

# Accelerated unknown compound annotation with confidence: from spectra to structure in untargeted metabolomics experiments

## Authors

Amanda Souza, Ioanna Ntai,  
and Ralf Tautenhahn,  
Thermo Fisher Scientific

## Keywords

Untargeted metabolomics,  
nontargeted metabolomics,  
discovery, metabolite profiling,  
unknown identification, compound  
annotation, qualitative analysis,  
elemental composition, database  
search, spectral library search,  
Compound Discoverer software,  
mzCloud spectral library,  
HighChem Fragmentation Library,  
high-resolution accurate-mass  
MS, Orbitrap ID-X Tribrid mass  
spectrometer, mzLogic algorithm

## Goal

Use Thermo Scientific™ Compound Discoverer™ 3.0 software to process untargeted metabolomics for unknown compound annotation. Demonstrate the utility of analyzing multiple analytical measurements from mass spectral data to apply a consensus approach across different annotation sources for confident annotation assignments.

## Introduction

Untargeted metabolomics aims to comprehensively detect endogenous metabolites to generate a metabolic profile of a given biological system. This unbiased approach is often without a priori knowledge of the molecular make-up of a sample. Untargeted metabolite profiling globally provides the potential association of a metabolite or set of metabolites to a biochemical phenotype that can be further associated with metabolic pathways and biological function. These associations are insightful in an array of research and discovery settings such as defining signatures of disease and mechanisms of cellular function in human health, defining metabolic response in plant and animals resulting from a changing environment, or evaluating food composition in a commercial setting. The application of untargeted metabolomics plays a critical role in understanding the molecular underpinnings in biological systems.

One of the greatest challenges in untargeted metabolomics analysis by mass spectrometry (MS) is the identification of unknown compounds.<sup>1-4</sup> The gold standard in the metabolomics field is to confirm the identification of an unknown analyte by comparing the generated mass spectra against that of a purified reference standard. Yet, there are numerous instances where this is not possible. It could be that the reference standard is commercially unavailable, and therefore requires chemical synthesis. Another instance may require isolation and purification from a select biological source. The most challenging scenario is when the unknown compound is truly novel and has yet to be characterized by researchers. Despite these practical setbacks, modern mass spectrometers and data processing tools provide the means for confident annotation of unknown compounds.<sup>5</sup>

Specifically, untargeted metabolomics analysis using high-resolution accurate-mass (HRAM) tandem Orbitrap™ mass spectrometry generates multiple analytical measurements that when taken collectively via consensus evaluation, builds confidence in the compound annotation or compound class association (Table 1). These multiple measurements start with the acquisition of ultra-high resolution mass spectra to distinguish between molecular ions of closely related mass and detection of associated adduct ions, isotope pattern, and, of particular importance, isotope fine structure. This information, combined with accurate mass measurements, provides confident elemental composition prediction. The elemental formulae are then used for chemical structure database searching.

Next, fragmentation spectra provide an added level of knowledge about the unknown compound. Fragmentation spectra are used to search against spectral libraries. Fragment ions can also be compared against theoretical fragments of putative structures by *in silico* fragmentation prediction. Multi-stage fragmentation (MS<sup>n</sup>) by ion trap instruments enables further structural characterization by establishing product ion relationships. Lastly, the novel mzLogic™ algorithm uses actual fragmentation spectra to prioritize putative chemical structure candidates. Collectively, this information facilitates the unknown annotation process by reducing the number of possible candidate compounds determined by the data and increases the confidence in annotation assignments.

The information-rich data generated by Thermo Scientific™ Orbitrap™ mass spectrometers must be efficiently extracted in a single data processing pipeline that streamlines the unknown compound annotation process and presents data in an intelligible format. Thermo Scientific Compound Discoverer 3.0 software expedites the annotation assignment process in untargeted metabolic profiling experiments from HRAM Orbitrap data using numerous annotation tools to fully integrate multiple analytical measurements. To start, the software predicts elemental composition using the power of HRAM Orbitrap data with isotope fine structure and incorporates fragmentation information. The elemental composition predicted from spectral data is then used to search against a chemical database for matching potential structures. The Compound Discoverer software is fully integrated with the ChemSpider™ chemical structure database providing access to over 250 data sources and

**Table 1. Multiple analytical measurements for unknown compound annotation. Data types include MS<sup>1</sup>, MS<sup>2</sup> and MS<sup>n</sup> mass spectral information along with chromatographic separation.**

| Data type             | Measured property              | Information                                                                                                                                                                                         |
|-----------------------|--------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Accurate mass         | Monoisotopic mass              | Elemental composition (mass tolerance)                                                                                                                                                              |
| Adducts               | Accurate mass difference       | Assignment of molecular species (M+H, M+NH <sub>4</sub> , 2M+H)                                                                                                                                     |
| Isotopic pattern      | Isotope distribution           | Constrain possible elemental formulas                                                                                                                                                               |
| Ultra-high resolution | Isotopic fine structure        | Separate isobaric species<br>Direct confirmation of isotopes: <sup>13</sup> C, <sup>15</sup> N, <sup>34</sup> S, <sup>18</sup> O, <sup>2</sup> H<br>Confirm elemental formula (isotopes and ratios) |
| MS <sup>2</sup>       | Product ions<br>Neutral losses | Sub-structures and their elemental composition<br>Compound class (signature neutral loss or product ions)                                                                                           |
| MS <sup>n</sup>       | Product ion relationships      | Mass spectral tree (precursor ion fingerprinting)<br>Collision energy profile (breakdown curves)                                                                                                    |
| LC-MS                 | Polarity<br>Hydrophobicity     | Retention time order—differentiate isomers<br>Compared to literature/reference standards                                                                                                            |

68 million chemical structures representing a vast chemical space of endogenous and exogenous origin (<http://www.chemspider.com/>). Furthermore, corresponding MS<sup>2</sup> fragmentation spectra are searched against the mzCloud<sup>™</sup> spectral library, which hosts an extensive array of HRAM MS<sup>n</sup> fragmentation spectra (<https://www.mzcloud.org/>).

For compounds lacking direct spectral library matches, the mzLogic algorithm takes advantage of actual experimental fragmentation data to prioritize candidate compounds resulting from the ChemSpider database search. The mzLogic algorithm utilizes sub-structure annotations from the mzCloud library mapped back to the experimental fragment ions to rank the most probable chemical structures in the ChemSpider database results list. Using actual data to rank chemical structures allows users to focus on reasonable compound candidates. Additionally, the Fragment Ion Search (FISh) scoring algorithm incorporates *in silico* fragmentation of a proposed chemical structure to explain fragment ion structures based on literature-defined chemical reactions using the HighChem<sup>™</sup> Fragmentation Library<sup>™</sup>.

Building cumulative evidence across different annotation sources for each individual compound enables consensus evaluation moving toward greater confidence, which in turn gives rise to greater certainty in the subsequent metabolic pathway analyses. If a purified standard is indeed available, the Compound Discoverer software considers chromatographic retention times and spectral matching using a customizable, in-house library to confirm compound identification. It should be noted that while the scope of this application note is to demonstrate unknown compound annotation, the Compound Discoverer software is a complete data analysis program for untargeted metabolomics experiments that operates by first applying a data reduction strategy to generate meaningful compounds that are experimentally related and subsequently providing statistical capabilities geared for differential analysis, visualization tools, and pathway analysis.<sup>6,7</sup>

Here we demonstrate the use of the Compound Discoverer 3.0 software for confident annotation of unknown compounds using the aforementioned annotation tools with an untargeted metabolomics analysis. The NIST Standard Reference Material (SRM) 1950, Metabolites in Frozen Human Plasma, was

analyzed with a novel, automated acquisition strategy to generate more fragmentation spectra using the Thermo Scientific<sup>™</sup> Orbitrap ID-X<sup>™</sup> Tribrid<sup>™</sup> mass spectrometer coupled to a Thermo Scientific<sup>™</sup> Vanquish<sup>™</sup> Horizon Ultra High Pressure Liquid Chromatograph (UHPLC) system.

