

# UC Davis

## UC Davis Previously Published Works

### Title

Using Accurate Mass Gas Chromatography–Mass Spectrometry with the MINE Database for Epimetabolite Annotation

### Permalink

<https://escholarship.org/uc/item/0kc4g41b>

### Journal

Analytical Chemistry, 89(19)

### ISSN

0003-2700

### Authors

Lai, Zijuan  
Kind, Tobias  
Fiehn, Oliver

### Publication Date

2017-10-03

### DOI

10.1021/acs.analchem.7b01134

Peer reviewed



Published in final edited form as:

*Anal Chem.* 2017 October 03; 89(19): 10171–10180. doi:10.1021/acs.analchem.7b01134.

## Using Accurate Mass Gas Chromatography–Mass Spectrometry with the MINE Database for Epimetabolite Annotation

Zijuan Lai<sup>†</sup>, Tobias Kind<sup>†</sup>, Oliver Fiehn<sup>\*†‡</sup>

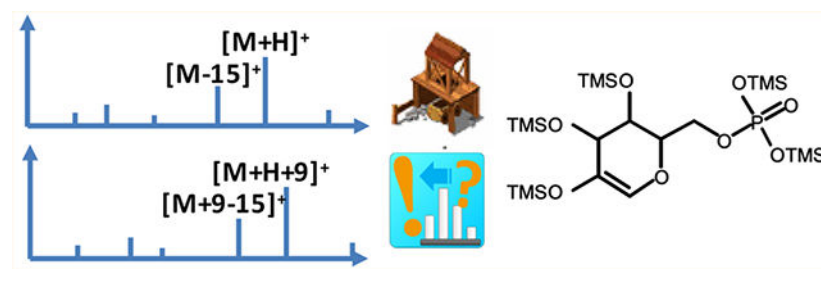
<sup>†</sup>West Coast Metabolomics Center, UC Davis, Davis, California 95616, United States

<sup>‡</sup>Department of Biochemistry, King Abdulaziz University, Jeddah 21589, Saudi Arabia

### Abstract

Mass spectrometry-based untargeted metabolomics often detects statistically significant metabolites that cannot be readily identified. Without defined chemical structure, interpretation of the biochemical relevance is not feasible. Epimetabolites are produced from canonical metabolites by defined enzymatic reactions and may represent a large fraction of the structurally unidentified metabolome. We here present a systematic workflow for annotating unknown epimetabolites using high resolution gas chromatography–accurate mass spectrometry with multiple ionization techniques and stable isotope labeled derivatization methods. We first determine elemental formulas, which are then used to query the “metabolic in-silico expansion” database (MINE DB) to obtain possible molecular structures that are predicted by enzyme promiscuity from canonical pathways. Accurate mass fragmentation rules are combined with in silico spectra prediction programs CFM-ID and MS-FINDER to derive the best candidates. We validated the workflow by correctly identifying 10 methylated nucleosides and 6 methylated amino acids. We then employed this strategy to annotate eight unknown compounds from cancer studies and other biological systems.

### Graphical Abstract



\*Corresponding Author: Cell: (+1) 530-754-8258. ofiehn@ucdavis.edu.

Author Contributions

Z.L. and O.F. designed the research. Z.L. performed experimental work, data acquisition, and data processing. T.K. instructed Z.L. for cheminformatics tools. Z.L., T.K., and O.F. interpreted the data and wrote the manuscript.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b01134.

The authors declare no competing financial interest.

For compound identification in metabolomics, experimentally obtained mass spectra are usually matched against reference data from mass spectral libraries.<sup>1,2</sup> Annotation confidence is improved if accurate masses, isotope abundance ratios, and retention index computations are added to combined scores.<sup>31</sup> While the expansion of mass spectral libraries has facilitated compound annotations in metabolomic studies, there are still many compounds that remain unidentified due to the variability of molecular structures and the scarcity of authentic standards.<sup>3,4</sup> Therefore, library-independent annotation tools need to be developed and validated.

Unidentified signals in metabolomics can have many origins.<sup>5</sup> Most importantly, the number of mass spectra of authentic compounds in current libraries is far smaller than the total number of natural products.<sup>6</sup> The PubChem repository comprises chemical structures of over 68 million small molecules,<sup>7</sup> while the NIST14 library, one of the largest mass spectral databases, only contains 435585 mass spectra.<sup>8</sup> Second, many enzymes show substrate promiscuity and perform more than one chemical reaction.<sup>9</sup> Third, not all enzymatic transformations have yet been fully discovered.<sup>10</sup> For example, many enzymes might perform repair metabolism functions to manage damaged metabolites.<sup>11</sup> Fourthly, enzymes have evolved to perform simple modifications of classic metabolites which are then removed from canonical pathways and obtain regulatory functions, such as oxylipins,<sup>12</sup> methylated and acetylated metabolites.<sup>13</sup> Such compounds have been termed “epimetabolites”.<sup>14</sup>

Gas chromatography–mass spectrometry (GC-MS) is a mature technique that offers a wide range of chemical classes to be screened simultaneously.<sup>15</sup> As we show here, it can also be used to discover new epimetabolites. Unlike collision-induced mass spectra in liquid chromatography–tandem mass spectrometry (LC-MS/MS), (hard) electron ionization mass spectra in GC-MS show a great number of fragment ions but low abundance or absence of molecular ions.<sup>16</sup> If (softer) chemical ionization is used in combination with accurate mass determination in GC-MS, molecular adduct ions can be obtained to calculate elemental compositions for the intact molecules.<sup>17</sup> These formulas can then be queried against structural databases of natural products or enzymatically transformed metabolites.

Here, we present a novel workflow for the structural annotation of unknown epimetabolites. Annotation is different to identification by the unavailability of chemical reference compounds.<sup>18</sup> Instead, the likelihood of structural annotation must be probed by prediction of mass spectra and retention times using *in silico* algorithms. The workflow largely extends our previous method<sup>17</sup> in three different ways: (a) we demonstrate that a greatly enlarged but biochemically possible suite of structures can be obtained from the MINE<sup>19</sup> database, a collection of virtual compounds that are predicted based on generalized enzymatic transformations as applied to KEGG<sup>20</sup> pathway small molecules; (b) we show that the use of derivatization methods, specifically ethoximation and trimethylsilylation-*de*, is necessary to enable filtering false positive isomer structures from such enlarged structure hit lists; and (c) for the first time, we show a combined use of the mass spectral prediction and annotation program CFM-ID<sup>21</sup> (that uses machine learning models trained by NIST<sup>8</sup> and Metlin<sup>22</sup> reference data sets) in addition to the MS-FINDER software<sup>23</sup> (that utilizes hydrogen rearrangement rules as validated by spectra from MassBank<sup>24</sup>). We first validate this novel, combined workflow using a set of commercially available epimetabolites, here, methylated

nucleosides and methylated amino acids. Subsequently, we employ the full workflow for structural annotation of unknown spectra in accurate mass GC-MS to a range of epimetabolites discovered in different biological studies.

## EXPERIMENTAL SECTION

### Reagents and Chemicals.

The following reagents and chemicals were obtained: water, isopropanol, and acetonitrile (FisherScientific, Pittsburgh PA); pyridine (Acros Organics, Geel, Belgium); C8–C30 fatty acid methyl esters [FAMES], methoxyamine hydrochloride [MeOX], ethoxyamine hydrochloride [EtOX], *N*-methyl-*N*-(trimethylsilyl)-trifluoroacetamide [MSTFA], and *N*-methyl-*N*-(trimethyl-*d*<sub>3</sub>-silyl)-trifluoroacetamide [MSTFA-*d*<sub>3</sub>] (Sigma-Aldrich, St Louis, MO). Reference compounds 1-methyladenosine, 1-methylguanosine, 1-methylpseudouridine, 2'-*O*-methylcytidine, 3'-*O*-methylcytidine, 3'-*O*-methylinosine, 3'-*O*-methyluridine, 5-methylcytidine, 5-methyluridine, and *N*-methyladenosine were purchased from Carbosynth (San Diego, CA); *N*-methylleucine, *N*-methyllysine, *N*-methylphenylalanine, *N*-methylserine, *N*-methylthreonine, and *N*-methyltyrosine were purchased from Bachem (Torrance, CA).

