possible glycosyl sequences based on the sequential loss of glycosyl and acyl moieties in MS/MS fragmentation. The calculation procedures for each step are detailed below.

Step 1: Combinatorial Enumeration of Aglycone, Glycosyl, and Acyl Groups. We use informed phytochemistry to better predict the potential specialized metabolic space for plant glycosides such as saponins and glycosylated flavonoids through in silico combinatorial enumeration of their building blocks or substructures, i.e., various aglycones, glycosyl, and acyl groups. This process involves the calculation of aglycone/glycosyl/acyl compositions followed by the generation of glycosyl sequences (Figure 2A).

(1) Calculation of all possible aglycone/glycosyl/acyl compositions: This step is similar to the calculation of elemental compositions using accurate mass. The total number of compositional possibilities increases with mass and aglycone/glycosyl/acyl varieties. Therefore, the users are enabled to define a set of constraints that place limits on the calculation, including (i) a set of allowed glycosyl/acyl units for the resulting compositions; (ii) the number of each glycosyl/acyl units (n_{max}) to be allowed in the resulting compositions; (iii) the total counts of all glycosyl/acyl units (N_{max}) to be allowed in the resulting compositions; (iv) searching criteria for

the aglycone structures, including the range of molecule weight, chemical class, and plant sources; and (v) calculation parameters related to the MS experiments, including mass error tolerance (σ in ppm), ionization mode, and adduct ions. The calculation process begins by importing the empirically measured *m/z* of the analyte. Initially, the molecular mass (M_m) is calculated according to the *m/z* value of the precursor ion as well as the mode of ionization and type of adduct ion as specified by the user. We use brute-force iteration to determine all possible glycosyl/acyl units and the aglycone. As shown in Figure 2A, the number of each glycosyl/acyl unit enumerates all integers from 0 to n_{max} while the total counts (N_g) of all glycosyl/acyl units is limited to N_{max}. For each enumeration, a mass balance (M_r) is calculated using the following equation:

$$M_r = M_m + N_g \times M_w - M_g \quad (1)$$

where M_g is the overall mass of glycosyl/acyl units and M_w is the mass of a water molecule (i.e., 18.0106 Da). When the mass balance is within the predefined range of molecular weight, it is then searched in the aglycone library for the metabolites within the mass error tolerance σ. If any matching aglycone is found, the name of the aglycone is returned with the counts of each glycosyl/acyl unit.

(2) Generation of all possible sugar chains: Plant specialized metabolites such as saponins and glycosylated flavonoids usually contain one to six monosaccharides. Sugar chains of these compounds can be highly diverse due to the order of monosaccharide units as well as a variety of glycosidic linkages in either linear or branched chains. In addition, many triterpenes and flavonoids have multiple hydroxyl and/or carboxylic groups all of which are potential positions for glycosylation. This results in additional isomeric possibilities related to the different positions of glycosyl and acyl moieties. It has been reported that for some isomeric flavonoid glycosides

the intensities of ions produced by the cleavage of glycosidic linkages were related to the positions of glycosyl moieties.[17] These studies were based on a few examples using a particular instrumental platform and thus, such fragmentation trends still require systematic and comprehensive evaluations. In addition, limited availability of authentic saponins for MS/MS experiments makes it difficult to investigate such spectra-structure correlations. In such case, MS is insufficient to provide confident positional identification and thus, positional isomers are not included in PlantMAT identifications. It is also necessary to note that, according to previously identified metabolites,[18] for triterpenes glycosylation occurs mostly at any one or two positions among C-3, C-24, and C-28. Incorporating this informed phytochemical knowledge in the structural elucidation can help refine the isomeric possibilities and prioritize the candidate list. Therefore, PlantMAT only considers the instances where the aglycone is attached with the maximum of two sugar chains in linear fashion.

For each group of aglycone/glycosyl/acyl units obtained from combinatorial enumeration, the algorithm reads the SMILES string of the aglycone structure, identifies all possible positions for glycosylation, such as hydroxyl and carboxyl groups, and calculates the total number of these functional groups (N_f). Then, permutation of all the glycosyl/acyl units is performed to generate all sequential possibilities when the aglycone is conjugated with a single sugar moiety. In case of more than one sugar moiety ($1 < x < N_f$), all the glycosyl/acyl units are then divided into x groups each containing n_i elements ($\sum_{i=1}^{x} n_i = N_g$), followed by permutation of the elements in each group. According to N_g, there may be more than one pattern in which the glycosyl units can be divided into x groups. For instance, three glycosyl units can be divided into two groups containing one and two elements (1, 2), respectively, whereas five glycosyl units can be divided into two groups in either pattern (1, 4) or (2, 3). As a result, the number of all sequential

possibilities is related to the number and variation of monosaccharides. Taking five different monosaccharides as an example, the calculation of the number of possible sequences is shown as following:

$$P(5, 5) = 120 \quad (2)$$

$$P(5, 1) \times P(4, 4) = 120 \quad (3)$$

$$P(5, 2) \times P(3, 3) = 120 \quad (4)$$

Equation (2) represents the number of possibilities with respect to an aglycone conjugated with a single sugar moiety. Equations (3) and (4) represent the number of possibilities when there are two separate sugar moieties attached to the aglycone. Thus, all together make the total number of 360 sequential possibilities. The foregoing permutations generate duplicates when there are the identical glycosyl units, and, therefore, it is necessary that these duplicates are identified and removed. For example, when there are three identical monosaccharides among the above five units, the total number of distinct possibilities is reduced to 360/3! = 60.

Step 2: Annotation of MS/MS Spectra. Combinatorial enumeration may generate a large number of candidates which can be refined using orthogonal analytical data. Tandem mass spectrometry (MS/MS) produces fragment ions of analytes, yielding much valuable information related to the chemical structures. Therefore, annotation of fragment ions helps with the structural elucidation of analytes. In MS/MS, fragment ions resulted from neutral losses are particularly characteristic and have been successfully used for structural identification of various small and macromolecules.[19–21] An exhaustive list of neutral losses in MS/MS was recently summarized and incorporated to a software entitled MS2Analyzer for molecule substructure annotation.[22] For saponins and glycosylated flavonoids, ESI-MS/MS analysis typically yields fragment ions corresponding to sequential loss of glycosyl/acyl units. Loss of CO_2 and/or H_2O

molecules is also commonly observed when carboxylic acid groups are present in the structures.[23,24] Such fragmentation trends have been also incorporated within PlantMAT logic for the annotation of MS/MS ions based on the calculated glycosyl/acyl components (Figure 3A). Subsequent to step 1, hypothetical neutral losses are generated through '*k*-combination' which selects 1 to *k* items ($k \leq N_g + 2$) from the set of candidate glycosyl/acyl units plus one CO_2 and one H_2O molecules, such that the order of selected items are not considered. This process is computationally achieved by using brute force iteration in which each unit enumerates from all integers from 1 to n_g (i.e., the number of each glycosyl/acyl unit as obtained in combinatorial enumeration). After each iteration, the masses of neutral losses are calculated and further used to calculate the *m/z* values of corresponding fragment ions. Each calculated *m/z* is searched in the measured MS/MS spectrum for the fragment ions above the specified noise level (minimum intensity) as well as within the mass error tolerance. The matched ions can, therefore, be putatively annotated based on the calculated neutral losses.