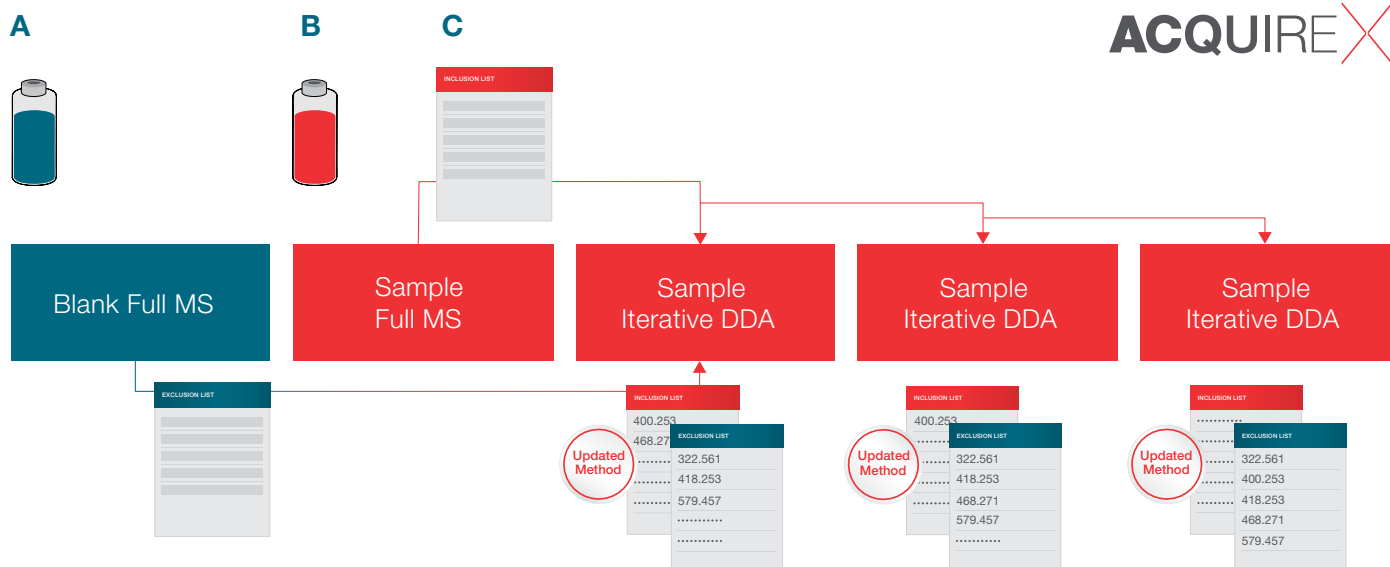
## Experimental conditions

### Sample preparation

SRM 1950 was purchased from the National Institute of Standards and Technology (NIST). The plasma sample was prepared via protein precipitation with the addition of four volumes of 80% methanol. The sample was centrifuged, and the supernatant collected. The sample extract was evaporated to dryness, then reconstituted in water containing 0.1% formic acid, and subsequently transferred to a deactivated autosampler vial. The solvent blank was prepared using the reconstitution solution with direct transfer to a deactivated autosampler vial. A total of 2 µL was injected onto the stationary phase.

### Instrument and method setup

The sample was analyzed using the Orbitrap ID-X Tribrid mass spectrometer (Table 2) coupled to a Vanquish Horizon UHPLC system. Data were acquired using Thermo Scientific<sup>™</sup> Xcalibur<sup>™</sup> 4.2 software and Thermo Scientific<sup>™</sup> Standard Integration Software (SII) for Xcalibur 1.4. The chromatographic separation was obtained with a Thermo Scientific<sup>™</sup> Hypersil GOLD<sup>™</sup> column (1.9 µm, 150 × 2.1 mm). The column was eluted isocratically at a flow rate of 300 µL/min with 100% mobile phase A (0.1% formic acid in water) for 3 min followed by a linear gradient to 50% mobile phase B (0.1% formic acid in methanol) over 8 min, and then to 98% mobile phase B over 1 min. The column compartment temperature was held at 45 °C. Sample analysis was performed using the AcquireX Deep Scan setup (Figure 1), an automated, data-informed acquisition strategy for real-time determination of ion inclusion and exclusion with repeated sample interrogation. In brief, full scan MS was acquired first on the solvent blank and next on the plasma extract to respectively generate an ion exclusion and inclusion list. Data-dependent acquisition with the automatically generated lists was performed on the plasma extract. The plasma extract was repeatedly injected with subsequent exclusion and inclusion list updates to generate the greatest number of unique fragmentation spectra for sample-related precursor ions. Detailed instrument parameters are provided in Tables 3 and 4.



**Figure 1. AcquireX Deep Scan acquisition strategy.** A) This approach employs data driven intelligence by first injecting a solvent blank (or experimental blank) to generate a three-dimensional matrix of precursor ions ( $m/z$ , retention time, and intensity) assigned as background ions that are unrelated to the experimental sample. Background ions are automatically added to an ion exclusion list for use in the data-dependent method. B) Next, the experimental sample is injected to generate a three-dimensional matrix whereby valid chromatographic peaks are detected and assigned as candidate ions for fragmentation by automatic addition to the ion inclusion list. Data generated from these two injections are then automatically incorporated into the subsequent data-dependent acquisition method, which is then applied to the sample. C) Upon completion of the first data dependent  $MS^2$  run, the acquisition process automatically moves ions already selected for fragmentation from the inclusion list to the exclusion list. The acquisition method is updated and subsequently applied to a repeat injection of sample. This iterative acquisition reduces redundant sampling of reoccurring precursor ions from injection to injection thereby generating more fragmentation spectra of unique compounds.

**Table 2. Mass spectrometer conditions for data acquisition using full scan mode and AcquireX Deep Scan setup, an intelligent data-dependent acquisition strategy using multiple injections.**

| MS conditions                 |                                           |
|-------------------------------|-------------------------------------------|
| Ion Source                    | Thermo Scientific™ OptaMax™ NG ion source |
| Ionization                    | ESI Positive Mode                         |
| Qualitative Acquisition       | AcquireX Deep Scan                        |
| Source conditions             |                                           |
| Sheath Gas Flow Rate          | 40 Arbitrary Units (AU)                   |
| Auxiliary Gas Flow Rate       | 8 AU                                      |
| Sweep Gas Flow Rate           | 1 AU                                      |
| Spray Voltage                 | 3500 V                                    |
| Ion Transfer Tube Temperature | 275 °C                                    |
| RF Lens                       | 35%                                       |
| Vaporizer Temperature         | 320 °C                                    |

**Table 3. Instrument method parameters for high-resolution full scan mode of the solvent blank and the NIST SRM 1950 human plasma to generate  $MS^1$  quantitation.**

| Full Scan acquisition parameters |                          |
|----------------------------------|--------------------------|
| Scan Range ( $m/z$ )             | 67–1000                  |
| Orbitrap Resolution              | 120,000 FWHM @ 200 $m/z$ |
| AGC Target                       | 1e <sup>5</sup>          |
| Max Injection Time               | 50 msec                  |

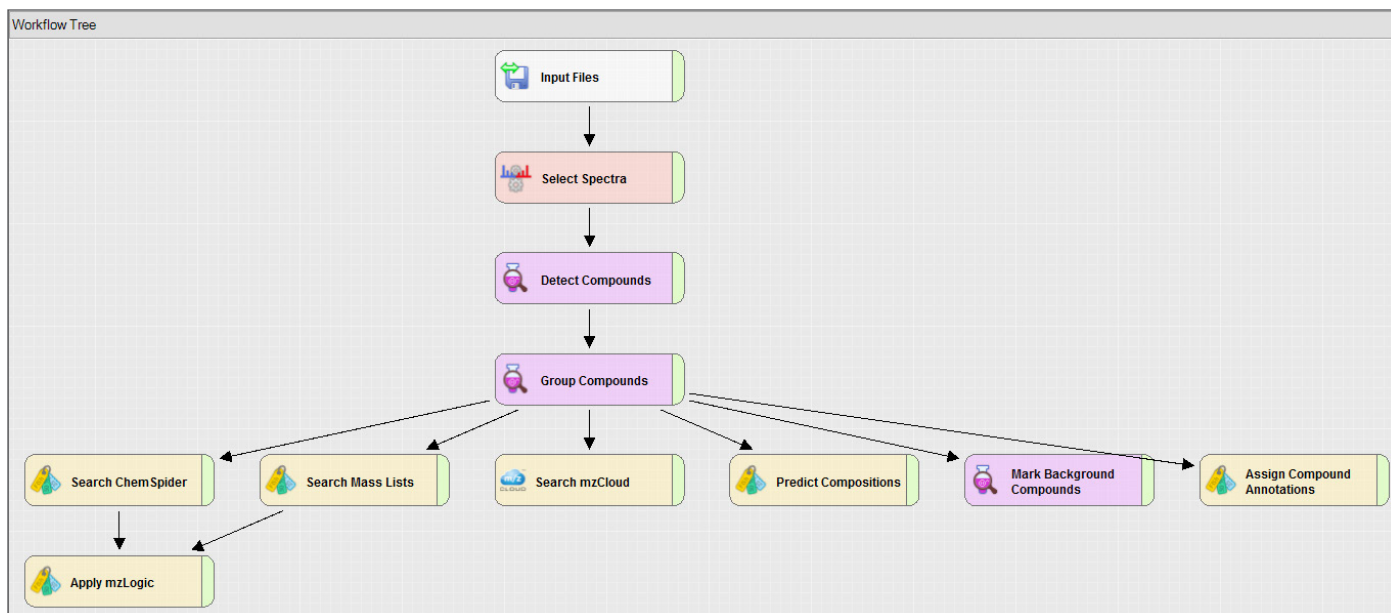
**Table 4. Instrument method parameters for the high-resolution data-dependent MS<sup>2</sup> fragmentation of the NIST SRM 1950 human plasma using the AcquireX Deep Scan setup for iterative injections and real-time dynamic updating of precursor ion exclusion and inclusion lists.** Data generated with this method exclusively provides the fragmentation spectra used for compound annotation in the data analysis process.