### Sample Preparation.

Samples were kept on ice during extraction procedures. Quantities used for sample extractions were 25  $\mu\text{L}$  for blood plasma,  $5 \times 10^6$  cells (e.g., algae cultures), or 5 mg fresh weight for tissues. Biological samples were extracted with 1000  $\mu\text{L}$  of degassed acetonitrile/isopropanol/water (3:3:2, v/v/v), and then homogenized, centrifuged, decanted, and evaporated. Extracts were cleaned by 500  $\mu\text{L}$  of degassed acetonitrile/water (1:1, v/v) to remove triglycerides and membrane lipids and evaporated again. The dried samples were derivatized with 10  $\mu\text{L}$  of MeOX (or EtOX) as 20 mg/mL solution in pyridine and subsequently by 90  $\mu\text{L}$  of MSTFA (or MSTFA-*d*<sub>3</sub>) for trimethylsilylation of acidic protons. Internal standards C8–C30 FAMES were added to determine the retention index. Samples were transferred to vials and submitted to instrumental analysis.

### Analytical Condition.

For accurate mass GC-MS analysis, we used an Agilent 7890A GC system with 7200 accurate mass Q-TOF mass spectrometer (Agilent Technologies, Santa Clara, CA, U.S.A.), maintaining the transfer line temperature at 290 °C. Chromatography was performed on a Rtx-5Sil MS column (30 m  $\times$  0.25 mm, 0.25  $\mu\text{m}$ ; Restek Corporation, Bellefonte, PA, U.S.A.) with helium (99.999%; Airgas, Radnor, PA, U.S.A.) at a constant flow of 1 mL/min. The GC temperature program was set as follows: initial temperature of 60 °C with a hold time of 1 min, a temperature ramp of 10 °C/min to 325 °C, and a final hold time of 9.5 min at 325 °C. Injection volume was 1  $\mu\text{L}$  in splitless mode at 250 °C. Mass spectra were acquired from *m/z* 50 to *m/z* 800 at 5 Hz scan rate and 750 V detector voltage in both electron ionization (EI) mode and chemical ionization (CI) mode. Other data acquisition parameters were EI ion source temperature, 230 °C; EI electron energy, 70 eV; CI ion source temperature, 300 °C; CI electron energy, 135 eV; CI gas flow rate, 20%; CI gas, methane (99.999%; Airgas, Radnor, PA, U.S.A.).

### Determination of Molecular Mass.

Data for derivatized samples were acquired by accurate mass GC-QTOF in both electron ionization (EI) mode and chemical ionization (CI) mode. Molecular masses for unknown compounds were deduced by aligning molecular adduct ions including  $[M - \text{CH}_3]^+$ ,  $[M + \text{H}]^+$ ,  $[M + \text{C}_2\text{H}_5]^+$ , and  $[M + \text{C}_3\text{H}_5]^+$ . The number and presence of carbonyl groups (ketones or aldehydes) in the derivatized molecules were computed by spectra comparison MeOX and EtOX derivatizations. The number of acidic protons were calculated by comparison of MSTFA versus MSTFA-*d*<sub>9</sub> derivatized spectra.

### Determination of Molecular Formula.

MS-FINDER software was downloaded from the PRIME web site (<http://prime.psc.riken.jp/>) and utilized to derive the molecular formulas. For each compound, the *m/z* intensity list was imported to MS-FINDER. Molecular adduct ions were determined along with the number of TMS and MeOX derivatization groups. In MS-FINDER software, the mass tolerance was set to 0.005 Da, the relative ion abundance cut off was set to 1% and isotopic ratio errors were set to 10%. The use of “LEWIS and SENIOR check” and “common range for element ratio check” filters was activated for determining elemental compositions. Target atoms were set to C, H, O, N, S, and P, including the option of “TMS-MeOX derivatized compound”. With valence rules and elemental ratio checks, the most likely molecular formula was yielded through MS-FINDER.

### Generation of Structure Candidates.

A list of isomeric structures was retrieved for each unknown compound by formula query in MINE DB, an open access database of computationally predicted enzyme promiscuity products. Raw structures were in silico derivatized in ChemAxon Instant JChem (<https://www.chemaxon.com/products/instant-jchem-suite/instant-jchem/>). This computational derivatization process also worked as a filter for removing unsuitable structures, especially for the number of acidic protons (TMS groups) and carbonyls (MeOX derivatization). Derivatized structures with free hydroxyl or carboxyl groups were deleted. We have reviewed accurate mass substructure rules for GC-MS spectra and validated 228 true-positive fragmentation patterns within 4 mDa mass accuracy window from 80 reports that were published over the past 50 years.<sup>25</sup> Characteristic substructures were obtained for each unknown compound by searching such accurate mass fragmentation rules. Subsequently, substructure constraints were applied in ChemAxon Instant JChem to further refine the list of structure candidates.

### Ranking of Structure Candidates.

The filtered structure candidates were interpreted by combining results from two in silico fragmentation algorithms for mass spectra prediction. For MS-FINDER<sup>23</sup> data processing, precursor ions were set as “[M]<sup>+</sup>” or “[M - CH<sub>3</sub>]<sup>+</sup>” with 10 mDa mass tolerance and a “fragmentation tree depth” of two fragmentations. For CFM-ID,<sup>21</sup> the web application was performed in <http://cfmid.wishartlab.com/>. We selected the “EI” spectra type, downloaded the top-20 results at 10 mDa mass tolerance, and used the dot-product scoring function. The

top-5 structure candidates from each program were manually investigated to propose the best hit.

## RESULTS AND DISCUSSION

### Overview of the Workflow for Annotation of Epimetabolites.

Metabolomic studies yield many more unidentified signals than peaks that are identified by mass spectral libraries. We largely extended our earlier workflow<sup>17</sup> for structural annotation of unknowns detected in untargeted GC-MS based metabolomics. We now employ a strategy consisting of labeled derivatization reagents, use of a much-enlarged database of biochemically possible metabolites (epimetabolites), and structure hit ranking by novel in-silico mass spectral prediction software. The complete workflow is shown in Figure 1. The first step was to select valuable targets from metabolomic chromatograms, often as result of statistical tests from biological studies. Next, we acquired high resolution GC-QTOF MS data in both electron ionization (EI) and chemical ionization (CI). The presence and identify of EI spectra was confirmed between high- and low resolution GC/MS chromatograms. Methane CI spectra were interpreted by the series of molecular adduct ions ( $[M - CH_3]^+$ ,  $[M + H]^+$ ,  $[M + C_2H_5]^+$ , and  $[M + C_3H_5]^+$ ) to identify the actual molecular mass. Each sample was further derivatized and compared for the presence of carbonyl groups (methoximation vs ethoximation) and for the number of acidic protons by comparison of trimethylsilylation (TMS) to deuterated trimethylsilylation (TMS-*d*<sub>6</sub>) spectra. Combining these data and using the MS-FINDER software led to the determination of top-ranking underivatized elemental formulas. These formulas were then used to scout for possible structure candidates in compound repositories. No hits were found in classic biochemical databases like KEGG. We here extended this search by using the Metabolic In Silico Network Expansion Database (MINE DB).<sup>19</sup> MINE is a database of potential molecular structures that could be produced by promiscuous enzymes, such as methyl-transferases and acetyl-transferases. MINE DB may therefore contain the structures of many novel epimetabolites that have not been reported previously. Such structures would be absent from regular compound repositories, including PubChem. Next, these structure lists were ranked by matching the experimental data against in-silico predicted mass spectra, utilizing the ‘competitive fragment modeling’ software CFM-ID<sup>21</sup> and MS-FINDER.<sup>23</sup> Other computational fragmentation programs, such as CSI:FingerID,<sup>26</sup> MAGMa,<sup>27</sup> and MIDAS,<sup>28</sup> do not support EI in silico fragmentation of trimethylsilylated small molecules in GC-MS. Since simple mass spectral matching did not provide sufficient differences in similarity scores, we further used accurate mass fragmentation rules with substructure assignments that we had amassed from the literature of the past 50 years.<sup>25</sup>

### Validation of the Workflow Using Authentic Epimetabolite Standards.

We tested ten methylated nucleosides and six methylated amino acids to validate this workflow (Table 1). For these standards, experimental mass spectra were not available in any public mass spectral libraries including NIST14, Metlin, MassBank or FiehnLib. Hence, these compounds could have indeed been unidentified signals in metabolomics profiling experiments. All structures were indeed predicted in the MINE<sup>19</sup> virtual metabolome database and are therefore suitable as test data sets to validate the performance of our

epimetabolites-focused discovery method. Reference mass spectra were acquired in accurate mass GC-QTOF MS (Figures S1 and S2). For each spectrum, MS-FINDER found the correct molecular formula as top hit, verifying that the mass accuracy of current high resolution instrumentation is good enough to unambiguously assign elemental compositions of unknown peaks.