Step 3: Scoring and Ranking of Glycosyl Sequences. The next challenge, after the aglycone and glycosyl/acyl residues are identified, is to determine how these building blocks are configured to form the glycoside. Various MS/MS techniques have been applied to elucidate the glycosylation patterns of macromolecules such as glycans and glycoproteins.[25,26] Specifically, the glycosyl sequence may be elucidated from the pattern of fragment ions corresponding to the sequential glycosyl elimination. Therefore, we developed a strategy for MS/MS prediction and matching for the scoring and ranking of all possible glycosyl sequences (Figure 4A).

Initially, hypothetical neutral losses are generated by sequentially cleaving the *O*-glycosidic bond of each glycosyl/acyl moiety. These neutral losses are further processed to make combinations that correlate to the combined loss of glycosyl/acyl residues of different

glycosyl/acyl moieties. Then, all these neutral losses (i) are used to calculate the m/z values of primary fragment ions (x_{i1}). A list of secondary ions (x_{ij}) for each of the fragment ions x_{i1} is further generated with an additional loss of CO_2 and/or H_2O molecules. Finally, the predicted fragmentation (X) consisting of both primary and secondary ions is compared with the measured fragmentation (Y) of the analyte. The number of matched ions is an important measure for ranking the candidates. However, it might be insufficient to distinguish true and false candidates because different candidates may have the same number of matched ions. Generally, in tandem MS, ions with high intensity are more characteristic than ions with low intensity. Therefore, the intensity of matched ions is also taken into account in the calculation of matching scores. The algorithm first finds all spectrum peaks (m_i) in Y that match the m/z of the predicted fragment ions in X within a user-defined mass error tolerance and then computes the spectral matching score using the equation as in the previous report:[27]

$$\text{matching score} = \sum_{m_i \in Y} \log_{10}(f \times I) \qquad (5)$$

where I is the relative intensity of each matched ions. For each primary fragment ion (x_{i1}) and the corresponding secondary ions (x_{ij}), only the matched ions with the highest relative intensity is included in the calculation of the matching score. A factor (f) of 10,000 is used to ensure that a match to a peak of significant intensity (≥0.01% relative intensity) will not contribute negatively to the overall score. Finally, all the sequential candidates are ranked in the descending order of their matching scores. The candidates with higher scores have higher possibility of being the correct glycosyl sequence.

Demonstration of the Three-Step Workflow of PlantMAT. To demonstrate the effectiveness of PlantMAT, we first created an aglycone library containing triterpenes and flavonoids which were used to predict saponins and glycosylated flavonoids in Medicago and

Arabidopsis spp. Traditionally, comprehensive libraries such as PubChem and ChemSpider are frequently used for metabolite identification. However, for particular plant species, they include many unrelated compounds and the coverage of species-specific metabolites are often limiting and incomplete. Thus, searching these libraries frequently yield no matches or worse, provide erroneous identifications. By contrast, a species-specific library can be more efficient and the enhanced specificity improves the identification accuracy. An extensive literature survey on SciFinder found 52 triterpenes and 156 flavonoids which were previously identified in Medicago and Arabidopsis spp. The chemical information of these compounds including their common names, plant origins, molecular formula, and SMILES codes were then entered in the aglycone library of PlantMAT. An additional 184 oleanane-type triterpenes from other plant species were extracted from Handbook of Spectroscopic Data of Saponins[18] and also imported to PlantMAT for an enhanced metabolite coverage as Medicago is known to biosynthesize oleanane derived triterpenes. For the identification of saponins and glycosylated flavonoids, we included various glycosyl and acyl groups in combinatorial enumeration, including hexose (Hex), deoxyhexose (dHex), uronic acid (HexA), and malonic acid (MA). For the identification of acylated flavonoids, we also include several hydroxycinnamic acids, such as coumaric acid (CA), ferulic acid (FA), and sinapinic acid (SA).

Initially, we used experimental ESI-MS and MS/MS spectral data of an authentic compound, soyasaponin I [soyasapogenol B 3-*O*-*α*-L-rhamnopyranosyl-(1→2)-*β*-D-galactopyranosyl-(1→2)-*β*-D-glucopyranosiduronic acid], to demonstrate how the three-step workflow of PlantMAT can be used for the identification of saponin structures. Based on the exact mass, i.e., 942.5188 Da, the parameters for combinatorial enumeration were defined as following: the number of each glycosyl group $0 \leq n_g \leq 3$, the total number of glycosyl groups $N_g = 3$, the number of malonic

acid $0 \leq n_a < 1$, and the mass error tolerance for searching the aglycones $\sigma \leq 5$ ppm. Using the exact mass as input, ten different combinations were generated and given as a list in the output spreadsheet showing the name of the aglycone and the counts of each glycosyl/acyl units (Figures 2B and S2). The display of each combinatorial candidate is accompanied by a drop-down list showing the glycosyl sequential possibilities emerged from the permutation of monosaccharide units. Taking the combination of soyasapogenol B, Hex, dHex, and HexA as an example (Figure 2C), permutation of the three distinct monosaccharides generates six sequential possibilities for a trisaccharide moiety. In case of two sugar moieties attached to the aglycone, the three monosaccharides are divided into two groups following three patterns, i.e., (Hex, dHex-HexA), (dHex, Hex-HexA), and (HexA, Hex-dHex). For each of these patterns, the pair of monosaccharides undergoes additional round of permutation, resulting in two sequential possibilities related to a disaccharide.

Next, we used MS/MS data to refine the candidates emerged from the foregoing combinatorial enumeration. For each predicted glycosyl composition, all possible neutral losses were initially obtained through a *k*-combination of the monosaccharide units plus a CO_2 and a H_2O molecule. In case of Hex, dHex, and HexA, there are 31 combinatorial possibilities, i.e., $\sum_{k=1}^{5} C(5, k)$, which were used to generate a list of fragment ions represented by their *m/z* values. The calculated *m/z* were then searched in the measured MS/MS spectrum and the ions which satisfy the matching criteria (e.g., relative intensity > 0.5, *m/z* error tolerance ≤ 5 ppm) were merged to a drop-down list showing their *m/z*, relative intensity, and annotations (see Figure S1). According to the number of annotated ions, the combination of soyasapogenol B, Hex, dHex, and HexA matched best with the measured MS/MS spectrum and all four observed product ions were annotated with respect to the neutral loss of predicted glycosyl components (Figure 3B).

Specifically, the anion, at *m/z* 457, represents the deprotonated ion that corresponded to the triterpene aglycone, soyasapogenol B, resulting from the loss of all glycosyl moieties (−484 Da). One anion observed at *m/z* 733 resulted from the loss of a deoxyhexose, a CO_2, and a H_2O (−208 Da). Two additional anions observed at *m/z* 633 and 615 correlated to the loss of a hexose and a deoxyhexose (−308 Da), and an additional loss of a H_2O (−18 Da). These annotated ions are sufficient to confirm the presence of three distinct monosaccharides, i.e., Hex, dHex, and HexA. Thus, the other nine compositional possibilities can be confidently eliminated, leaving only soyasapogenol B, Hex, dHex, and HexA as probable components. The remaining challenge is the differentiation of monosaccharide stereoisomers, such as different hexoses including glucose, galactose and mannose. While it is still possible to elucidate the absolute configurations using LC-MS, it usually requires chiral separation and/or multistage tandem fragmentation of analytes which involves laborious sample preparation procedures and specific instrumentation techniques.[28–30] Chemical complexity of plant crude extracts containing large number of metabolites makes such determination even more challenging. Nevertheless, these difficulties can be confidently addressed by using NMR via various well-established homonuclear and heteronuclear correlation experiments. For example, the axial and equatorial positions of protons of sugar rings can be determined based on their multiplicity patterns and coupling constants, which lead to the differentiation of monosaccharide stereoisomers.