| Data-dependent Scan Mode Acquisition Parameters |                               |
|-------------------------------------------------|-------------------------------|
| Full scan MS                                    |                               |
| Scan Range ( <i>m/z</i> )                       | 67–1000                       |
| Orbitrap Resolution                             | 120,000 FWHM @ 200 <i>m/z</i> |
| AGC Target                                      | 1e <sup>5</sup>               |
| Max Injection Time                              | 50 msec                       |
| Data-dependent MS <sup>2</sup> Fragmentation    |                               |
| Top-speed Mode                                  | 0.6 sec cycle time            |
| Activation                                      | HCD                           |
| dd-MS <sup>2</sup> Resolution                   | 30,000 FWHM @ 200 <i>m/z</i>  |
| AGC Target                                      | 5e <sup>4</sup>               |
| Max Injection Time                              | 54 msec                       |
| Quad Isolation Width                            | 1.5 Daltons                   |
| Normalized Stepped Collision Energy             | 20, 35, 50%                   |
| Intensity Threshold                             | 2e <sup>4</sup>               |
| Dynamic Exclusion (s)                           | 2.5 sec                       |

## Data processing

Data were processed using the Compound Discoverer 3.0 software. To expedite the data processing setup, a pre-defined processing template was used. A modified version of the Metabolomics Max ID workflow template was employed to provide exhaustive compound annotation from multiple measurements following unknown peak detection. This template is useful for a limited number of raw data files where the peak intensity threshold is set to a pre-determined minimum low value, and detection filters are turned off to accommodate very low abundant compounds. Briefly, the workflow includes unbiased unknown compound detection, elemental composition prediction, database searching at the precursor level against the ChemSpider database, application of the mzLogic algorithm to rank putative candidates generated from the ChemSpider database, searching against a custom-built metabolomics database, as well as MS<sup>2</sup> spectral matching against the mzCloud spectral library. The Compound Discoverer 3.0 software is fully integrated with the ChemSpider database and the mzCloud spectral library for automated and expedited data processing. Ten data sources were selected via the ChemSpider database comprising both endogenous and exogenous entries: Aggregated Computational Toxicology Resource (ACToR), BioCyc, Drug Bank, EAWAG Biocatalysis/Biodegradation Database, Environmental Protection Agency (EPA) DSSTox, EPA ToxCast, Federal Drug Administration Unique Ingredient Identifier (UNII), FooDB, Human Metabolome Database, and the Kyoto Encyclopedia of Genes and Genomes™ (KEGG). Data were processed with and without mzLogic for comparison purposes. A more detailed illustration of the workflow and associated nodes is depicted in Figure 2.





**Figure 2. A workflow tree from the Compound Discoverer 3.0 software displaying select data processing nodes and the associated workflow connections.** Included are preliminary data processing nodes like Input Files and Select Spectra nodes. Unknown peak detection is implemented via the Detect Compounds node. Information across multiple raw data files are integrated through the Group Compounds node. Numerous unknown compound annotation nodes are utilized for both MS<sup>1</sup> and MS<sup>2</sup> spectral data: Predict Compositions node, Search Mass List node, Search ChemSpider node, Apply mzLogic node, and the Search mzCloud node. The Assign Compound Annotations node prioritizes the annotation source. Lastly, the Mark Background Compounds node incorporates a solvent blank (experimental blank) to indicate compounds arising from the sample preparation.

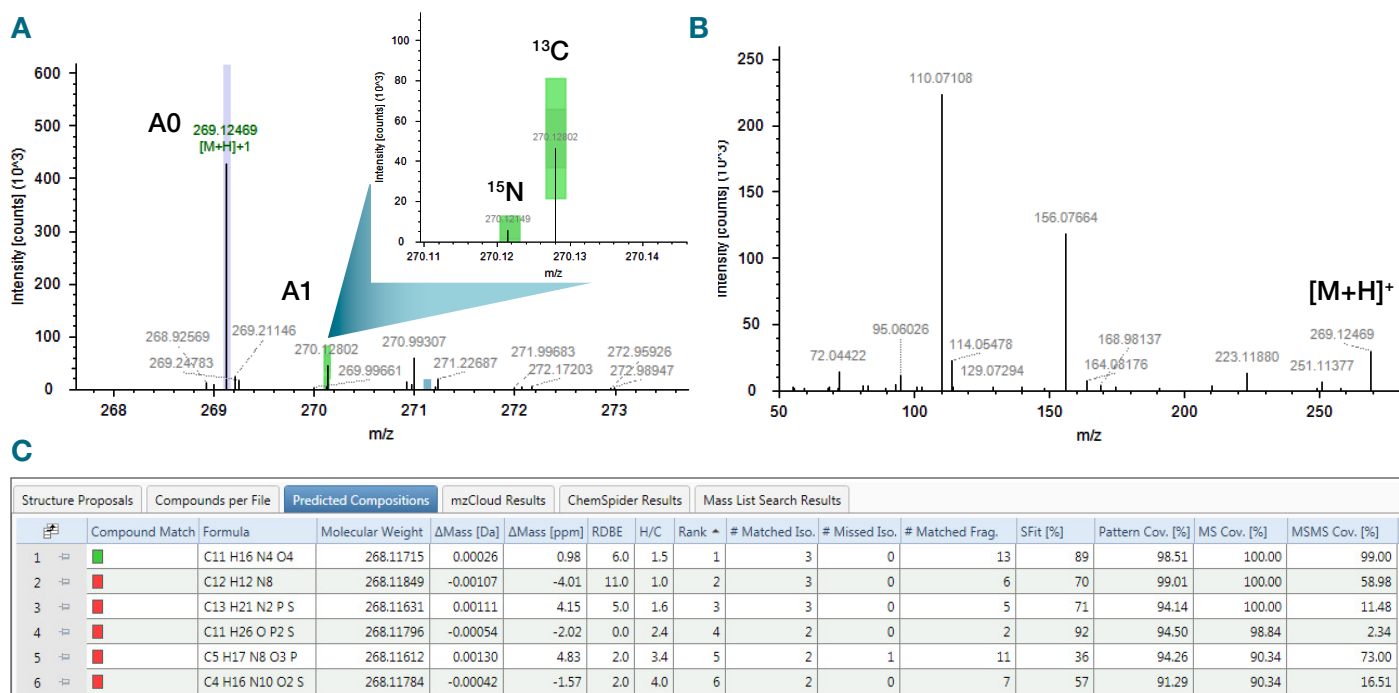
## Results and discussion

### The role of HRAM in unknown compound annotation

High-resolution mass spectrometers increase the ability to distinguish between two molecules of closely related mass, which is highly advantageous for untargeted metabolomics analyses of complex matrices such as human plasma extracts. Additionally, information obtained from accurate mass lends the ability to apply narrow search criteria against theoretical values of exact mass and sufficiently high resolution allowing for the detection of isotope pattern and isotopic fine structure. HRAM obtained with Orbitrap mass spectrometers and isotopic fine structure increases confidence in elemental composition prediction, which is further aided with fragmentation data. Performing database searches using elemental formula is preferred over molecular weight or  $m/z$  due to increased specificity, which in turn reduces complexity. Taken together, these attributes are beneficial for the annotation of unknown compounds.

Figure 3A shows a narrow mass range of the full scan spectrum displayed in the Compound Discoverer 3.0 software indicating the monoisotopic peak and respective isotope pattern for the expected compound at

$m/z$  269.1247. Of particular interest is the isotope fine structure achieved with high resolution to differentiate the nitrogen 15 isotope from the carbon 13 isotope. Mass and intensity tolerances are applied. In the example, the measurement reveals the presence of the nitrogen atom, which is applied to the elemental composition prediction. This confidently eliminates chemical formulas lacking nitrogen, thus, reducing the number of candidate formulas. The Compound Discoverer software additionally considers fragment ions to rank putative chemical formulas by confirming fragment ions that match a subset of the precursor ion's elemental composition. Figure 3B shows the MS<sup>2</sup> spectrum of the isolated precursor ion using an isolation window of 1.5 Daltons (Da). The presence of the [M+H]<sup>+</sup> ion confirms the precursor ion isolation. Figure 3C displays the predicted composition results for the expected compound, acetylcarnosine. Supporting data is provided for each possible candidate formula including delta mass, the number of matched isotopes, the number of matched fragment ions, spectral fit (Sfit) and coverage values. The elemental formula predicted for  $m/z$  269.1247 is C<sub>11</sub>H<sub>16</sub>N<sub>4</sub>O<sub>4</sub>. High resolution, accurate mass, and fragmentation data collectively contribute increased confidence for the predicted chemical formula.