For these 16 test cases, we obtained between 18 and 162 isomeric structures per elemental formula through querying the MINE database. Next, all potential candidates were subjected to *in silico* derivatization as published previously.<sup>17</sup> The number of TMS groups directly informed about the minimum number of acidic protons, providing a powerful constraint for removing impossible isomer structures. Accurate mass fragmentation rules<sup>25</sup> extracted from literature further filtered the list of candidates with characteristic substructures. From a total of 1886 derivatized structures in the 16 test data sets, overall 1298 isomers were excluded by the combined constraints of derivatization status and diagnostic substructures. The remaining candidates were then imported into mass spectra prediction software CFM-ID and MS-FINDER for further annotation. The correct structures of the 16 test epimetabolites were found at an average top-2.5 rank, a suitably small number for final verification by synthesizing authentic chemical standards. While there was no statistical difference in structure ranking between these two programs, CFM-ID yielded better performance for methylated amino acids (average rank 1.8), while MS-FINDER was superior for identifying methylated nucleosides (average rank 2.2). Mass spectral interpretations for two example cases are shown in Figure 2. For 2'-*O*-methylcytidine (Figure 2a), all intense fragment ions were annotated with high fragment scores, and the correct structure was ranked as top-hit by both programs. In contrast, for 5-methylcytidine (Figure 2b), CFM-ID and MS-FINDER annotated the true structure as second-best and fifth-best, respectively, due to the low confidence of the base ion and the absence of several major ions. These data prove that *in silico* fragmentation programs reduce the vast number of potential isomers to a very small number of candidate structures, albeit cannot rank the true structures as top 1 with 100% accuracy.

### Selecting Unidentified Signals in GC-MS-Based Metabolomics.

We then used this validated workflow for annotating eight unknown compounds that were found to be statistically significant in four different metabolomics studies (Table 2): a breast cancer tissue project<sup>29</sup> with 284 subjects, an *Escherichia coli* biochemical experiment analyzing the knockout of RidA enzyme (Reactive Intermediate Deaminase A),<sup>30</sup> an algae metabolome investigation comparing *Chlamydomonas reinhardtii* to *Chlorella minutissima* and *Euglena gracilis*,<sup>31</sup> and a plant chemotaxonomy research study.<sup>32</sup> These studies were initially conducted on a Leco Pegasus IV low resolution GC-TOF MS instrument using mass spectral deconvolution software for data processing and the BinBase database for compound identification.<sup>33</sup> The BinBase database includes NIST14 and FiehnLib mass spectral libraries, and automatically adds novel unknowns with high quality mass spectra that have not been reported previously. The eight unknown compounds were selected based on their statistical significances, statistical effect sizes, and their mass spectra purities as determined by the Leco ChromaTOF software. For example, unknown BinBase IDs 54 and 592 were found at 12.9- and 6.6-fold increases in breast tumors compared to nonmalignant breast



tissues.<sup>29</sup> Similarly, BinBase IDs 3122 and 18588 were detected with 2.4- and 3.2-fold enrichments in *Escherichia coli* RidA knockouts than in wild-types.<sup>30</sup> For algae and plant specific BinBase IDs 21695, 25801, 16833, and 17746, we found that these compounds were exclusively present in certain species but not in others.<sup>31</sup> These findings support the notion that such unknown compounds represent genuine metabolites and are not likely to be random artifacts or chemical contaminants.

### Determining Molecular Formulas with High Resolution Mass Spectrometry.

In order to determine the molecular formula of unknowns, a high resolution GC-QTOF accurate mass instrument was employed for data collection. Hence, we analyzed the samples with regular derivatization procedures in electron ionization mode to ensure that the unknown target peaks were found in high resolution GC-MS at the same retention indices as had been observed in low resolution Leco Pegasus IV GC-TOF MS profiling studies. Subsequently, we followed the workflow as discussed above using ethoximation and deuterated TMS reagents with methane-based chemical ionization. Herein, we showcased the annotation of unknown BinBase ID 54 as the example to illustrate our strategy. For BinBase ID 54, the molecular mass was calculated as 602.218 Da by aligning a series of ions at  $m/z$  587.194 ( $[M - CH_3]^+$ ),  $m/z$  603.224 ( $[M + H]^+$ ),  $m/z$  631.255 ( $[M + C_2H_5]^+$ ), and  $m/z$  643.256 ( $[M + C_3H_5]^+$ ) (Figure 3). Additionally, comparison of TMS and TMS-*d*<sub>9</sub> derivatized mass spectra showed a 45 Da mass shift, inferring this unknown had five trimethylsilyl (TMS) groups (Figure 3). Similarly, molecular masses and the numbers of TMS groups were computed for other unknowns (Table 2). The comparison of spectra under methoximation versus ethoximation proved that none of the unknown compounds comprised ketone (or aldehyde) functional groups.

Next, we utilized MS-FINDER<sup>23</sup> with valence and elemental ratio checks of the adduct and fragment ions to yield the most likely molecular formulas for the unknown compounds, adopting the Seven Golden Rules<sup>34</sup> algorithm. For BinBase ID 54, the elemental composition C<sub>6</sub>H<sub>11</sub>O<sub>8</sub>P was derived as top-hit among 32 potential formulas, using the molecular mass  $602.218 \pm 0.005$  Da and 5 TMS groups as input. The second-best formula had no hits in structure database and therefore was excluded. In an analogous manner, formulas were confirmed as C<sub>6</sub>H<sub>9</sub>NO<sub>4</sub> for BinBase ID 592, C<sub>9</sub>H<sub>17</sub>NO<sub>4</sub>S for ID 3122, C<sub>8</sub>H<sub>16</sub>N<sub>2</sub>O<sub>4</sub>S for ID 18588, C<sub>5</sub>H<sub>11</sub>NO<sub>4</sub> for ID 21695, C<sub>5</sub>H<sub>11</sub>NO<sub>3</sub> for ID 25801, C<sub>6</sub>H<sub>10</sub>O<sub>6</sub> for ID 16833, and C<sub>11</sub>H<sub>12</sub>O<sub>6</sub> for ID 17746 (Table 3).

### Annotating Molecular Structures in CFM-ID and MS-FINDER with the MINE DB.

We then retrieved molecular structures by searching elemental formulas for the unknowns. The metabolome database HMDB<sup>35</sup> only yielded 0–6 compounds per query, while the nonmetabolome small molecule repositories PubChem<sup>7</sup> and ChemSpider<sup>36</sup> returned 20–2,463 structures per formula. However, most of these structures were xenobiotic compounds and highly unlikely to be present in our samples. Instead, we assumed that most unknown features that are detected in metabolomics experiments are genuine biochemical compounds that could be derived by substrate ambiguity, enzyme promiscuity, or intracellular chemical damage.<sup>10</sup> We therefore used the MINE<sup>19</sup> database to obtain in silico metabolites that are predicted by canonical enzyme reactions. For unknown BinBase ID 54, 80 structure



candidates with high natural product likeness scores were exported from MINE DB using formula query and imported to Instant JChem for in silico derivatization. An initial 84 derivatized structures were generated and filtered to 15 suitable candidates by the following substructure constraints (Figure 4): (1) Isomers that still showed underivatized hydroxyl or carboxyl groups were removed because the reactivity of MSTFA is known to completely derivatize such functional groups. In contrary, amines were retained as fully, partly, or nonderivatized structures.<sup>25</sup> (2) According to the results by TMS-*d*<sub>9</sub> labeling, only structure candidates with five TMS remaining in the structure hit lists. (3) Isomers lacking a phosphate substructure were excluded because the mass spectral pattern of intense *m/z* 299.072 and *m/z* 315.102 ions clearly indicated this unknown as a phosphate compound.<sup>25</sup> Similarly, accurate mass GC-MS fragmentation rules were applied to confirm the presence of dihydroxyl group in BinBase ID 3122, 18588, 21695, and 16833, amine group in BinBase ID 592, amino acid group in BinBase ID 25801, as well as phenol group for BinBase ID 17746. Such characteristic substructures (Table 2) contributed to the significant reduction of structure candidates.