The final step is to refine isomeric glycosyl sequences corresponding to the three distinct monosaccharides. In case of a trisaccharide moiety such as -HexA-Hex-dHex, sequential cleavage of the glycosidic bonds may yield three different neutral losses, i.e., -dHex, -Hex-dHex, and -HexA-Hex-dHex (Figure 4B). In case of more than one sugar moiety, an additional step of *k*-combination is necessary to generate combined neutral losses of glycosyl units from different

sugar moieties. Using -HexA and -Hex-dHex as an example, the disaccharide moiety -Hex-dHex may yield two different neutral losses, i.e., -dHex and -Hex-dHex, both of which can be combined with the neutral loss of monosaccharide moiety -HexA, resulting in additional two possibilities including (-dHex, -HexA) and (-Hex-dHex, -HexA). For each glycosyl sequence, the *m/z* values were calculated based on the hypothetical neutral losses and then compared with the observed fragment ions in the measured MS/MS spectrum. Finally, all the sequential possibilities were listed in the descending order of their matching scores calculated by the logarithmic equation (Eq. 5). As shown in Figure 4C, four out of 12 candidates (33% of all candidates), including the correct sequence (-HexA-Hex-dHex), were equally ranked as the best matches with the identical scores of 13.77. Hypothetically, these four glycosyl sequences produce the same pattern of fragment ions, including *m/z* 733 $[M–dHex–CO_2–H]^-$, *m/z* 633 $[M–Hex–dHex–H]^-$, and *m/z* 457 $[M–HexA–Hex–dHex–H]^-$, all of which matched to the measured MS/MS spectrum. Finally, NMR can facilitate the further differentiation of these four glycosyl sequences as well as the elucidation of glycosidic linkages which are essential for the determination of unambiguous structure. As shown in the above example, the third step of PlantMAT workflow is useful for the sequencing of sugar chains which consists of different monosaccharides. However, when there are only identical monosaccharides, sequencing is not necessary and thus, only the first and second steps can be performed.

Application of PlantMAT for the Identification of Metabolites in *Medicago truncatula*. *M. truncatula* is a model species for legume biology and it is an excellent system to study the specialized metabolism of legumes.[31] Several studies describing the metabolite compositions of *M. truncatula* have been reported over the past decade. Metabolite profiling using LC-MS-based methods revealed the presence of a large number of saponins,[32–42] flavonoids,[43–48] and isoflavonoids[46] in various Medicago spp. tissues. These compounds have been found to exhibit

several important biological functions, such as allelopathy, poor digestibility in ruminants, deterrence to insect foraging, and antifungal properties.[49,50] In addition, plant species containing saponins have long been used in traditional medicines and have shown numerous pharmacological activities, such as anti-inflammatory, hemolytic, cholesterol lowering, and anticancer properties.[51,52] We applied PlantMAT in the systematic annotation of saponins and glycosylated flavonoids in *M. truncatula* extracts in order to further demonstrate its efficiency and accuracy in LC-MS-based metabolomics experiments.

Initially, UHPLC-ESI-QToFMS/MS was used to profile 80% aqueous methanol extracts of aerial and root materials of *M. truncatula* (Jemalong A17). The deconvoluted total ion chromatograms of the two materials showed a total of ~300 peaks within the range of retention time 2−20 min (Figure 5). ESI-QToFMS analysis indicated that the m/z values of these peaks ranged from 400 to 1400. Based on the average mass of triterpene aglycones (480±30 Da) and monosaccharides (170±20 Da), it can be assumed that saponins with molecular mass between 1300 and 1400 Da may contain up to six monosaccharides. Therefore, the parameters related to combinatorial enumeration were set as following: the number of each glycosyl group $0 \leq n_g \leq 6$, the total number of all glycosyl groups $0 \leq N_g \leq 6$, the number of each acyl group $0 \leq n_a \leq 1$, and the mass error tolerance $\sigma \leq 5$ ppm. Then, the measured exact masses and MS/MS spectra of the observed peaks were analyzed following the three-step workflow as described previously. A batch query was executed by importing the peak list containing the exact mass of each peak to PlantMAT. As a result, 125 peaks were initially highlighted as possible saponins or glycosylated flavonoids via combinatorial enumeration using their measured *m/z* values of precursor ions. In silico annotation of their MS/MS spectra predicted 59 saponins and 14 flavonoids (Tables S1 and S2) out of these 125 peaks. These saponins are glycosides of six different oleanane-type

triterpenes, including bayogenin, hederagenin, zanhic acid, medicagenic acid, soyasapogenol B, and soyasapogenol E. The structural diversity of these compounds is the result of modification varieties, e.g., glycosylation, acylation, etc., to aglycones. As shown, PlantMAT enabled a systematic, high-throughput identification of saponin and flavonoid compositions of *M. truncatula*. It is worth noting that the entire computational workflow took only ~10 minutes to putatively identify 73 out of more than 300 observed peaks, demonstrating the high performance of PlantMAT in batch processing of metabolomics data. Such comprehensive annotation also allows the evaluation of metabolite variety and distribution in different parts of plants which are important to understand their presence to various biological processes, such as plant growth, stress, defense, and plant-microbe interactions. Results indicate that zanhic acids (11 glycosides) are only produced in aerial parts of *M. truncatula*, whereas hederagenins (8) and soyasapogenol E (4) are unique to roots. Other aglycones are found in both aerial and root parts with different varieties. For example, soyasapogenol B are mainly distributed in aerial parts (8 vs. 2), whereas bayogenins are more common in roots (5 vs. 1). Medicagenic acids are the most abundant among all triterpene aglycones and found in both aerial and root parts (8 vs. 6).

Validation of PlantMAT Identifications Using NMR. Ten of the putatively identified peaks (Figure 5; Table 1) were isolated for structural elucidation using NMR in order to characterize the stereochemistry of these compounds and assess the identification accuracy of PlantMAT. The sensitivity of NMR is the major concern for the analysis of mass-limited natural products such as plant specialized metabolites. However, we have developed two strategies to overcome these challenges. First, targeted analytes are purified and concentrated in an automated fashion using repetitive (15−20) injections of the same extract, repetitive separations, and repetitive solid-phase extractions of target analytes using UPLC-MS-SPE. Targeted analytes were thus

repetitively separated and collected onto SPE cartridges so that sufficient materials (1−5 μg) could be obtained for NMR analyses. Second, recent developments in micro-cryoprobe technologies have dramatically improved detection sensitivity of NMR which enables the analysis of natural products at micromolar concentration levels.[53,54] Thus, for the ten selected peaks, both conventional 1D- and 2D-NMR spectra could be acquired with sufficient signal-to-noise ratio using a 600 MHz NMR equipped with a 1.7 mm cryoprobe. As demonstrated by the NMR identifications (Figures 6 and S2−S10), aglycone/glycosyl/acyl components of the ten peaks were correctly predicted through combinatorial enumeration followed by MS/MS annotation (Table S3). In instances where there are multiple different monosaccharide subunits, there could be numerous glycosyl sequential possibilities as shown by several peaks such as A-84, A-107, and A-145 (see Table 1). Nevertheless, MS/MS prediction and matching has good capability to distinguish true and false matches, and the correct glycosyl sequences were found in the top 1–7 candidates (Table 1). While it is not possible to assign unambiguous structures using MS/MS and PlantMAT only, they together offer an efficient approach for putative identification which is useful in dereplication or more efficient structural elucidations using additional analytical methods such as NMR.