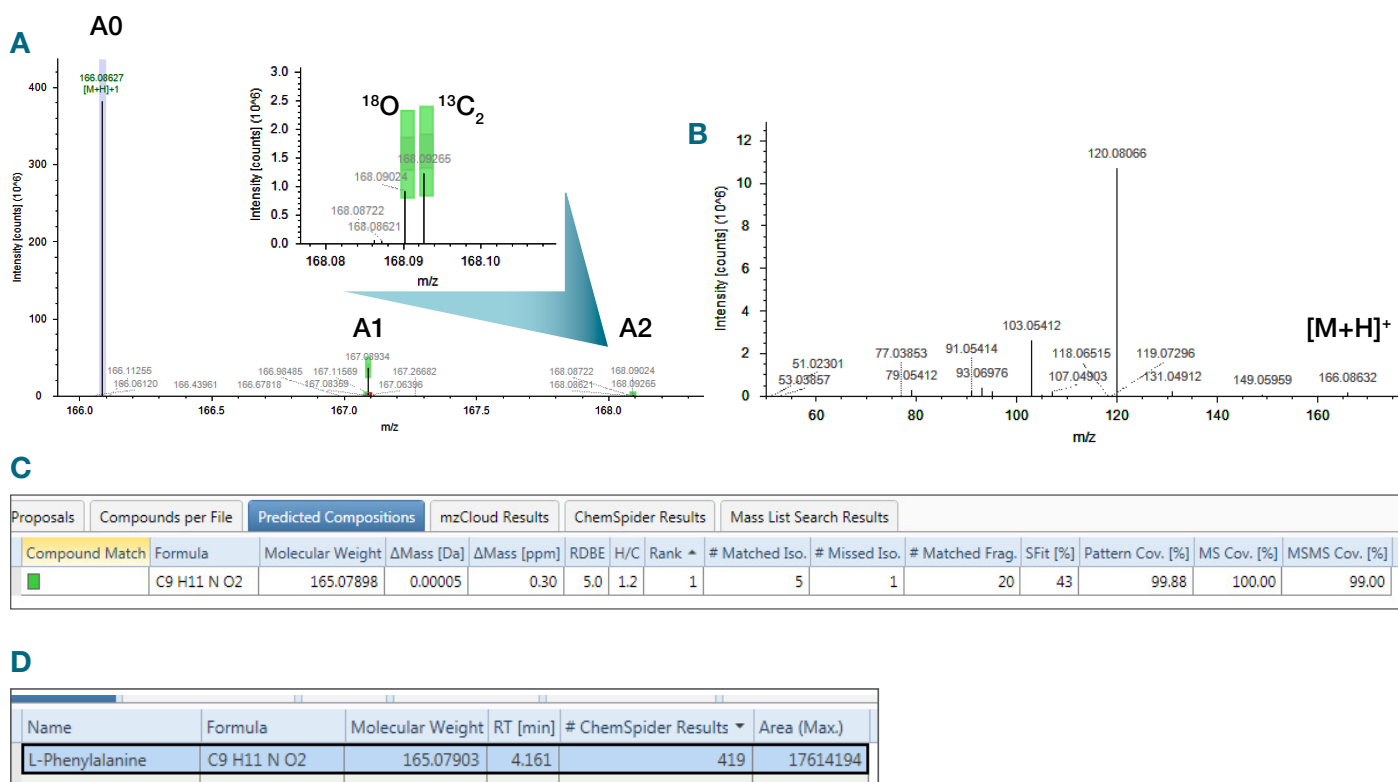
**Figure 3. HRAM and fragmentation data to predict elemental composition.** A) MS<sup>1</sup> survey scan showing color-coded isotope pattern fit for the detected compound of  $m/z$  269.1247. The violet vertical box indicates the monoisotopic protonated molecule of the expected compound, acetylcarnosine, matching the centroid. Green boxes represent the theoretical isotope pattern with defined tolerances for mass and relative intensity. Inset is the region of the A1 isotope cluster displaying the <sup>13</sup>C and <sup>15</sup>N isotopes. B) MS<sup>2</sup> fragment ion spectrum for the isolated precursor ion of  $m/z$  269.1247. C) Predicted composition results table for the selected compound displaying multiple candidates in rank order. Several variables are considered including delta mass, the number of matched isotopes, and the number of matched fragment ions.

## Database search as a starting point

Chemical structure databases aid in generating putative candidate compounds. Public repositories containing known, well-defined chemical compounds can be searched against for annotation. Information found in these databases generally include chemical structure, related chemical characteristics, metadata like biological activity, associations, and may include fragmentation spectra either from actual data or *in silico* prediction of fragmentation. The Compound Discoverer software is fully integrated with the ChemSpider database for automatic searching specifically for precursor mass information. Two search modes can be utilized. The searches can be performed either by elemental formula or by mass. The advantage of searching by elemental formula over mass is that the elemental composition prediction incorporates several variables from the mass measurement, as described in the previous section for more confident formulas. A search by formula can reduce the number of possible matches when searching databases compared to mass alone. Nevertheless, database searching provides a good starting point for compound annotation via precursor matching.

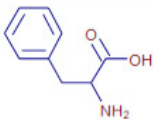
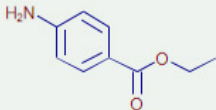
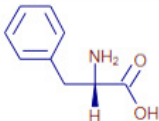
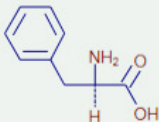
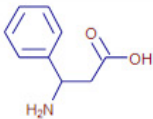
Figure 4A shows a narrow mass range from the full scan spectrum in the Compound Discoverer 3.0 software for the expected compound phenylalanine. The molecular ion and associated isotopes for the amino acid are detected. Elemental composition is predicted using both full scan data and associated fragment ions (Figure 4B) resulting in the formula C<sub>9</sub>H<sub>11</sub>NO<sub>2</sub> (Figure 4C). Applying this formula to the ChemSpider database search against the selected data sources generates 419 possible candidate matches (Figure 4D). DL-phenylalanine was ranked highest based on the number of references in the ChemSpider database, which is over 13,000 (Figure 5). The ChemSpider results table in the Compound Discoverer 3.0 software displays an interactive ChemSpider ID for easy network access in addition to chemical structure, molecular formula, and delta mass. While a database search finds putative candidate compounds, caution should be used when this is the sole annotation source because the search is simply based on precursor information only. As shown here, over 400 possible annotations were generated for this single chemical formula. This results from the fact that one molecular formula can represent various possible chemical structures, such as structural isomers where there is the same number of atoms for each element but with different spatial arrangement.

Searching data sources containing endogenous and exogenous substances can be highly useful in an untargeted metabolomics experiment, particularly when analyzing samples obtained from “free-ranging” organisms such as humans and animals. The SRM 1950 human plasma sample was pooled from a collection of 100 fasted donors, both male and female.<sup>8</sup> A putative ChemSpider match for benzoylecgonine, a metabolite of the drug compound cocaine, was detected in the sample (Figure 6); illustrating the value of comprehensive small molecule analysis considering more than just endogenous metabolites.

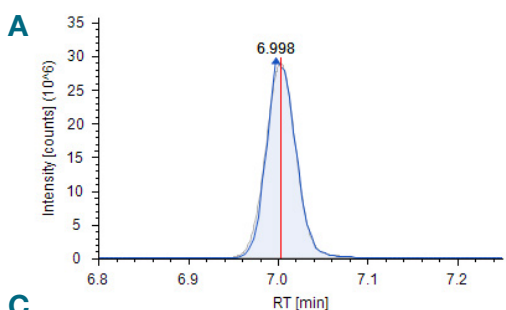


**Figure 4. Elemental composition for database searching.** A) MS<sup>1</sup> survey scan showing color-coded isotope pattern fit for the expected compound phenylalanine,  $m/z$  166.0863. The inset showing the A2 isotope cluster containing the  $^{13}\text{C}_2$  and  $^{18}\text{O}$  isotopes. B) MS<sup>2</sup> fragment spectrum for the isolated precursor ion of  $m/z$  166.0863. C) One candidate elemental formula for the expected compound and associated variables to support this: delta mass, the number of matched isotopes, and the number of matched fragment ions. D) Applying the elemental formula, C<sub>9</sub>H<sub>11</sub>NO<sub>2</sub> to the ChemSpider database search resulted in 419 candidate annotations.