For BinBase ID 54, 15 filtered isomeric structures were submitted to CFM-ID and MS-FINDER for mass spectral interpretation. These programs complement each other by using different in silico fragmentation approaches, machine learning algorithm (CFM-ID) or chemical rule methodology (MS-FINDER). The best structure for BinBase ID 54 was annotated as 1-dehydro-1-deoxy-glucose-6-phosphate with rank of Top 1 in MS-FINDER and Top 2 in CFM-ID. The retention index of this peak was close to other known phosphorylated sugars, including glucose-6-phosphate. The predicted mass spectrum for 1-dehydro-1-deoxy-glucose-6-phosphate matched well with the experimental spectrum within 0.005 Da mass accuracy in both software. All major fragment ions in the accurate mass GC-MS spectrum of BinBase ID 54 were explained with assigned substructures (Figure 5). The ion series at *m/z* 315.102, *m/z* 299.072, *m/z* 243.064, *m/z* 227.033, and *m/z* 211.001 represented the phosphate substructure.<sup>25</sup> The mass difference of *m/z* 90.050 between fragments *m/z* 587.194, *m/z* 497.142, and *m/z* 407.092 indicated the sequential loss of TMSOH from the intact molecule, which is a characteristic fragmentation pattern for sugar-like compounds.<sup>25</sup> The MINE database suggests that a kinase might exist that phosphorylates 1-dehydro-1-deoxy-glucose,<sup>19</sup> a reference metabolite in the biochemical database KEGG as KEGG ID C00478.<sup>20</sup> However, it is also possible that the detected peak is a thermolytic cleavage product during the GC-MS injection, for example from glucose-1,6-bisphosphate.<sup>37</sup> This reaction mechanism is similar to the cleavage of UDP-*N*-acetylglucosamine that generates a dehydro-deoxy substructure.<sup>17</sup>

In the same manner, we confidently annotated other unknown BinBase IDs 592, 3122, 18588, 21695, 25801, 16833, and 17746 as 4-hydroxy-4-methylglutamate lactone (Figure S3), 3-(4'-methylthio)butylmalate amide (Figure S4), 1-methylcystathionine (Figure S5), ribosyl-1-amine (Figure S6), 2-amino-5-hydroxyvalerate (Figure S7), ribosyl-1-carboxylic acid (Figure S8), and 2-hydroxy-2-carboxylate-(2'-methyl-3',4'-dihydroxycyclohexene)-pyran (Figure S9), respectively. For each compound, the fragmentation pathways and fragment ion annotations were interpreted and cross-validated by accurate mass fragmentation rules, CFM-ID, and MS-FINDER. The d<sub>9</sub>-TMS derivatized mass spectra yielded the number of acidic protons for both molecular ions and fragment ions in an

unambiguous way (Figure S10), which served as important confirmation of the *in silico* predictions. The proposed structures for these eight unknowns were generated by known metabolites with known enzymes such as kinase, hydrolyase, oxidoreductase, or methyltransferase (Table 4), giving high probability that these compounds might actually exist. However, none of these molecules were available as authentic standards either from academic or commercial sources. Hence, one cannot name these compounds as identified but only as annotated, according to the nomenclature established by the Metabolomics Society.<sup>18</sup> Alternatively, unknown peaks might be collected by fraction collector and subjected to NMR structural elucidation.<sup>38</sup> However, this procedure is not well established for gas chromatography-based fraction collection. Given the thousands of unknowns observed in metabolomics, we here offer a systematic route that could be suitable for large scale analyses.

## CONCLUSIONS

We here present a novel workflow for *in silico* structural annotation of unknowns by using accurate mass GC-QTOF MS with multiple ionization methods and derivatization approaches. Such annotations could be very useful for discovering novel epimetabolites.<sup>14</sup> This workflow was validated by correctly identifying 10 methylated nucleosides and 6 methylated amino acids that were absent from public mass spectral libraries. We selected eight unknown compounds from four GC-MS based metabolomics studies that were first determined by elemental formulas and then annotated for best molecular structures with mass spectral interpretations from lists of isomeric candidates. A true challenge in untargeted metabolomics is to annotate novel molecules that have never been reported in literature and that are not covered in chemical or biochemical structure databases.<sup>39</sup> We hence utilized the MINE database to retrieve possible structures of epimetabolites that might exist naturally. We also show that a workflow benefits from combining two different *in-silico* fragmentation software packages as no single software is sufficient today. We expect this pipeline to be the basis for *in silico* identification of new epimetabolites in future metabolomics studies, and foresee that further unknown discoveries will allow the sphere expansion of untargeted metabolomics.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGMENTS

We appreciate funding by the U.S. National Science Foundation Projects MCB 113944 and MCB 1611846, as well as the U.S. National Institutes of Health Grants U24 DK097154 and S10 RR031630.

## REFERENCES

- (1). Dunn WB; Ellis DI *TrAC, Trends Anal. Chem* 2005, 24, 285–294.
- (2). Fiehn O *Plant Mol. Biol* 2002, 48, 155–171. [PubMed: 11860207]
- (3). Patti GJ; Yanes O; Siuzdak G *Nat. Rev. Mol. Cell Biol* 2012, 13, 263–269. [PubMed: 22436749]
- (4). Vinaixa M; Schymanski EL; Neumann S; Navarro M; Salek RM; Yanes O *TrAC, Trends Anal. Chem* 2016, 78, 23–35.