The advantage of PlantMAT is the use of combinatorial enumeration with informed phytochemistry to expand the queryable chemical space of plant metabolites which improves the efficiency of dereplicating known metabolites and increases the possibility of identifying novel structures. This strategy uses known metabolites to intelligently predict the metabolic potential of a system and is not simply a library search tool. Combinatorial enumeration yields all structural possibilities related to a group of predefined substructures which cannot be obtained by searching the traditional chemical or spectral libraries. For example, for peaks A-46 and R-44,

neither a PubChem nor ChemSpider search of their molecular formulas found any structures matched to PlantMAT identifications, indicating they are likely novel compounds. These two peaks were isolated from aerial and root extracts of *M. truncatula*, respectively. As expected, structural elucidation confirmed that peak A-46 was a novel flavonoid glycoside [apigenin 4'-*O*-*β*-(2'-*E*-coumaroyl)glucuronopyranosyl-(1→2)-*O*-*β*-glucuronopyranoside] and peak R-44 was a novel saponin [bayogenin 3-*O*-*β*-glucopyranosyl-(1→4)-*O*-*β*-(3'-malonyl)glucopyranosyl-28-*O*-*β*-glucopyranoside]. Another example is related to the identification of glycosyl/acyl sequential isomers. Initially, the peak A-145 was putatively identified as a malonylated soyasapogenol B glycoside consisting of a hexose, a deoxyhexose, and an uronic acid. There are four probable glycosylation patterns which matched equally to the measured MS/MS spectrum, e.g., a malonylated trisaccharide -HexA-Hex-dHex-Mal. This peak was then isolated for structural elucidation and confirmed as soyasapogenol B 3-*O*-(4'-malonyl)rhamnopyranosyl-(1→2)-*O*-*β*-galactopyranosyl-(1→2)-*O*-*β*-glucuronopyranoside. However, PubChem or ChemSpider searches of its molecular formula $C_{51}H_{80}O_{21}$ found only an isomeric soyasapogenol B glycoside having a trisaccharide moiety (-HexA-Hex-dHex) at C-3 and a malonyl group at C-28 position. As these structurally related saponins may have similar spectral features, the unknown analyte could be erroneously assigned with an isomeric structure when using only public database searches for the candidate structures.

It is important to note that several factors, e.g., analyte concentration, ionization polarity, collision energy, and instrumental platforms, have great influence on the MS/MS fragmentation of analytes, and they may be optimized so that MS/MS can give the most abundant and characteristic fragments that may improve the refinement of candidate structures. We used the MS/MS data reported in peer-reviewed literature to further examine the fragmentation patterns of

glycosides on different instrumental platforms and further assess the prediction accuracy of PlantMAT. Initially, published MS/MS spectra were manually extracted from the literature and converted into digitalized formats as ion lists in text files. This included 180 MS/MS spectra of saponins from 4 Medicago spp. (*M. arabica*, *M. hybrid*, *M. sativa*, and *M. truncatula*) covering 3 different types of instrumental platforms including fast atom bombardment mass spectrometer,[34,36] ion trap mass spectrometer,[33,35,38,39] and Fourier transform ion cyclotron resonance mass spectrometer.[40] These data were then imported to PlantMAT for the identification using the workflow as described previously. Results indicate that PlantMAT predictions were fully consistent with the reported identifications which demonstrates that PlantMAT is compatible with the analyses of other and different instrumental platforms. It was also observed that MS/MS fragmentation of saponins appears to be highly similar on these platforms, showing predominantly neutral losses of glycosyl residues.

CONCLUSION

UHPLC-MS/MS-based metabolomics is a powerful technique for the large-scale and high-throughput metabolite profiling. However, a large proportion of confident metabolite identifications are currently still performed through manual interpretation of LC-MS data. In addition, searches of public database structures/spectral libraries is insufficient for full metabolome identifications within LC-MS profiles due to the limited metabolite coverage of searched libraries.

We introduce here a new computational tool, PlantMAT, for predicting the metabolic potential of a system and for large-scale metabolite identifications. It is demonstrated that combinatorial enumeration, which sets PlantMAT apart from other metabolite identification

tools, substantially increased the possibility of identifying specialized metabolites that are currently not present in conventional databases. Using orthogonal MS/MS spectral data, a large number of combinatorial possibilities can be reduced to a single or small set of putative identifications. The streamlined, automated workflow of PlantMAT enables batch processing and annotation of large-scale metabolomics data. These improvements can yield many benefits to metabolomics studies. For example, PlantMAT can be implemented within MS/MS-based plant metabolomics pipeline for the dereplication of previously identified metabolites as well as the search for structurally novel metabolites. Structural information obtained from PlantMAT is also helpful in final structural validation/elucidation using NMR spectroscopy. The application of PlantMAT in the systematic annotation of saponins and glycosylated flavonoids in *M. truncatula*, led to an increased number of metabolite identifications which are critical to improve our basic understanding of their biosynthesis and biological functions. We have further successfully applied PlantMAT to the identification of flavonoid conjugates in *Arabidopsis thaliana*, steroidal saponins in Asparagus, triterpene saponins in liquorice (*Glycyrrhiza uralensis*), triterpene saponins in soybean (*Glycine max*) and alkaloid saponins in potato (*Solanum tuberosum*).

PlantMAT also has additional potential to identify other classes of plant metabolites such as lignins[55] and complex polyphenols, e.g., tannins and procyanidians.[56] These compounds are composed of multiple identical or different subunits and, thus, combinatorial enumeration is a possible approach for their structural identification. For example, the basic structures of lignins are generally composed of three monolignol mononers including *p*-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol. These monolignols are incorporated in lignins as phenylpropanoid derivatives, i.e., *p*-hydroxyphenyl, guaiacyl, and syringyl subunits, respectively. These subunits can be used in combinatorial enumeration to predict the possible compositions of lignins. Then,

the in silico predictions can be refined by implementing the typical MS/MS fragmentation patterns of lignins in the PlantMAT algorithms.[57] The expanded chemical coverage of PlantMAT will further increase our ability to conceptualize and query the potential chemical space for the systematic annotation of plant metabolites.

ASSOICATED CONTENT

Ⓢ Supporting Information

Supporting Information Available: PlantMAT graphical user-interfaces and illustration of a single query; Experimental conditions for UHPLC-MS-SPE-NMR analyses; Putative identification of glycosides in *M. truncatula*; MS/MS spectra of glycosides used for the validation of PlantMAT; and ^{1}H NMR spectra of glycosides used for the validation of PlantMAT. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel: +1 (573) 882-5486. Fax: +1 (573) 224-4743. Email: sumnerlw@missouri.edu.

Notes

The authors declare no competing financial interest.

ALKNOWLEDGEMENTS

This work was supported by the NSF-JST Metabolomics for a Low Carbon Society Award #1139489, NSF MRI Award #1126719 for the NMR, and NSF RCN Award #1340058. We thank Alan Pilon and Treyon Grant at Noble Foundation for their assistance in the construction

of the metabolite libraries. The authors also gratefully acknowledge helpful UPLC-MS-SPE-NMR discussions with Aiko Barsch at Bruker Daltonik GmbH, Bremen,Germany, and Ulrich Braumann and Markus Godejohann from Bruker BioSpin GmbH, Rheinstetten, Germany.

REFEFENCES

(1) Tachibana, C. *Science* **2014**, *345*, 1519–1521.

(2) Hildebrandt, C.; Wolf, S.; Neumann, S. *J. Integr. Bioinform.* **2011**, *8*, 157, 1–16.

(3) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*, 703−714.

(4) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27*, 747−751.