| roposals                                                                          | Compounds per File                                                                | Predicted Compositions       | mzCloud Results | ChemSpider Results | Mass List Search Results |                     |                          |              |
|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|------------------------------|-----------------|--------------------|--------------------------|---------------------|--------------------------|--------------|
| Compound Match                                                                    | Structure                                                                         | Name                         | Formula         | Molecular Weight   | $\Delta$ Mass [Da]       | $\Delta$ Mass [ppm] | CSID                     | # References |
|  |  | DL-Phenylalanine             | C9 H11 N O2     | 165.07898          | -0.00005                 | -0.29               | <a href="#">969</a>      | 13669        |
|  |  | Benzocaine                   | C9 H11 N O2     | 165.07898          | -0.00005                 | -0.29               | <a href="#">13854242</a> | 1589         |
|  |  | L-(-)-Phenylalanine          | C9 H11 N O2     | 165.07898          | -0.00005                 | -0.29               | <a href="#">5910</a>     | 942          |
|  |  | D-(+)-Phenylalanine          | C9 H11 N O2     | 165.07898          | -0.00005                 | -0.29               | <a href="#">64639</a>    | 408          |
|  |  | 3-Amino-phenylpropionic acid | C9 H11 N O2     | 165.07898          | -0.00005                 | -0.29               | <a href="#">62403</a>    | 285          |

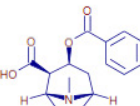
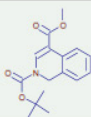
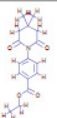
**Figure 5. Top five candidate annotations based on the ChemSpider database search for the expected compound phenylalanine, *m/z* 166.0863.** Here, putative candidates are ranked based on the number of references found in the ChemSpider database. Using this logic, the phenylalanine isomers, benzocaine, and 3-aminophenylpropionic acid are ranked highest.



**B**

| Name             | Formula                                          | Molecular Weight | RT [min] | # ChemSpider Results | Area (Max.) |
|------------------|--------------------------------------------------|------------------|----------|----------------------|-------------|
| Benzoyllecgonine | C <sub>16</sub> H <sub>19</sub> N O <sub>4</sub> | 289.13170        | 7.003    | 64                   | 1148631     |

**C**

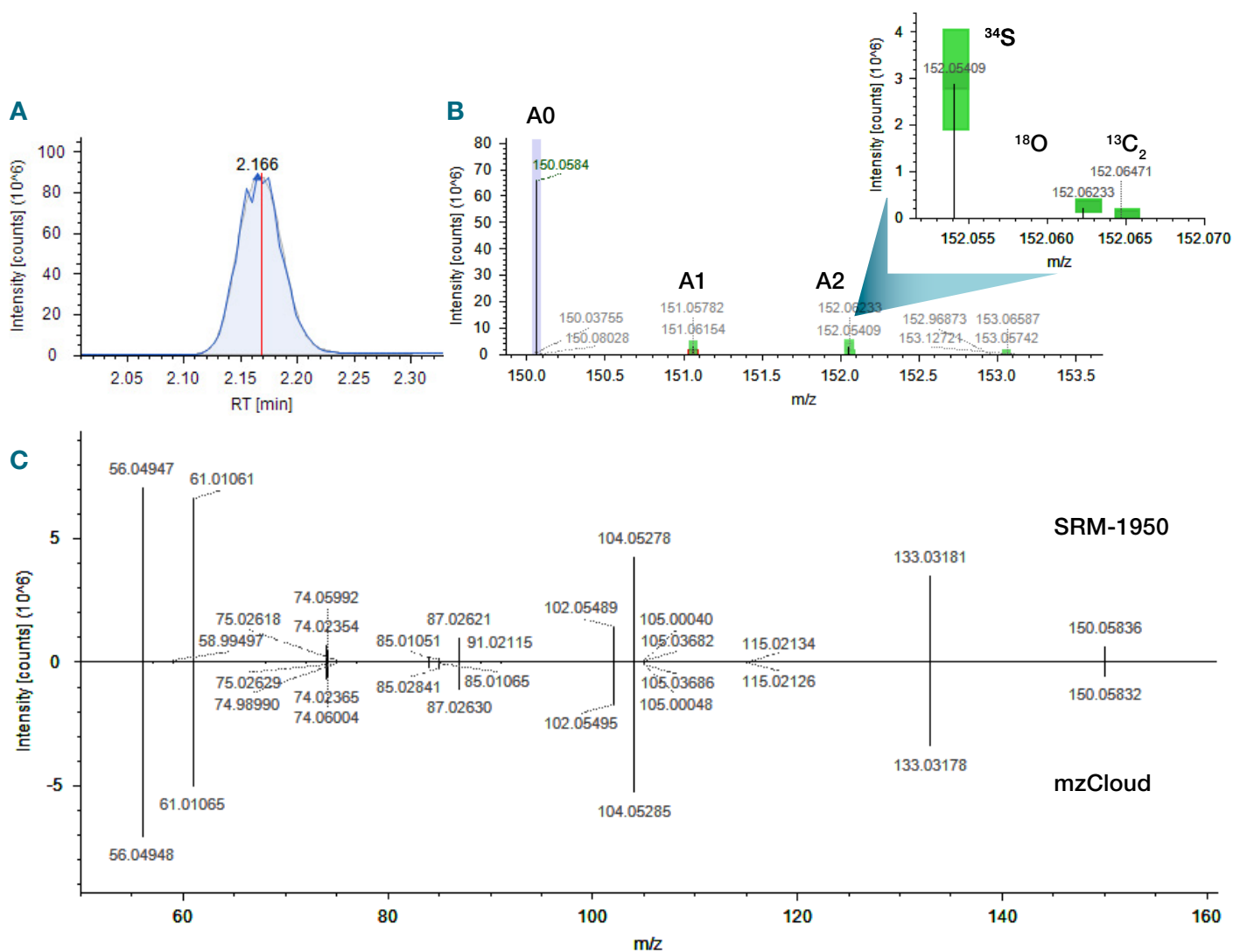
| Proposals                                                                           | Compounds per File                                                                  | Predicted Compositions                                   | mzCloud Results | ChemSpider Results | Mass List Search Results |                     |                          |              |
|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|----------------------------------------------------------|-----------------|--------------------|--------------------------|---------------------|--------------------------|--------------|
| Compound Match                                                                      | Structure                                                                           | Name                                                     | Formula         | Molecular Weight   | $\Delta$ Mass [Da]       | $\Delta$ Mass [ppm] | CSID                     | # References |
|  |  | Benzoyllecgonine                                         | C16 H19 N O4    | 289.13141          | -0.00029                 | -1.01               | <a href="#">395095</a>   | 75           |
|  |  | 2-tert-Butyl 4-methyl isoquinoline-2,4(1H)-dicarboxylate | C16 H19 N O4    | 289.13141          | -0.00029                 | -1.01               | <a href="#">24534363</a> | 65           |
|  |  | ethyl 4-(4,4-dimethyl-2,6-dioxopiperidino)benzoate       | C16 H19 N O4    | 289.13141          | -0.00029                 | -1.01               | <a href="#">2104375</a>  | 59           |