- (5). da Silva RR; Dorrestein PC; Quinn RA Proc. Natl. Acad. Sci. U. S. A 2015, 112, 12549–12550. [PubMed: 26430243]
- (6). Fiehn O TrAC, Trends Anal. Chem 2008, 27, 261–269.
- (7). Kim S; Thiessen PA; Bolton EE; Chen J; Fu G; Gindulyte A; Han L; He J; He S; Shoemaker BA; et al. Nucleic Acids Res 2016, 44, D1202–13. [PubMed: 26400175]
- (8). Wallace WE; Ji W; Tchekhovskoi DV; Phinney KW; Stein SE J. Am. Soc. Mass Spectrom 2017, 28, 733–738. [PubMed: 28127680]
- (9). Tawfik OK; S D Annu. Rev. Biochem 2010, 79, 471–505. [PubMed: 20235827]
- (10). Khersonsky O; Roodveldt C; Tawfik DS Curr. Opin. Chem. Biol 2006, 10, 498–508. [PubMed: 16939713]
- (11). Linster CL; Van Schaftingen E; Hanson AD Nat. Chem. Biol 2013, 9, 72–80. [PubMed: 23334546]
- (12). Howe GA; Schillmiller AL Curr. Opin. Plant Biol 2002, 5, 230–236. [PubMed: 11960741]
- (13). Su X; Wellen KE; Rabinowitz JD Curr. Opin. Chem. Biol 2016, 30, 52–60. [PubMed: 26629854]
- (14). Showalter MR; Cajka T; Fiehn O Curr. Opin. Chem. Biol 2017, 36, 70–76. [PubMed: 28213207]
- (15). Fiehn O; Kopka J; Dörmann P; Altmann T; Trethewey RN; Willmitzer L Nat. Biotechnol 2000, 18, 1157–1161. [PubMed: 11062433]
- (16). McLafferty FW; Turek F Interpretation of Mass Spectra; University Science Books, 1993.
- (17). Kumari S; Stevens D; Kind T; Denkert C; Fiehn O Anal. Chem 2011, 83, 5895–5902. [PubMed: 21678983]
- (18). Sumner LW; Amberg A; Barrett D; Beale MH; Beger R; Daykin CA; Fan TW-M; Fiehn O; Goodacre R; Griffin JL; et al. Metabolomics 2007, 3, 211–221. [PubMed: 24039616]
- (19). Jeffries JG; Colastani RL; Elbadawi-Sidhu M; Kind T; Niehaus TD; Broadbelt LJ; Hanson AD; Fiehn O; Tyo KE; Henry CS J. Cheminf 2015, 7, 44.
- (20). Kanehisa M; Goto S Nucleic acids research 2000, 28, 27–30. [PubMed: 10592173]
- (21). Allen F; Pon A; Greiner R; Wishart D Anal. Chem 2016, 88, 7689–7697. [PubMed: 27381172]
- (22). Smith CA; O’Maille G; Want EJ; Qin C; Trauger SA; Brandon TR; Custodio DE; Abagyan R; Siuzdak G Ther. Drug Monit 2005, 27, 747–751. [PubMed: 16404815]
- (23). Tsugawa H; Kind T; Nakabayashi R; Yukihiro D; Tanaka W; Cajka T; Saito K; Fiehn O; Arita M Anal. Chem 2016, 88, 7946–7958. [PubMed: 27419259]
- (24). Horai H; Arita M; Kanaya S; Nihei Y; Ikeda T; Suwa K; Ojima Y; Tanaka K; Tanaka S; Aoshima K; et al. J. Mass Spectrom 2010, 45, 703–714. [PubMed: 20623627]
- (25). Lai Z; Fiehn O Mass Spectrom. Rev 2016, na.
- (26). Dührkop K; Shen H; Meusel M; Rousu J; Böcker S Proc. Natl. Acad. Sci. U. S. A 2015, 112, 12580–12585. [PubMed: 26392543]
- (27). Ridder L; van der Hoof JJ; Verhoeven S Mass Spectrom 2014, 3, S0033–S0033.
- (28). Wang Y; Kora G; Bowen BP; Pan C Anal. Chem 2014, 86, 9496–9503. [PubMed: 25157598]
- (29). Denkert C; Bucher E; Hilvo M; Salek R; Orešič M; Griffin J; Brockmüller S; Klauschen F; Loibl S; Barupal DK; et al. Genome Med 2012, 4, 37. [PubMed: 22546809]
- (30). Niehaus TD; Gerdes S; Hodge-Hanson K; Zhukov A; Cooper AJ; ElBadawi-Sidhu M; Fiehn O; Downs DM; Hanson AD BMC Genomics 2015, 16, 382. [PubMed: 25975565]
- (31). Tsugawa H; Cajka T; Kind T; Ma Y; Higgins B; Ikeda K; Kanazawa M; VanderGheynst J; Fiehn O; Arita M Nat. Methods 2015, 12, 523–526. [PubMed: 25938372]
- (32). Tushingham S; Ardura D; Eerkens JW; Palazoglu M; Shahbaz S; Fiehn O Journal of Archaeological Science 2013, 40, 1397–1407.
- (33). Fiehn O; Wohlgemuth G; Scholz M International Workshop on Data Integration in the Life Sciences; Springer, 2005; pp 224–239.
- (34). Kind T; Fiehn O BMC Bioinf 2007, 8, 105.
- (35). Wishart DS; Tzur D; Knox C; Eisner R; Guo AC; Young N; Cheng D; Jewell K; Arndt D; Sawhney S; et al. Nucleic Acids Res 2007, 35, D521–D526. [PubMed: 17202168]
- (36). Pence HE; Williams AJ Chem. Educ 2010, 87, 1123.
- (37). Beitner R Trends Biochem. Sci 1979, 4, 228–230.

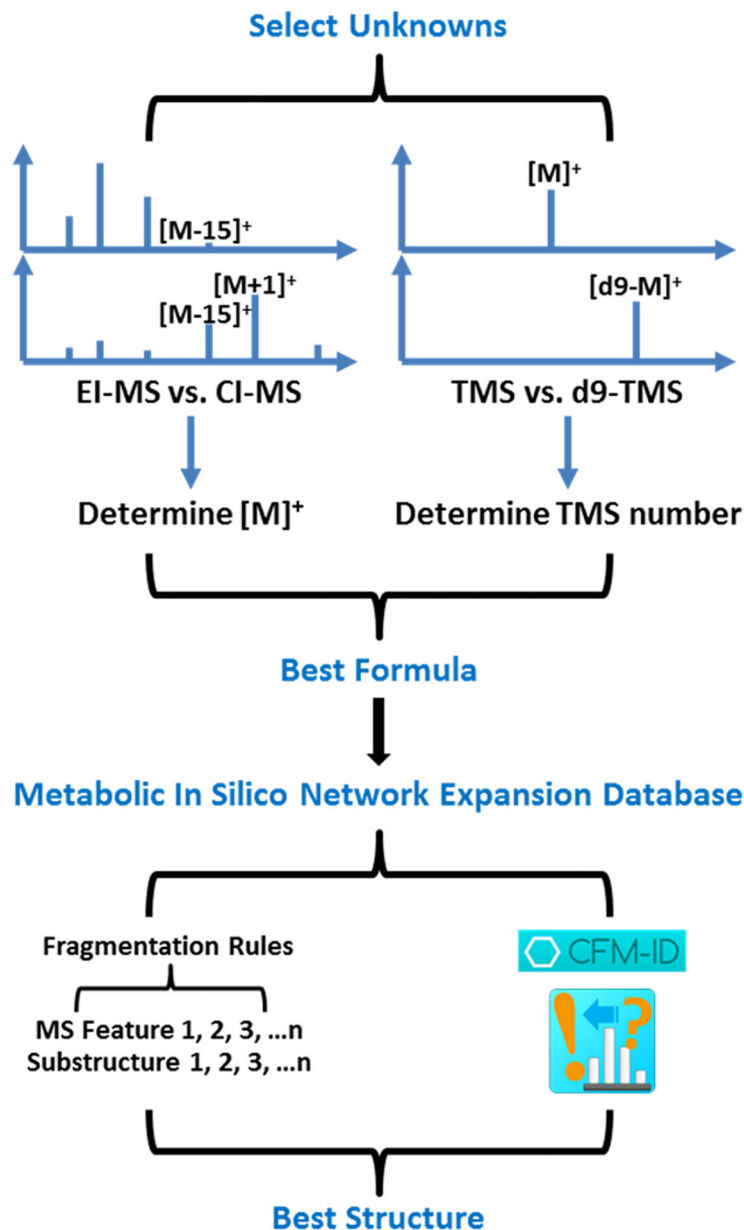
- (38). Corcoran O; Spraul M Drug Discovery Today 2003, 8, 624–631. [PubMed: 12867148]  
(39). Kind T; Fiehn O Bioanalytical reviews 2010, 2, 23–60. [PubMed: 21289855]