(5) The National Institute of Standards and Technology. http://www.nist.gov/srd/nist1a.cfm. Aug. 9, 2016.

(6) Wolf, S.; Schmidt, S.; Muller-Hannemann, M.; Neumann, S. *BMC Bioinformatics* **2010**, *11*, 148, 1-12.

(7) Ridder, L.; van der Hooft, J. J. J.; Verhoeven, S.; de Vos, R. C. H.; Bino, R. J.; Vervoort, J. *Anal. Chem.* **2013**, *85*, 6033–6040.

(8) Allen F.; Pon, A.; Wilson M.; Greiner, R.; Wishart, D. *Nucleic Acids Res.* **2014**, *6*, W94–W99.

(9) Kind, T.; Liu, K.-H.; Lee, D. Y.; DeFelice, B.; Meissen, J. K.; Fiehn, O. *Nat. Methods* **2013**, *10*, 755−758.

(10) Pichersky, E.; Gang, D. *Trends Plant Sci.*, **2000**, *5*, 439–445.

(11) Binoa, R. J.; Hall, R. D.; Fiehn, O.; Kopka, J.; Saito, K.; Draper, J.; Nikolau, B. J.; Mendes, P.; Roessner-Tunali, U.; Beale, M. H.; Trethewey, R. N.; Lange, B. M.; Wurtele, E. S.; Sumner, L. W. *Trends Plant Sci.*, **2004**, *9*, 418–425.

(12) Dührkopa, K.; Shenb, H.; Meusela, M.; Rousub, J.; Böckera, S. *P. Natl. Acad. Sci. USA* **2015**, *112*, 12580–12585.

(13) Allard, P.-M.; Peresse, T.; Bisson, J.; Gindro, K.; Marcourt, L.; Pham, V. C.; Roussi, F.; Litaudon, M.; Wolfender, J.-L. *Anal. Chem.* **2016**, *88*, 3317–3323.

(14) Patiny, L.; Borel, A.; *J. Chem. Inf. Model.* **2013**, *53* 1223–1228.

(15) Zhang, M.; Sun, J.; Chen, P. *Anal. Chem.* **2015**, *87*, 9974–9981.

(16) Wang, X. *FEBS Letters* **2009**, *583*, 3303–3309.

(17) Geng, P.; Sun, J.; Zhang, R.; He, J.; Abliz, Z. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 1519–1524.

(18) Ahmad, V. U; Basha, A. In *Handbook of Spectroscopic Data or Saponins*, Taylor & Francis, United Kingdom, 2000, 993–2830.

(19) Tian, Q.; Giusti, M. M.; Stoner, G. D.; Schwartz, S. J. *J. Chromatogr. A* **2005**, *1091*, 72–82.

(20) Murphy, R. C.; James, P. F.; McAnoy, A. M.; Krank, J.; Duchoslav, E.; Barkley, R. M. *Anal. Biochem.* **2007**, *366*, 59–70.

(21) Ficarro, S. B.; McCleland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M. *Nat. Biotechnol.* **2002**, *20*, 301–305.

(22) Ma, Y.; Kind, T.; Yang, D.; Leon, C.; Fiehn, O. *Anal. Chem.* **2014**, *86*, 10724–10731.

(23) Muir, A. D.; Ballantyne, K. D.; Hall, T. W. In *Proceedings of the Phytochemical Society of Europe*, Springer, Germany, 2000, 45, 35–41.

(24) Lei, Z.; Jing, L.; Qiu, F.; Zhang, H.; Huhman, D.; Zhou, Z.; Sumner, L. W. *Anal. Chem.* **2015**, *87*, 7373–7381.

(25) Zaia, J. *Chem. Bio.* **2008**, *15*, 881–892.

(26) Dell, A.; Morris, H. R. *Science* **2001**, *291*, 2351–2356.

(27) Ibrahim, A.; Yang, L.; Johnston, C.; Liu, X.; Ma, B.; Magarvey, N. A. *P. Natl. Acad. Sci. USA* **2012**, *109*, 19196–19201.

(28) Desaire, H.; Leary, J. A. *Anal. Chem.* **1999**, *71*, 1997–2002.

(29) Nagy, G.; Pohl, N. L. B. *Anal. Chem.* **2015**, *87*, 4566–4571.

(30) Konda, C.; Londry, F. A.; Bendiak, B.; Xia, Y. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 1441–1450.

(31) Dixon, R. A.; Sumner, L. W. *Plant Physiol.* **2003**, *131*, 878–885.

(32) Bialy, Z.; Jurzysta, M.; Oleszek, W.; Piacente, S.; Pizza, C. *J. Agric. Food Chem.* **1999**, *47*, 3185–3192.

(33) Huhman, D. V.; Sumner, L. W. *Phytochemistry* **2002**, *59*, 347–360.

(34) Bialy, Z.; Jurzysta, M.; Mella, M.; Tava, Aldo. *J. Agric. Food Chem.* **2004**, *52*, 1095–1099.

(35) Kapusta, I.; Janda, B.; Stochmal, A.; Oleszek, W. *J. Agric. Food Chem.* **2005**, *53*, 7654–7660.

(36) Bialy, Z.; Jurzysta, M.; Mella, M.; Tava, A. *J. Agric. Food Chem.* **2006**, *54*, 2520–2526.

(37) Schliemann, W.; Ammer, C.; Strack, D. *Phytochemistry* **2008**, *69*, 112–146.

(38) Tava, A.; Mella, M.; Avato, P.; Biazzi, E.; Pecetti, L.; Bialy, Z.; Jurzysta, M. *J. Agric. Food Chem.* **2009**, *57*, 2826–2835.

(39) Tava, A.; Pecetti, L.; Romani, M.; Mella, M.; Avato, P. *J. Agric. Food Chem.* **2011**, *59*, 6142–6149.

(40) Pollier, J.; Morreel, K.; Geelen, D.; Goossens, A. *J. Nat. Prod.* **2011**, *74*, 1462–1476.

(41) Abbruscato, P.; Tosi, S.; Crispino, L.; Biazzi, E.; Menin, B.; Picco, A. M.; Pecetti, L.; Avato, P.; Tava, A. *J. Agric. Food Chem.* **2014**, *62*, 11030–11036.

(42) Huhman, D. V.; Berhow, M. A.; Sumner, L. W. *J. Agric. Food Chem.* **2005**, *53*, 1914–1920.

(43) Stochmal, A; Piacente, S.; Pizza, C.; De Riccardis, F.; Leitz, R.; Oleszek, W. *J. Agric. Food Chem.* **2001**, *49*, 753–758.

(44) Stochmal, A.; Simonet, A. M.; Macias, F. A.; Oleszek, W. *J. Agric. Food Chem.* **2001**, *49*, 5310–5314.

(45) Kowalska, I.; Stochmal, A.; Kapusta, I.; Janda, B.; Pizza, C.; Piacente, S.; Oleszek, W. *J. Agric. Food Chem.* **2007**, *55*, 2645–2652.

(46) Farag, M. A.; Huhman, D. V.; Lei, Z.; Sumner, L. W. *Phytochemistry* **2007**, *68*, 342–354.

(47) Jasiński, M.; Kachlicki, P.; Rodziewicz, P.; Figlerowicz, M.; Stobiecki, M. *Plant Physiol. Biochem.* **2009**, *9*, 847–853.

(48) Staszków, A.; Swarcewicz, B.; Banasiak, J.; Muth, D.; Jasiński, M.; Stobiecki, M. *Metabolomics **2011***, *7*, 604–613.