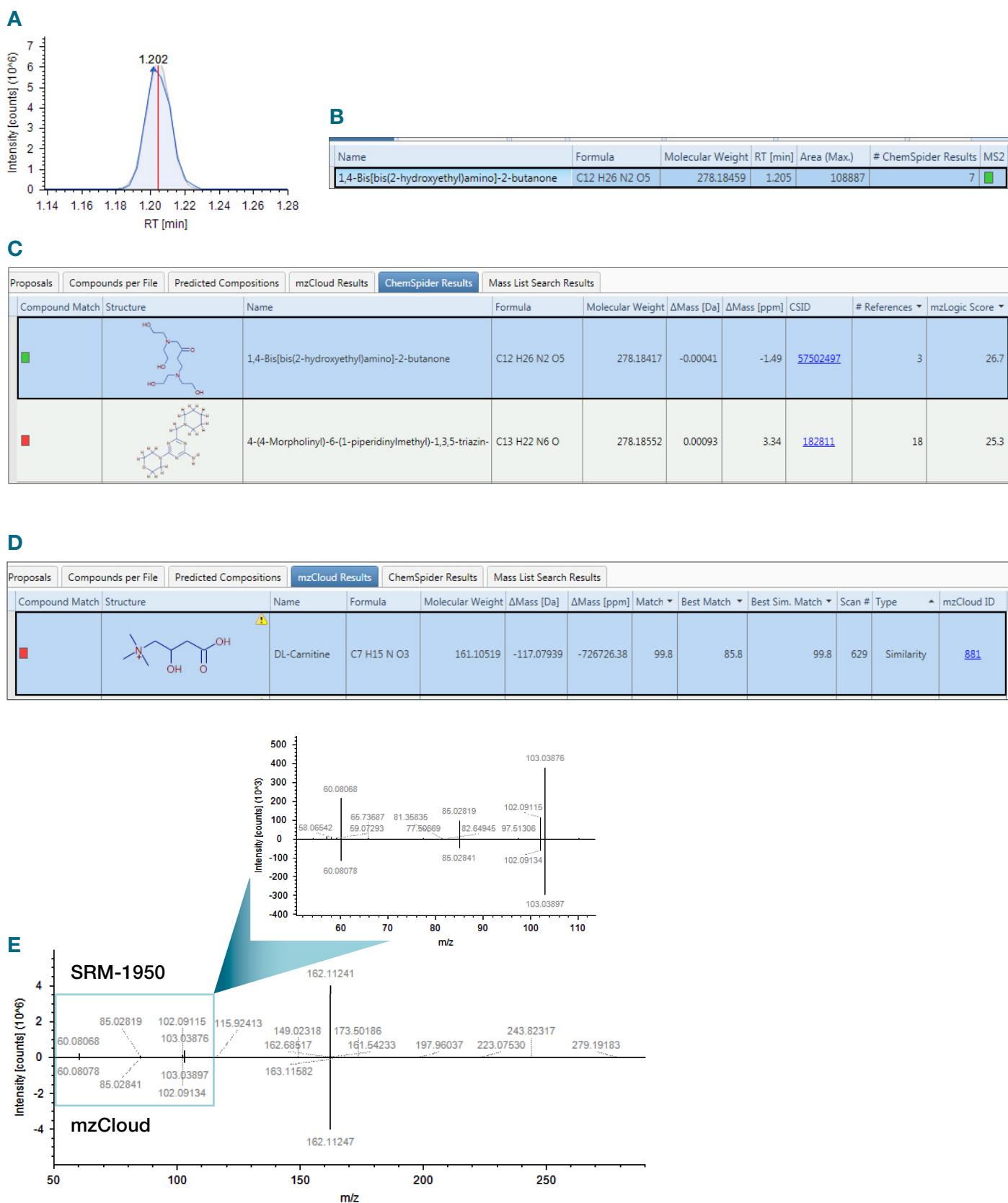
**Figure 6. Choosing data sources for endogenous and exogenous compounds.** A) Chromatogram of the expected compound, benzoyllecgonine. B) Searching the elemental formula, C<sub>16</sub>H<sub>19</sub>NO<sub>4</sub>, against the ChemSpider database using data sources containing both endogenous and exogenous compounds resulted in a total of 64 candidate annotations. C) The top-ranked candidate hit is benzoyllecgonine, a primary metabolite of the drug cocaine. Entries listed here are ranked based on the number of references for each compound in the ChemSpider database.

## The power of fragmentation spectra in parallel with a library hosting an immense number of spectra

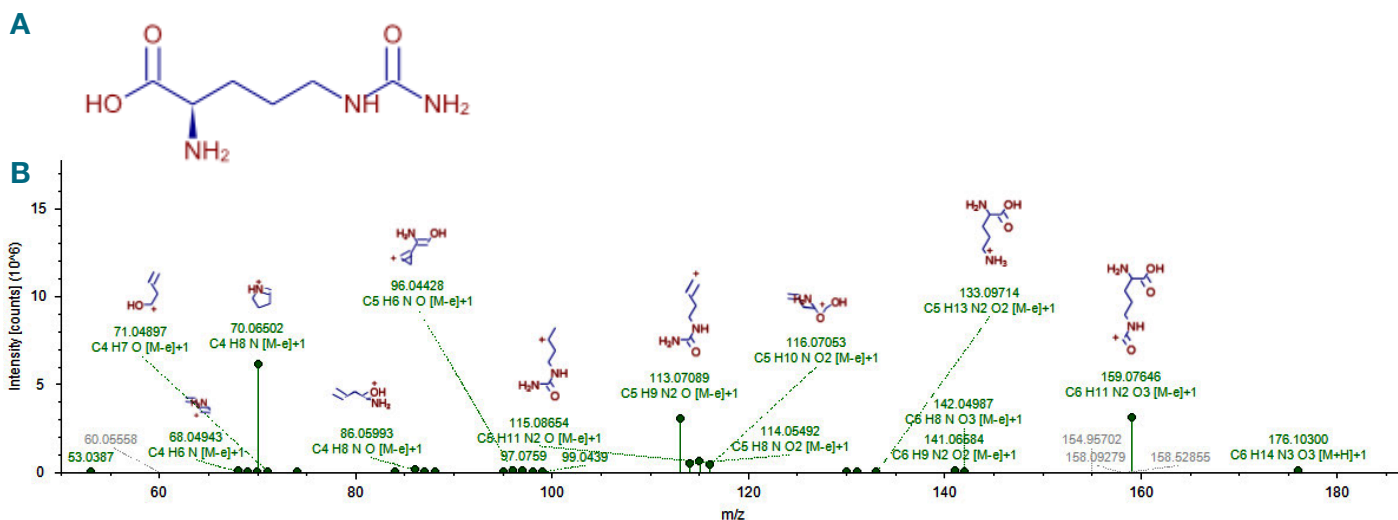
Coupling ion dissociation techniques to HRAM MS provides another measurement for compound characterization to build confidence toward unknown compound annotation. Isolation of the precursor ion population via a narrow isolation window improves the quality giving purer fragmentation spectra. For this reason, a data-dependent acquisition was chosen for the analysis and the high-quality fragmentation data was then used to search against a spectral library. Unlike a database, a spectral library is a collection of actual fragmentation spectra. Typically, each compound entry contains multiple spectra representing several different collision energies. The mzCloud spectral library is a highly curated, public library of endogenous and exogenous small molecules containing almost 3 million fragmentation spectra. Each compound entry in the library generally includes two collisional techniques: higher energy collision-induced dissociation (HCD) and ion trap resonant collision-induced dissociation (CID). For each dissociation technique, fragmentation spectra span a wide range of collision energies in iterations of 5 or 10%, thus generating a complete breakdown curve showing the reduction of precursor ion intensity with increasing collision energy while fragment ions increase concurrently. This systematic collection of spectra eliminates constraints in how data are acquired in terms of collision energy for subsequent spectral library matching. Furthermore, ion trap technology provides MS<sup>n</sup> with repeated isolation and fragmentation of product ions beyond MS<sup>2</sup>, such as MS<sup>3</sup>, MS<sup>4</sup> to MS<sup>n</sup>, producing spectral trees for each library entry. Finally, the mzCloud spectral library was generated using actual spectra collected from purified reference material. Each spectrum is recalibrated for exact mass and noise removed and is further structurally annotated making this a very high-quality spectral library. The mzCloud library can be searched by uploading individual fragmentation spectrum; however, the Compound Discoverer 3.0 software is fully integrated with the mzCloud library, enabling batch searching for a fully automated analysis. Searching against a high-quality, fully curated spectral library with ample fragmentation spectra from existing compounds provides reinforcing knowledge about the molecular makeup of an unknown compound.

Two search types were implemented: identity search and similarity search. An identity search matches both precursor ion and fragment ions, while a similarity search only searches for fragment ions, allowing for matches indicative of substructure. Figure 7 shows a mzCloud library match for the amino acid methionine of the chromatographic peak at 2.16 min (Figure 7A) and a monoisotopic peak at  $m/z$  150.0584 (Figure 7B). The full scan spectrum and the isotope fine structure is in complete accord with the elemental formula C<sub>5</sub>H<sub>11</sub>NO<sub>2</sub>S. Figure 7C displays the mirror plot for the MS<sup>2</sup> spectrum of the experimental data (top) and the matching spectrum found in the mzCloud library (bottom) where multiple ions overlap including the precursor ion and several product ions. This identity match further points toward a confident annotation of methionine. There are instances when the library may not contain an expected compound yet, the fragmentation data is still highly informative toward annotation, particularly when product ions match between the library and experimental data for substructure elucidation. This approach is commonly known as Precursor Ion Fingerprinting (PIF),<sup>9</sup> which is immensely insightful for determining the degree of similarity between unknown and candidate compounds. Figure 8 follows this evaluation process. A chromatographic peak at 1.20 min (Figure 8A) with a predicted chemical composition of C<sub>12</sub>H<sub>26</sub>N<sub>2</sub>O<sub>5</sub> generated seven ChemSpider database matches (Figure 8B). Of the seven candidates, the top-ranked entry is 1,4-bis[bis (2-hydroxyethyl)amino]-2-butanone (Figure 8C) based on the molecular weight of the precursor ion. At the same time, taking the fragmentation data into account resulted in the mzCloud library similarity match to carnitine (Figure 8D) suggesting this compound is related to carnitine given that there are several overlapping product ions in addition to the presence of the precursor ion for carnitine as seen in the mirror plot (Figure 8E). Though the molecular weight of the expected compound is heavier by 117.0794 Da compared to carnitine, the fragmentation data indicates a carnitine-like compound rather than the top-ranked hit from the ChemSpider database. Fragmentation spectra are meaningful to further build confidence toward unknown compound annotation.