Author Manuscript

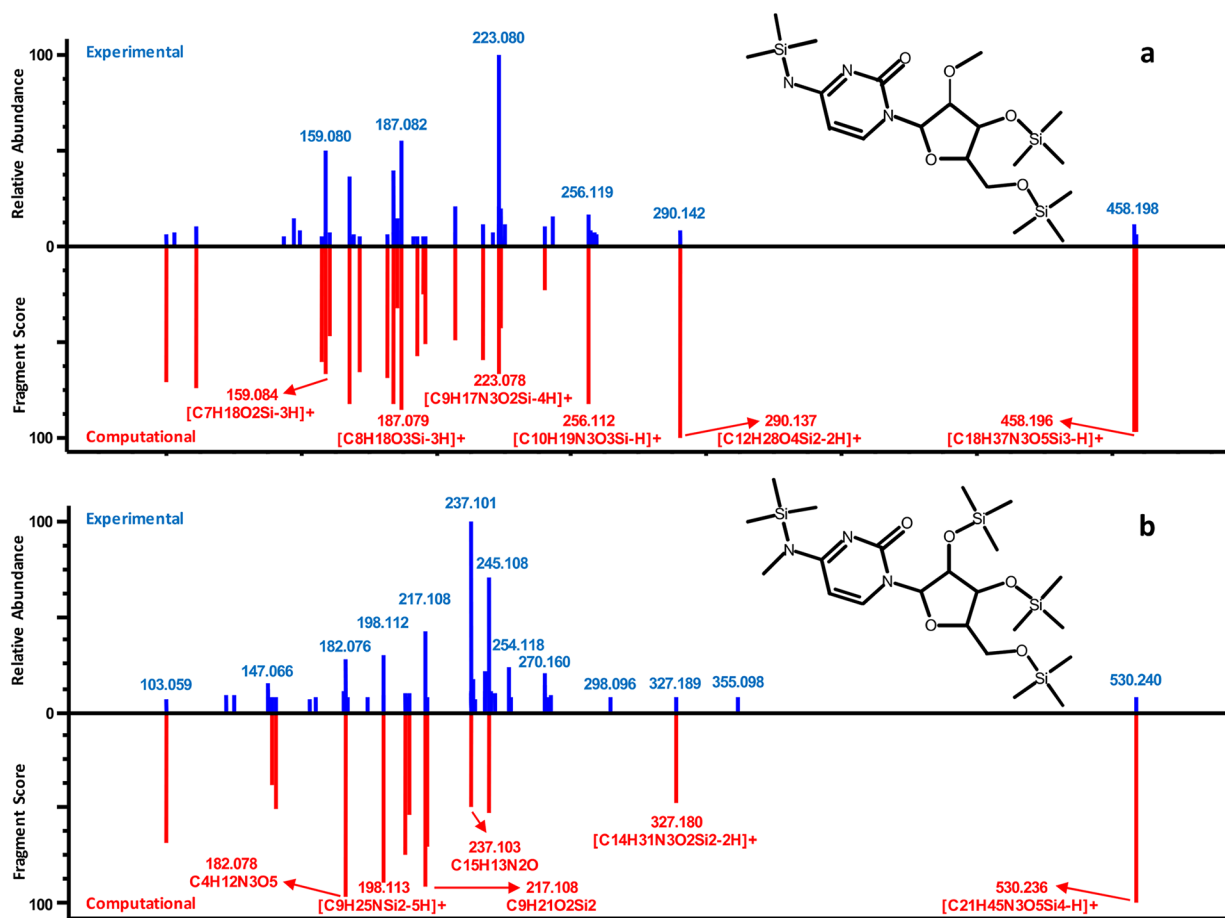
Author Manuscript

Author Manuscript

Author Manuscript

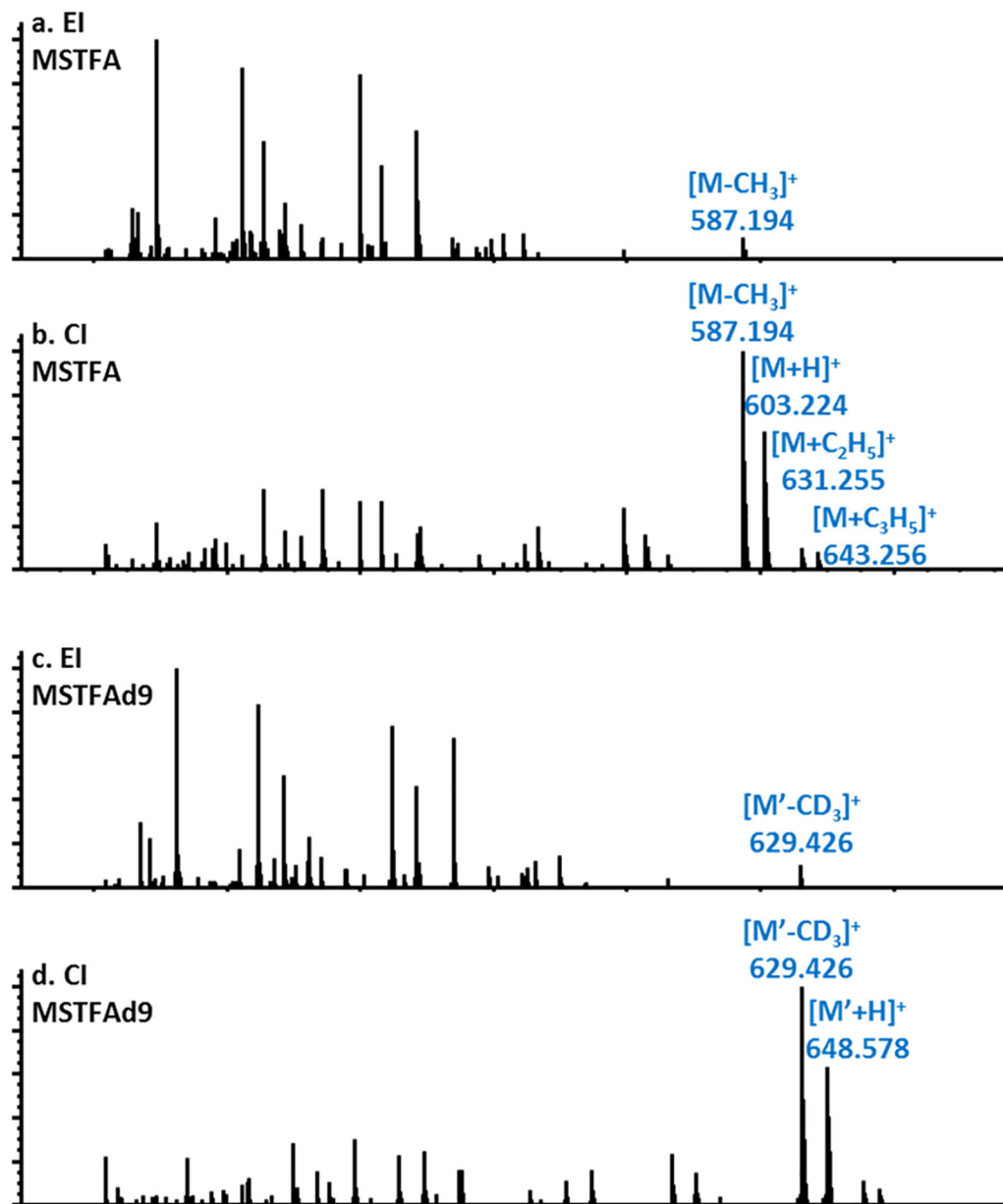


**Figure 1.** Workflow for identifying unknown peaks in the discovery of epimetabolites. High resolution accurate mass GC-MS was used to generate data from different ionization modes and derivatization methods. The molecular mass was determined by aligning molecular adduct ions in EI and CI spectra. The number of TMS groups was calculated based on the mass unit shift between TMS and TMS-*d*<sub>9</sub> spectra. For each peak, the best formula was derived by matching accurate masses, isotope abundance ratios, and TMS derivatization status. The MINE database was queried to obtain molecular structures that expand the currently known number of metabolites by assuming enzyme substrate ambiguity. Subsequently, accurate mass fragmentation rules were combined with in silico spectra simulation programs CFM-ID and MS-FINDER to annotate the best structure.