(49) Papadopoulou, K.; Melton, R. E.; Leggett, M.; Daniels, M. J.; Osbourn, A. E. *P. Natl. Acad. Sci. USA* **1999**, *96*, 12923–12928.

(50) Massad, T. J.; Trumbore, S. E.; Ganbat, G.; Reichelt, M.; Unsicker, S.; Boeckler, A.; Gleixner, G.; Gershenzon, J.; Ruehlow, S. *New Phytol.* **2014**, *203*, 607–619.

(51) Rao A, V.; Gurfinkel, D. M. *Drug Metab. Pers. Ther*. **2000**, *17*, 211–236.

(52) Sparg, S. G.; Light, M. E.; van Staden, J. *J. Ethnopharmacol.* **2004**, *94*, 219–243.

(53) Kovacsa, H.; Moskaua, D.; Spraulb, M. *Prog. Nucl. Mag. Res. Sp.* **2005**, *46*, 131–155.

(54) Dalisay, D. S.; Molinski, T. F. *J. Nat. Prod.* **2009**, *72*, 739–744.

(55) Vanholme, R.; Demedts, B.; Morreel, K.; Ralph, J.; Boerjan, W. *Plant Physiol.* **2010**, *153*, 895–905.

(56) Khanbabaee, K.; van Ree, T. *Nat. Prod. Rep.* **2001**, *18*, 641–649.

(57) Morreel, K.; Dima, O.; Kim, H.; Lu, F.; Niculaes, C.; Vanholme, R.; Dauwe, R.; Goeminne, G.; Inzé, D.; Messens, E.; Ralph, J.; Boerjan, W. *Plant Physiol.* **2010**, *153*, 1464–1478.

Table 1. Putative Identifications of Ten Glycosides Selected for the Validation of PlantMAT[a]

Peak	Rt (m)	Exact mass	n_C	Aglycone	H	A	D	P	C	M	n_G	n_G'
A-46	7.6	767.1474	3	Apigenin	0	2	0	0	1	0	7	4
A-84	11.7	1251.5642	4	Zanhic acid	2	0	1	2	0	0	90	2
A-107	13.4	1087.4933	5	Medicagenic acid	0	1	1	2	0	0	30	2
A-139	17.4	941.5095	10	Soyasapogenol B	1	1	1	0	0	0	12	1
A-145	18.1	1027.5080	15	Soyasapogenol B	1	1	1	0	0	1	60	2
R-34	9.7	515.1156	5	Formononetin	1	0	0	0	0	1	2	1
R-44	12.1	1059.4980	5	Bayogenin	3	0	0	0	0	1	10	2
R-58	13.5	911.4264	4	Medicagenic acid	2	0	0	0	0	1	6	3
R-71	14.7	927.4935	5	Hederagenin	2	0	0	1	0	0	6	3
R-103	18.8	939.4963	13	Soyasapogenol E	1	1	1	0	0	0	12	4

[a]The table gives the following results for each PlantMAT-identified peak: the number of all the possible aglycone/glycosyl/acyl combinations (n_C) from combinatorial enumeration; the most probable combination suggested by in silico MS/MS annotation (H–hexose, A–uronic acid, D–6-deoxyhexose, P–pentose, C–coumaric acid, M–malonic acid); the number of all the possible glycosyl sequences (n_G); and the number of glycosyl sequences (n_G') which have the highest matching scores.

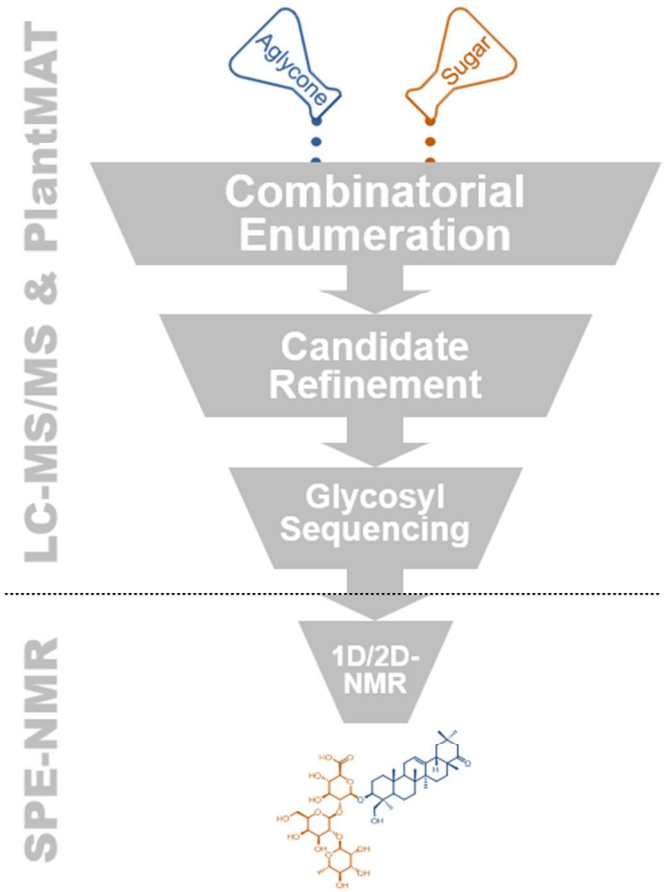
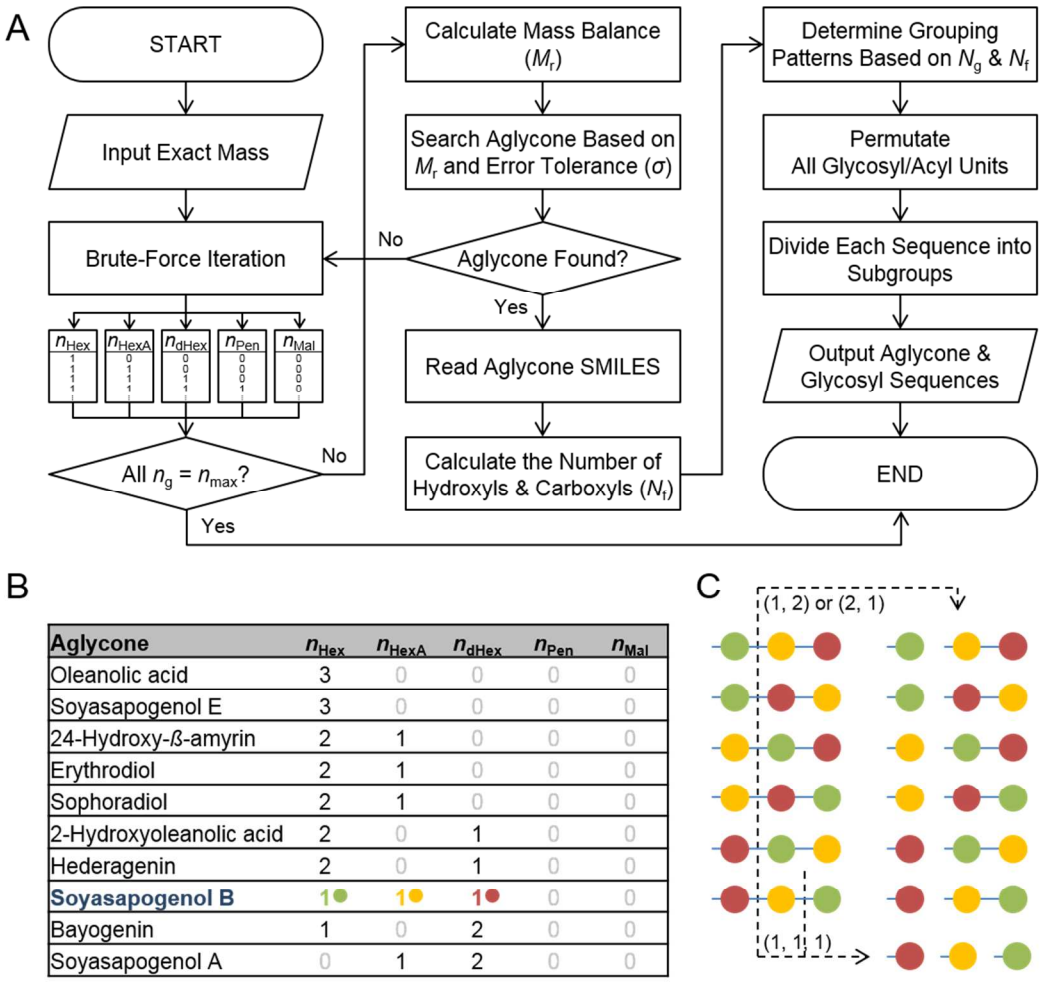