**Figure 8. Similarity match—MS<sup>2</sup> spectral search against the mzCloud spectral library.** A) Chromatogram of the expected compound, 1,4-bis[bis(2-hydroxyethyl)amino]-2-butanone. B) Searching the elemental formula, C<sub>12</sub>H<sub>26</sub>N<sub>2</sub>O<sub>5</sub>, against the ChemSpider database resulted in a total of 7 candidate annotations. C) The top ranked candidate compound shows three references, with a delta mass of 1.29 ppm. D) The mzCloud library search resulted in a similarity match to carnitine, suggesting that this compound is rather related to carnitine, and thus carnitine-like. The molecular weight of the expected compound is 278.1846 Daltons (Da) while the molecular weight of carnitine is 161.1052 Da, a delta mass of 117.0794 Da. The caution triangle indicates a naturally positively charged compound. E) Mirror plot of the MS<sup>2</sup> fragmentation spectrum from the SRM-1950 (top) and the matched spectrum from the reference library in mzCloud (bottom). Several overlapping product ions are displayed including the precursor ion and lower molecular weight ions like  $m/z$  103.0397, 85.0284, and 60.0808 (zoomed area).



**Figure 9. FISH scoring using *in silico* prediction.** A) The chemical structure for the expected compound citrulline, molecular weight 175.0959 Da. B) MS<sup>2</sup> fragmentation spectrum for the isolated precursor ion of  $m/z$  176.1030. The FISH coverage score is 77.1% where 27 product ions were successfully matched (green) and 8 ions remain unmatched. Structural annotations are displayed for all matched ions when sufficient computer monitor display space is allowed. The FISH coverage score is determined by the number of matched centroids divided by the number of used (matched and unmatched) centroids multiplied by 100.

### Knowing where to break the chemical bond

Another way to evaluate fragmentation data of an unknown compound is to apply an *in silico* fragmentation prediction algorithm to parent structures in order to generate potential ion fragments. Applying reaction logic, *in silico* fragmentation prediction in the Compound Discoverer 3.0 software is generated based on the HighChem Fragmentation Library, which consists of 31,901 fragmentation schemes and 136,169 decoded mechanisms taken from literature publications based on mass spectrometry. Erroneous fragmentation mechanisms are eliminated using this practical approach since reaction mechanisms are rigorously evaluated, both manually and automatically, producing a high-quality database. The *in silico* processing method in the Compound Discoverer 3.0 software is by the FISH scoring algorithm. Incorporating FISH capability in an untargeted metabolomics analysis aids to structurally explain product ions present in the fragmentation spectrum via a proposed chemical structure. Figure 9 demonstrates the applicability of the FISH processing method. An expected compound in the SRM 1950 extract is proposed as citrulline. Both the ChemSpider database and mzCloud library matching support this annotation. Taking this one step further, FISH was then applied to the chemical structure citrulline (Figure 9A) and associated with the fragmentation spectrum generated for the corresponding precursor mass of  $m/z$  176.1030. Of the possible product ions assessed, more than 75% in the spectrum can be structurally related to citrulline (Figure 9B). While the selected example illustrates FISH applied to MS<sup>2</sup>, this processing method can be applied to any level of MS<sup>n</sup> data. The ability to define a product

ion by structural annotation for a proposed compound contributes additional knowledge in evaluating annotations for unknown compounds.

### The mzLogic algorithm—combining the database search with the mzCloud Spectral Library

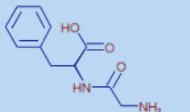
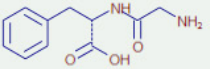
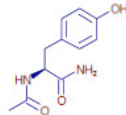
Real fragmentation data can also be used to rank order results from the ChemSpider database search. As previously described, a search against the ChemSpider database may result in numerous putative candidate compounds in which the number of references prioritizes the results list. This is a plausible approach; however, it does not take into consideration any structural information. Leveraging sub-structural knowledge and the extensive spectral fragmentation information in the mzCloud spectral library, the new mzLogic algorithm can prioritize the list of candidate compounds from the ChemSpider database based on the unknown's fragmentation spectra. The mzLogic algorithm first obtains all possible parent chemical structures resulting from the ChemSpider database search. Next, the results from the similarity search against the mzCloud library are included to map structural annotations of matched fragment ions back to the experimental data. The partial structures are overlaid to the parent structures generated from the database search to obtain the best fit. The candidate structures that can best be explained with the maximum common sub-structure and the highest spectral match score are then prioritized. The mzLogic score is easily viewed in the Compound Discoverer 3.0 software in a column format within the ChemSpider database results table, or any other result table containing chemical structure, such as the Mass List Search results table.



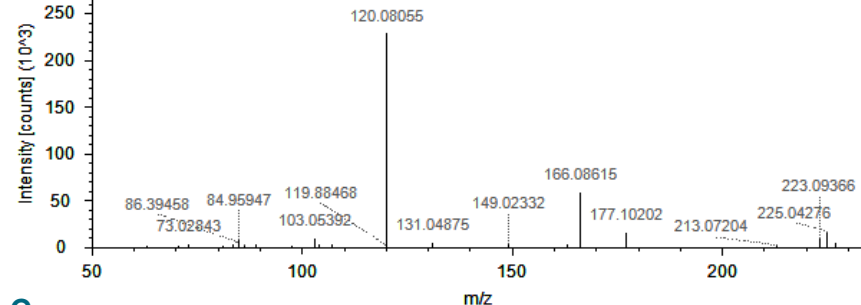
Figure 10 shows the application of the mzLogic algorithm to rank order the ChemSpider database search results to generate the annotation of Gly-Phe. The compound was top ranked with a score of 45.7 (Figure 10A) out of a total of 185 candidates (not shown). The fragmentation spectrum (Figure 10B) was used to perform a similarity search against the mzCloud library to generate fragment structures of similar compounds. The structural

annotations are used to determine the maximum common substructure to rank the ChemSpider database search results (Figure 10C). The mzLogic algorithm uses fragmentation information for structural explanation via the mzCloud library, thus providing another annotation tool contributing to confident annotations based on mass spectral fragmentation data.

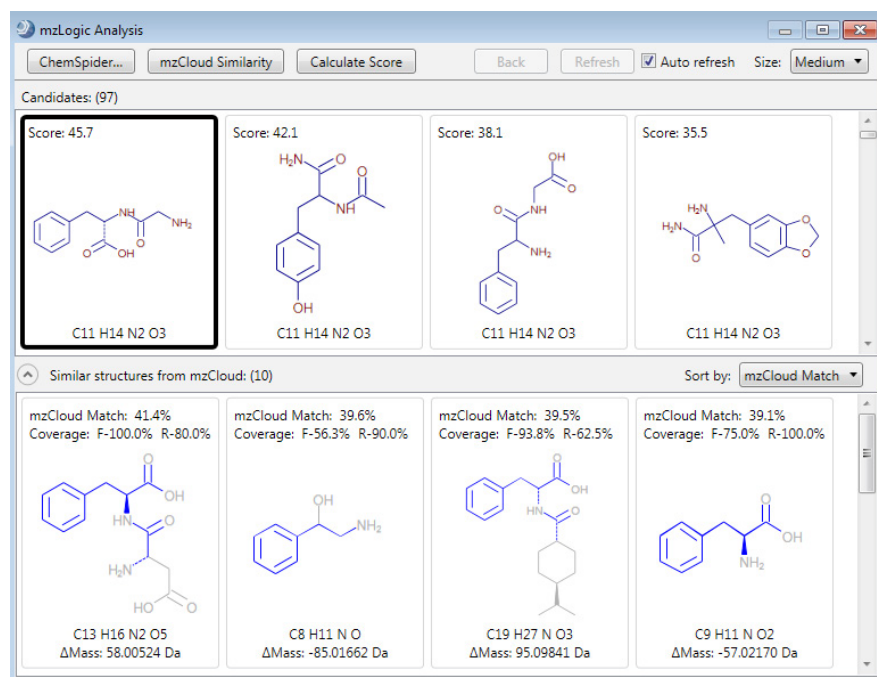
**A**

| Proposals                                                                         | Compounds per File                                                                | Predicted Compositions | mzCloud Results | ChemSpider Results | Mass List Search Results |                     |                        |              |               |
|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|------------------------|-----------------|--------------------|--------------------------|---------------------|------------------------|--------------|---------------|
| Compound Match                                                                    | Structure                                                                         | Name                   | Formula         | Molecular Weight   | $\Delta$ Mass [Da]       | $\Delta$ Mass [ppm] | CSID                   | # References | mzLogic Score |
|  |  | Gly-Phe                | C11 H14 N2 O3   | 222.10045          | -0.00023                 | -1.04               | <a href="#">87926</a>  | 116          | 45.7          |
|  |  | Glycyl-L-phenylalanine | C11 H14 N2 O3   | 222.10045          | -0.00023                 | -1.04               | <a href="#">83909</a>  | 102          | 45.7          |
|  |  | AC-TYR-NH2             | C11 H14 N2 O3   | 222.10045          | -0.00023                 | -1.04               | <a href="#">643100</a> | 56           | 42.1          |