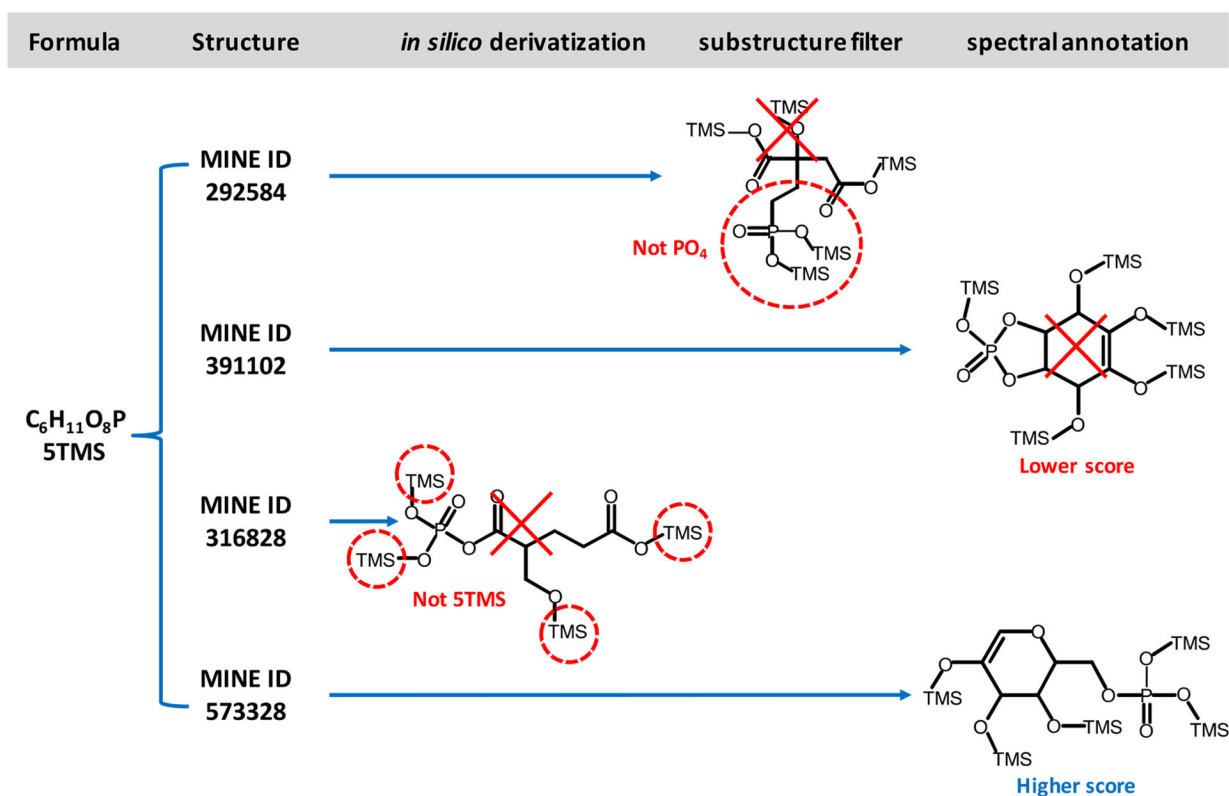
**Figure 2.**

In silico mass spectra simulation and scoring in MS-FINDER. (a) For 2'-*O*-methylcytidine, all intense fragment ions were annotated with high fragment scores, and the correct structure was ranked as top-best hit. (b) For 5-methylcytidine, the most abundant ion  $m/z$  237.101 yielded a low fragment score. Major fragment ions  $m/z$  254.118,  $m/z$  270.160,  $m/z$  298.096, and  $m/z$  355.098 were absent for interpretation, and the true structure was ranked as fifth-best hit.



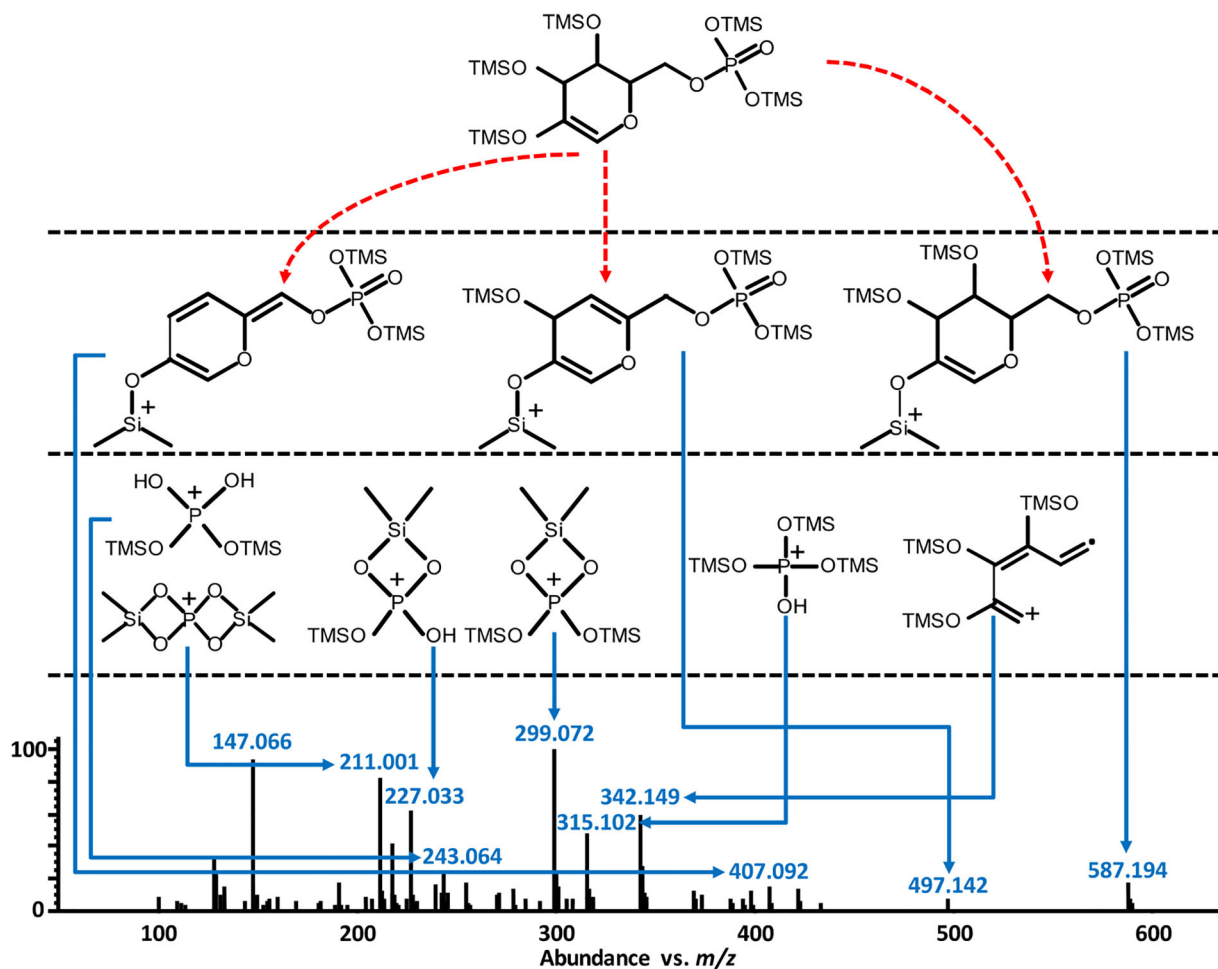
**Figure 3.** Molecular mass derivation for unknown BinBase ID 54. When investigating the mass spectra of the compound in (a) EI mode with MSTFA derivatization, (b) methane-CI mode with MSTFA derivatization, (c) EI mode with MSTFA- $d_9$  derivatization, and (d) methane-CI mode with MSTFA- $d_9$  derivatization, the molecular mass was calculated as 602.218 Da by aligning a series of ions at  $m/z$  587.194 ( $[M - CH_3]^+$ ),  $m/z$  603.224 ( $[M + H]^+$ ),  $m/z$  631.255 ( $[M + C_2H_5]^+$ ), and  $m/z$  643.256 ( $[M + C_3H_5]^+$ ). (c) EI mode using TMS- $d_9$  and (d) methane-CI mode using TMS- $d_9$ .





**Figure 4.**

Annotating molecular structures for unknown BinBase ID 54. C<sub>6</sub>H<sub>11</sub>O<sub>8</sub>P was confirmed as the best formula and used to query structures from the MINE DB. The isomeric structures were *in silico* trimethylsilylated. Candidates without five TMS or PO<sub>4</sub> substructure were excluded. Computational spectra were simulated and scored in CFM-ID and MS-FINDER. The best structure was deduced as 1-dehydro-1-deoxy-glucose-6-phosphate by software ranking and manual investigation.



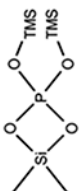


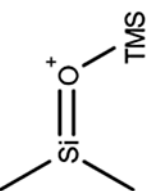

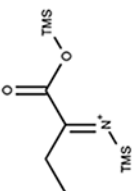

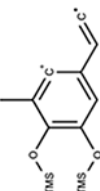
**Figure 5.** Mass spectral annotation of unknown BinBase ID 54. The in silico fragmentation pathways of 1-dehydro-1-deoxy-glucose-6-phosphate 5 TMS were interpreted with neutral loss fragments in the upper part of half of the fragment structures and rearrangement structures below. The theoretical exact masses of fragment ions with assigned substructures were matched against experimental accurate masses within 5 mDa.