Figure 1. Integration of LC-MS/MS, PlantMAT, and SPE-NMR for accelerated metabolite identification. This first step of PlantMAT involves combinatorial enumeration of metabolite building blocks including various aglycone, glycosyl and acyl groups. Then, all possible aglycone/glycosyl/acyl combinations are refined by comparing the experimentally measured MS/MS fragments to the predicted fragment ions resulting from the combinatorial neutral losses of glycosyl and acyl moieties. The third step involves the scoring and ranking of all possible glycosyl sequences based on the hypothetical fragmentation of glycosides. Finally, the compounds are isolated using UHLPC-MS-SPE and their stereochemistry and glycosidic linkages within the structure are elucidated using 1D/2D-NMR.

[single-column figure]



Aglycone	n_{Hex}	n_{HexA}	n_{dHex}	n_{Pen}	n_{Mal}
Oleanolic acid	3	0	0	0	0
Soyasapogenol E	3	0	0	0	0
24-Hydroxy-ß-amyrin	2	1	0	0	0
Erythrodiol	2	1	0	0	0
Sophoradiol	2	1	0	0	0
2-Hydroxyoleanolic acid	2	0	1	0	0
Hederagenin	2	0	1	0	0
Soyasapogenol B	1	1	1	0	0
Bayogenin	1	0	2	0	0
Soyasapogenol A	0	1	2	0	0

Figure 2. Panel A shows the algorithmic procedure of combinatorial enumeration which involves the calculation of aglycone/glycosyl/acyl compositions and the generation of all possible glycosyl sequences. In Panel B, a saponin with exact mass of 942.5188 was predicted to have ten combinatorial possibilities of different aglycones and glycosyl units. Structure diversity of glycosides is also represented by the variety of glycosyl moieties. As shown in Panel C, permutation of three monosaccharides, Hex, HexA, and dHex, generates six sequential isomers related a trisaccharide. Two of the three monosaccharides can form a disaccharide, resulting in additional six sequential possibilities.

[two-column figure]

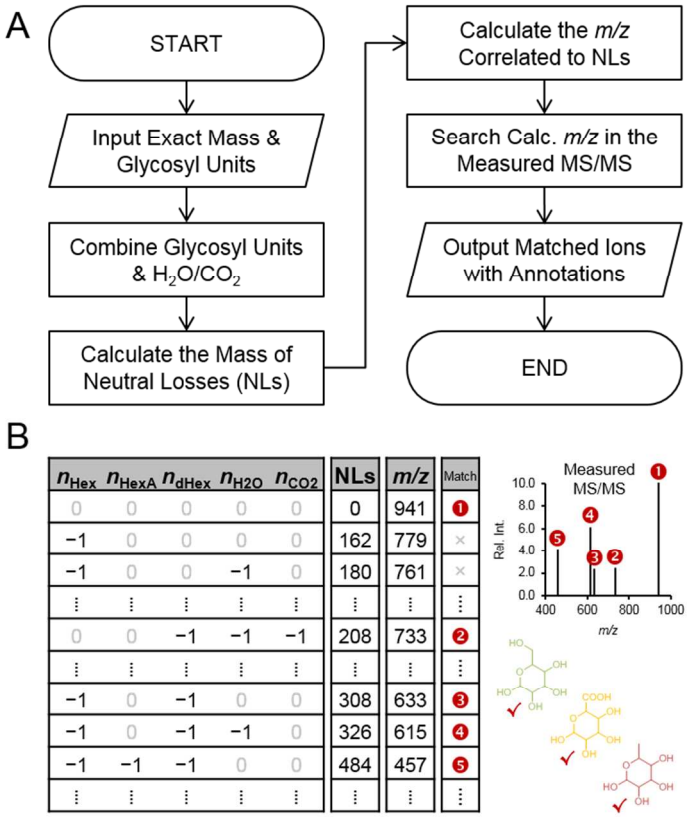


Figure 3. Panel A shows the algorithmic procedure for MS/MS spectral annotation. In Panel B, hypothetical MS/MS neutral losses from the reference saponin are initially generated through the combination of three monosaccharides as well as CO_2 and H_2O. The *m/z* values of fragment ions correlated to these neutral losses are then calculated and searched in the measured MS/MS spectrum. As a result, the four matched ions (denoted by ❷–❺) adequately confirmed the presence of Hex, HexA, and dHex.

[single-column figure]

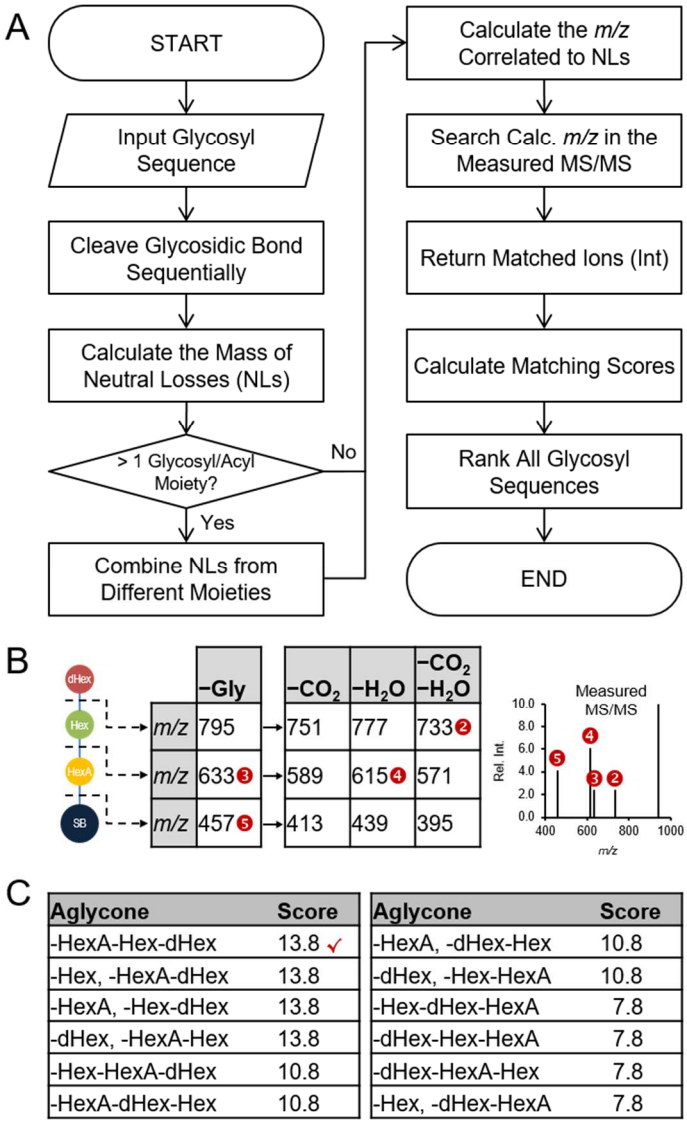


Figure 4. Panel A shows the algorithmic procedure for glycosyl sequencing. In Panel B, hypothetical fragment ions of the test saponin are initially generated by sequentially cleaving the glycosidic bonds within the trisaccharide moiety, followed by additional losses of CO_2 and H_2O. The predicted ions are then compared with the measured MS/MS spectrum and a matching score is calculated using Eq. 5. Panel C shows the twelve glycosyl sequential isomers which are ranked in the descending order of their matching scores.