**B**



**C**



**Figure 10. The mzLogic algorithm ranks database search results using fragment structures obtained via the mzCloud library.**

A) A ChemSpider database search resulted in 185 matched compounds (not shown) where the top match is the dipeptide Gly-Phe. With the mzLogic algorithm in place, putative candidates are first ranked based on mzLogic scoring followed by the number of references in the ChemSpider database. B) Fragmentation data were acquired for this compound, but spectral library search did not result in an identity match. C) Results from the mzLogic analysis. The top panel shows the top ranked structural candidates based on the mzLogic score. The bottom panel shows compound structures from the mzCloud library where structural similarity overlapping with the selected candidate compound is highlighted in blue. The black box shows Gly-Phe selected with a score of 45.7 and the corresponding structural similarity colored in blue.

## Consensus for increased annotation confidence and pathway certainty

Confidence in unknown compound annotation increases when multiple analytical measurements are collectively evaluated.<sup>10,11</sup> Several annotation tools were used for the analysis of the SRM 1950 human plasma extract. In the Assign Compound Annotation processing node, users can define the preference order of the annotation source for the assigned annotation. For this analysis, annotation assignments were based in the following order of data sources starting with the most preferred: mzCloud spectral library search, predicted elemental composition, mass list search, and the ChemSpider database search. Accordingly, complete matching across all sources for a given annotation using maximum information (full scan spectra, fragmentation spectra, and chemical structure) increases the confidence in the annotation. The consensus approach is easily viewed in the Compound Discoverer 3.0 software via a color-coded status indicator (Figure 11) illustrating the value of achieving agreement across multiple annotation sources. While not all compounds detected in the analysis may have full matches to all sources, there is still confirmatory information contributing to compound annotation as with partial matching or chemical relatedness (i.e., similarity search). Confidence in an unknown compound annotation is essential in untargeted metabolomics analysis. Using multiple analytical measurements to gain greater compound knowledge increases confidence in compound annotation. Confident compound annotations facilitate reliable biochemical pathway analyses.

| Name                                                  | Formula      | Annotation Source  |                |                   |                 |
|-------------------------------------------------------|--------------|-------------------------------------------------------------------------------------------------------|----------------|-------------------|-----------------|
|                                                       |              | Predicted Compositions *                                                                              | mzCloud Search | ChemSpider Search | MassList Search |
| 7-Methylguanine                                       | C6 H7 N5 O   |                                                                                                       |                |                   |                 |
| 4-Indolecarbaldehyde                                  | C9 H7 N O    |                                                                                                       |                |                   |                 |
| Kahweol                                               | C20 H26 O3   |                                                                                                       |                |                   |                 |
| 2-[[[(2-Benzoylphenyl)sulfonyl]methyl]-2-ethylhexanal | C22 H26 O2 S |                                                                                                       |                |                   |                 |
| [Similar to: Oxycarboxin; ΔMass: 29.0424 Da]          | C7 H10 O7 S  |                                                                                                       |                |                   |                 |

**Figure 11. Application of the consensus approach for confident compound annotation using results by several annotation sources generated from multiple analytical measurements.** Shown are five compounds listed in the Main Table results. The annotation source results are displayed in column format for each compound via colorized rectangles representing match status. The color green indicates a full match while the color orange is a partial match. The color red represents a mismatched mass in the case of the mzCloud library search (similarity match) and represents no match in the case of ChemSpider database search. The third compound listed is Kahweol, a diterpenoid molecule found in the beans of *Coffea arabica*. This compound matched all four sources used in the analysis: elemental composition prediction, mzCloud library search, ChemSpider database search, and the mass list search. Other compounds listed in the table, 7-methylguanine and 4-indolecaraldehyde, matched three sources, while the last two entries resulted in the mzCloud library similarity search only.

## Conclusion

The SRM 1950 human plasma extract was analyzed using the Orbitrap ID-X Tribrid MS and the Compound Discoverer 3.0 software. HRAM data were generated for MS<sup>1</sup> and MS<sup>2</sup>. Fragmentation spectra were acquired using the novel AcquireX Deep Scan acquisition strategy. The data-driven, intelligent acquisition approach prioritized precursor ion selection for subsequent fragmentation using ion inclusion and exclusion lists via preliminary full scan analysis of the blank and matrix sample. The unknown compounds were annotated using multiple annotation sources in the processing workflow to increase confidence in the assignment of unknown compound annotations. Predicted elemental composition from the high-quality mass spectra obtained using the HRAM Orbitrap instrument was used to search the ChemSpider database. The database provides access to several relevant metabolomics databases such as HMDB, BioCyc, KEGG, Yeast Metabolome Database in addition to chemically relevant sources of synthetic origin like ACToR, DrugBank, and EAWAG. Structure candidates produced from the ChemSpider database search were ranked using the mzLogic algorithm that takes into account experimental fragment ions. The fragmentation spectra were searched against the

mzCloud library using both identity and similarity search. For further confirmation, *in silico* fragmentation of proposed chemical structures using the FISH algorithm was used to structurally explain fragment ions.

Results generated from this experiment demonstrate the utility of analyzing multiple analytical measurements from mass spectral data against different data sources. Both endogenous and exogenous compounds were annotated in the human plasma extract. Expected endogenous metabolites like acetylcarnosine and phenylalanine were annotated by the predicted elemental composition and ChemSpider database search. Incorporating databases of synthetic origin allowed for the annotation of benzoylecgonine, a primary metabolite of the drug cocaine. The mzCloud library search generated an identity match to methionine while the similarity search detected carnitine-like compounds. *In silico* fragmentation from the FISH algorithm structurally explained more than 75% of the fragment ions for citrulline. The mzLogic algorithm enabled the Gly-Phe annotation to be prioritized out of 185 candidate compounds. In some instances, the compound annotation was the same across different sources indicating agreement. In other instances lacking an identity match, a partial match associated compound class. Employing different annotation tools in the Compound Discoverer 3.0 software to analyze the SRM 1950 extract for multiple analytical measurements built confidence in the annotation assignments of the unknown compounds.

## References

1. Patti, G., Dunn, W., Creek, D., and Sumner, L. (2018). The journey from features to compound identification in metabolomics when will we get there? *The Scientist*, eBook #65354.
2. Dunn, W.B., Erban, A., Weber, R.J.M., Creek, D.J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J., and Viant, M.R. (2013). Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, 9 (Suppl1), 44–66.
3. Loulou Peisl, B.Y., Schymanski, E.L., and Wilmes, P. (2018). Dark matter in host-microbiome metabolomics: tackling the unknowns—a review. *Analytica Chimica Acta*, 1037, 13–27.
4. Brown, M., Dunn, W.B., Dobson, P., Patel, Y., Winder, C.L., Francis-McIntyre, S., Begley, P., Carroll, K., Broadhurst, D., Tseng, A., Swainston, N., Spasic, I., Goodacre, R., and Kell, D.B. (2009). Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst*, 134, 1322–1332.
5. Peake, D. (2018). Review paper 65356: High-resolution compound identification in metabolomics: a review of current practices.
6. Stratton, T., and Tautenhahn, R. (2018). White paper 65210: Compounding insights for small molecule research Thermo Scientific Compound Discoverer software.
7. Souza, A., and Tautenhahn, R. (2018). Technical note 65204: Features or compounds? A data reduction for untargeted metabolomics to generate meaningful data.
8. Phinney, K.W. et al. (2013). Development of a Standard Reference Material for Metabolic Research. *Analytical Chemistry*, 85, 11732–11738.
9. Sheldon, M.T., Mistrik, R., and Croley, T.R. (2009). Determination of ion structures in structurally related compounds using precursor ion fingerprinting. *Journal of the American Society of Mass Spectrometry*, 20, 370–376.
10. Rochat, B. (2017). Proposed confidence scale and ID score in the identification of known-unknown compounds using high resolution MS data. *Journal of the American Society of Mass Spectrometry*, 28 (4), 709–723.
11. Blaženović, I., Kind, T., Ji, J., and Fiehn, O. (2018). Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites*, 8, 31.

Find out more at

[www.thermofisher.com/compounddiscoverer](http://www.thermofisher.com/compounddiscoverer)

**ThermoFisher**  
SCIENTIFIC