**Table 1.**  
Annotation of 10 Methylated Nucleosides and 6 Methylated Amino Acids for Workflow Validation

No.	name	RT (min)	RI (Fiehn)	MINE ID	formula	mass	CFM-ID rank	MS-FINDER rank
1	1-methyladenosine	24.9	942720	602371	C <sub>11</sub> H <sub>15</sub> N <sub>5</sub> O <sub>4</sub> -3TMS	497.231	3	4
2	1-methylguanosine	28.8	1102998	259583	C <sub>11</sub> H <sub>15</sub> N <sub>5</sub> O <sub>5</sub> -4TMS	585.265	8	1
3	1-methylpseudouridine	23.1	872424	284529	C <sub>10</sub> H <sub>14</sub> N <sub>2</sub> O <sub>6</sub> -3TMS	474.204	2	2
4	2'-O-methylcytidine	24.2	916966	602133	C <sub>10</sub> H <sub>15</sub> N <sub>3</sub> O <sub>5</sub> -3TMS	473.220	1	1
5	3'-O-methylcytidine	23.9	905689	602134	C <sub>10</sub> H <sub>15</sub> N <sub>3</sub> O <sub>5</sub> -3TMS	473.220	2	2
6	3'-O-methylinosine	23.1	870695	550471	C <sub>11</sub> H <sub>14</sub> N <sub>4</sub> O <sub>5</sub> -3TMS	498.215	2	2
7	3'-O-methyluridine	22.0	828405	507523	C <sub>10</sub> H <sub>14</sub> N <sub>2</sub> O <sub>6</sub> -3TMS	474.204	2	1
8	5-methylcytidine	24.2	914212	153249	C <sub>10</sub> H <sub>15</sub> N <sub>3</sub> O <sub>5</sub> -4TMS	545.259	2	5
9	5-methyluridine	22.6	852669	521419	C <sub>10</sub> H <sub>14</sub> N <sub>2</sub> O <sub>6</sub> -4TMS	546.243	5	2
10	N-methyladenosine	23.8	901455	52386	C <sub>11</sub> H <sub>15</sub> N <sub>5</sub> O <sub>4</sub> -3TMS	497.231	1	2
11	N-methylleucine	10.8	384857	18421	C <sub>7</sub> H <sub>13</sub> NO <sub>2</sub> -2TMS	289.189	2	2
12	N-methyllysine	16.9	634380	23416	C <sub>7</sub> H <sub>16</sub> N <sub>2</sub> O <sub>2</sub> -3TMS	376.240	5	2
13	N-methylphenylalanine	15.4	570942	23901	C <sub>10</sub> H <sub>13</sub> NO <sub>2</sub> -2TMS	323.174	1	2
14	N-methylserine	12.0	430901	5876	C <sub>6</sub> H <sub>9</sub> NO <sub>3</sub> -3TMS	335.177	1	6
15	N-methylthreonine	12.5	451638	979	C <sub>5</sub> H <sub>11</sub> NO <sub>3</sub> -3TMS	349.192	1	6
16	N-methyltyrosine	18.8	704844	5520	C <sub>10</sub> H <sub>13</sub> NO <sub>3</sub> -3TMS	411.208	1	2

Table 2.

Metadata of Eight Selected Unknowns for BinBase IDs, Retention Indices, Kingdoms, Species, Molecular Masses, TMS Numbers, and Characteristic Fragment Ion with Substructure

No.	BinBase ID	Kovats RI	Fiehn RI	Kingdom	Species	Mass	TMS	characteristic experimental masses	Fragment
1	54	2229	773272	Human	Cancer Cell	602.218	5	299.072	
2	592	1432	514932	Human	Cancer Cell	303.135	2	142.105	
3	3122	1812	654564	Bacteria	Escherichia coli	523.249	4	205.108	
4	18588	2393	814481	Bacteria	Escherichia coli	524.242	4	147.066	
5	21695	1421	510542	Algae	Chlamydomonas reinhardtii	437.229	4	217.108	
6	25801	1243	438455	Algae	Chlamydomonas reinhardtii	349.194	3	246.135	
7	16833	1662	603638	Plant	Artemisia douglasiana	466.205	4	217.108	
8	17746	2114	743080	Plant	Artemisia douglasiana	528.226	4	292.135	

**Table 3.**

## Best Formula Derivation of Eight Selected Unknowns

No.	BinBase ID	best formula			
		original	derivatized	molecular mass	mass error
1	54	C <sub>6</sub> H <sub>11</sub> O <sub>8</sub> P	C <sub>21</sub> H <sub>51</sub> O <sub>8</sub> PSi <sub>5</sub>	602.217	0.001
2	592	C <sub>6</sub> H <sub>9</sub> NO <sub>4</sub>	C <sub>12</sub> H <sub>25</sub> NO <sub>4</sub> Si <sub>2</sub>	303.132	0.003
3	3122	C <sub>9</sub> H <sub>17</sub> NO <sub>4</sub> S	C <sub>18</sub> H <sub>45</sub> N <sub>5</sub> O <sub>5</sub> Si <sub>4</sub>	523.25	0.001
4	18588	C <sub>8</sub> H <sub>16</sub> N <sub>2</sub> O <sub>4</sub> S	C <sub>20</sub> H <sub>48</sub> N <sub>2</sub> O <sub>4</sub> SSi <sub>4</sub>	524.241	0.001
5	21695	C <sub>5</sub> H <sub>11</sub> NO <sub>4</sub>	C <sub>17</sub> H <sub>43</sub> NO <sub>4</sub> Si <sub>4</sub>	437.227	0.002
6	25801	C <sub>5</sub> H <sub>11</sub> NO <sub>3</sub>	C <sub>14</sub> H <sub>35</sub> NO <sub>3</sub> Si <sub>3</sub>	349.192	0.002
7	16833	C <sub>6</sub> H <sub>10</sub> O <sub>6</sub>	C <sub>18</sub> H <sub>42</sub> O <sub>6</sub> Si <sub>4</sub>	466.206	0.001
8	17746	C <sub>11</sub> H <sub>12</sub> O <sub>6</sub>	C <sub>23</sub> H <sub>44</sub> O <sub>6</sub> Si <sub>4</sub>	528.222	0.004

**Table 4.** Best Structure Hits for the Annotation of Eight Selected Unknowns Peaks from BinBase

No.	BinBase ID	No. of query hits						best structure					
		HMDB	ChemsSpider	PubChem	MINE	name	reactant	enzyme	MINE ID	CFM-ID	MS-FINDER		
1	54	1	20	88	80	1-dehydro-1-deoxy-glucose-6-phosphate	1-dehydro-1-deoxy-glucose	phosphokinase	573328	top 2	top 1		
2	592	2	398	1201	84	4-hydroxy-4-methylglutamate lactone	4-hydroxy-4-methylglutamate	hydrolyase	523142	top 1	top 2		
3	3122	0	1577	2463	8	3-(4'-m)-butylmalate amide	3-(4-methylthio) butylmalate	amide synthase	449196	top 2	top 1		
4	18588	2	752	1254	10	1-methylcystathionine	cystathionine	methyltransferase	52919	top 2	top 1		
5	21695	0	190	577	29	ribosyl-1-amine	5-phosphoribosylamine	phosphatase	53784	top 1	top 1		
6	25801	3	421	1197	60	2-amino-5-hydroxyvalerate	norvaline	oxidoreductase	1459	top 4	top 1		
7	16833	6	200	670	145	ribosyl-1-carboxylic acid	gluconic acid	hydrolyase	19760	top 1	top 1		
8	17746	0	163	643	80	2-hydroxy-2-carboxylate-(2',-methyl-3',4'-dihydroxycyclohexene)-pyran	2-hydroxy-8-methylchromene-2-carboxylate	oxidoreductase	137082	top 2	top 2		