[single-column figure]

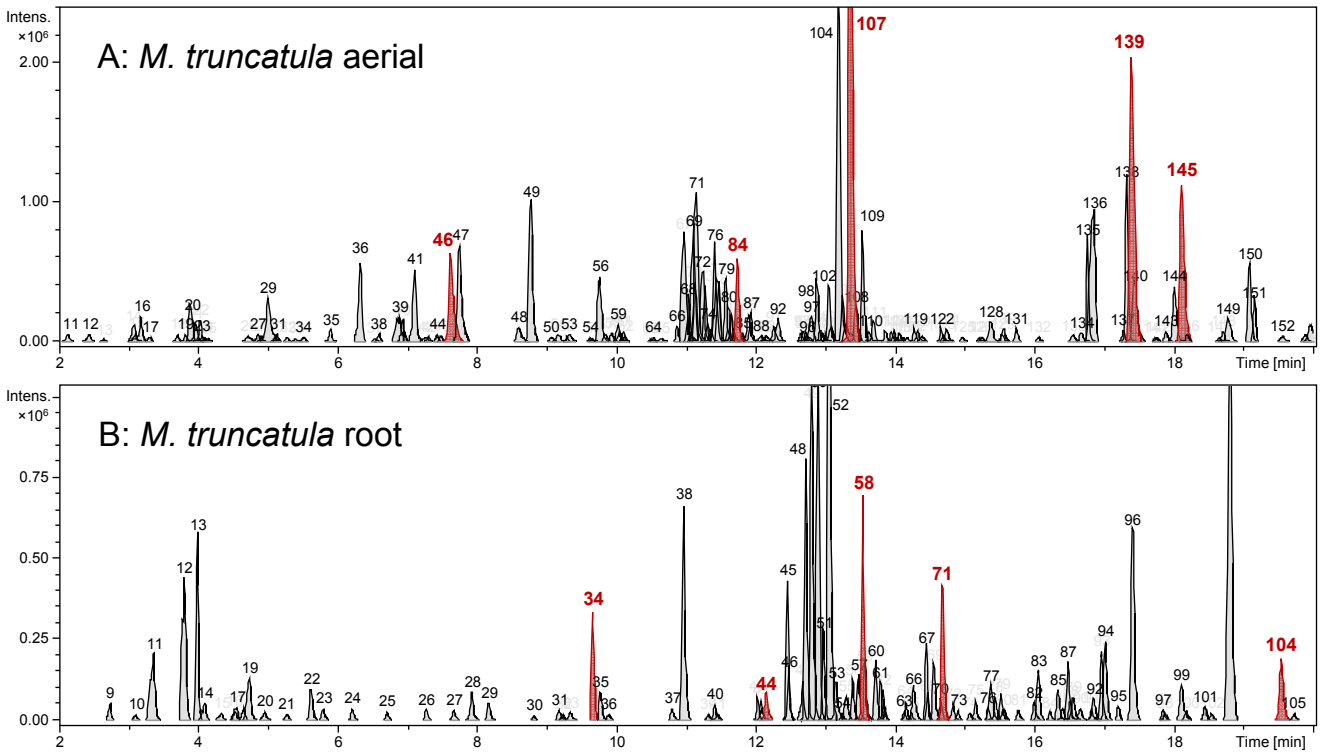


Figure 5. Deconvoluted total ion chromatograms (TICs) of *M. truncatula* aerial (A) and root (B) extracts in the range of retention time 2–20 min. Peaks shown in red were selected for the validation of PlantMAT. These peaks were trapped and concentrated via UHPLC-MS-SPE and their structures as given in Figure 6 were elucidated by a combination of 1D and 2D-NMR.

[two-column figure]

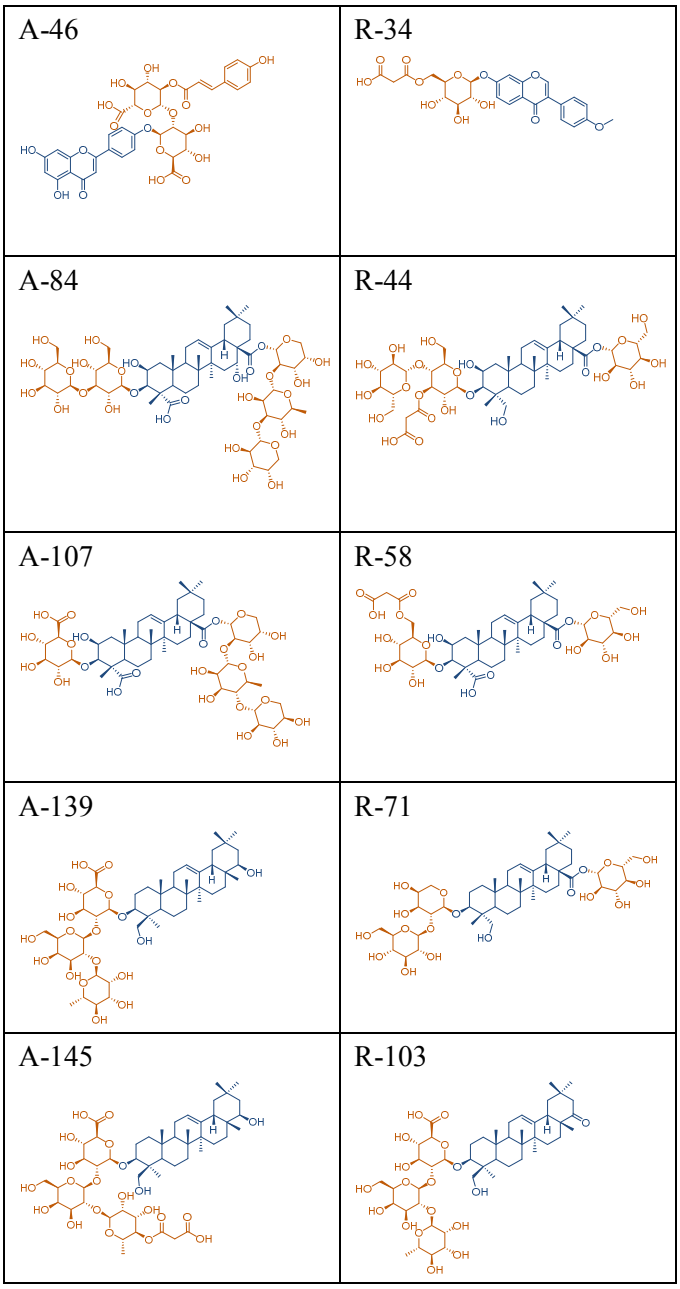


Figure 6. NMR-identified structures of ten glycosides selected for the validation of PlantMAT.

[single-column figure]

Graphical Abstract