REVIEW PAPER 65356

High-resolution compound identification in metabolomics: a review of current practices

Author

David Peake, Thermo Fisher Scientific, San Jose, CA, USA

Keywords

Metabolomics, untargeted metabolomics, features, compounds, data reduction, compound identification, compound annotation, structural elucidation, spectral libraries, MS^n

Introduction

This article is intended for scientists who are thinking about the quality of their metabolomics data and how it is used in the context of compound identification. The goal is to discuss the challenges that researchers often face in the real-world identification of known and unknown small molecules. Specifically, we discuss the unique toolset provided by ultra-high-resolution accurate mass measurements and the dissociation techniques available on modern Thermo Scientific™ Orbitrap™ mass spectrometers.

In this review we consider the annotation of small molecule metabolites from high-resolution mass spectral methods including GC-MS and $LC\text{-}MS^2$ experimental approaches. We will explore what it means to identify compounds and how annotation of mass spectral features differs from the accepted standards for organic structure elucidation. We discuss the intrinsic value of chromatographic separation combined with high-resolution mass spectral measurements, the current methods used for compound annotation, and workflow strategies including new tools that are now coming to fruition including next generation software and very high-quality mass spectral libraries.

ThermoFisher SCIENTIFIC

Accepted methods for synthetic small molecule characterization

Modern synthetic organic chemistry and the well-established methods used to provide definitive characterization of new molecular structures provide a benchmark for small molecule characterization. In most cases, several orthogonal analytical methods, in addition to the known synthetic methods, are required to provide unequivocal proof of molecular structure. Spectroscopic methods typically used to characterize molecular structure include infrared FT-IR, UV-VIS and nuclear magnetic resonance spectroscopies (^{1}H and ^{13}C FT-NMR, and two-dimensional correlation methods such as TOCSY and HSQC). Chromatographic methods combined with mass spectrometry (GC-MS, LC-MS, IC-MS) provide a measure of the chromatographic purity and identity of the synthetic compound along with characterization of unwanted by-products and impurities. High resolution accurate mass spectral information also provides molecular mass and confirmation of elemental composition for molecular species present in a synthetic sample.

Full characterization of unknown molecules requires sufficient quantity of a purified sample to unequivocally establish the molecular structure. The highest level of confidence for any small molecule identification strategy therefore, includes isolating the pure compound obtained through a well-defined synthetic route and validation with various spectroscopic methods that include all of the compounds structural properties including stereochemistry.

Introduction to metabolomics

Metabolomics[1] is a newly emerging field of 'omics' research. Metabolites are the small molecule by-products produced by enzymatic processes of living cells. The metabolome, the complete set of endogenous metabolites, intermediates and metabolism products in a biological system, is connected to the genome, transcriptome and proteome and provides an instantaneous snapshot of the physiological state of any living being. Phenotypical changes to the metabolome in response to stress such as the environment or disease can inform research in life sciences, medicine and biomarker discovery.[2]

Typically, metabolomics analyses are performed by GC-MS or LC-MS[3-5] Full Scan mode using a high-resolution mass spectrometer. Unbiased studies are conducted in an untargeted manner in order to reveal the most complete fingerprint of metabolites related to either a normal or diseased state of an organism. In an untargeted profiling experiment the goal is to detect the relative changes in metabolite concentration and annotate metabolites initially based on accurate mass-to-charge ratio (*m/z*) and retention time.

GC-MS and LC-MS profiling methods provide complementary data needed to perform metabolomics studies on a wide range of analytes. There is also some overlap of molecular classes amenable to either the GC-MS or LC-MS approach (Figure 1) which also provides useful cross-validation. Some of the advantages and disadvantages of using GC-MS and LC-MS approaches are summarized in Table 1.

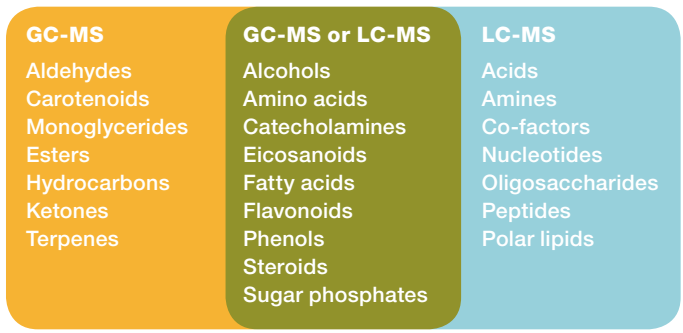


Figure 1. Compound classes amenable to GC-MS vs LC-MS.

In a typical GC-MS or LC-MS analysis, the choice of chromatographic technique and ionization method limits the metabolite coverage for large-scale discovery metabolomics. Endogenous metabolites are extremely diverse in their physical-chemical properties and concentrations leading to several analytical challenges in any metabolomics study. Several orthogonal sample preparation and chromatographic separations are needed to cover compounds with a very broad range of polarity and hydrophobicity. The choice of ionization method will be one of several factors determining the dynamic range of mass spectral analysis. Utilizing both positive and negative ion modes, or chemical ionization, improves the probability that more diverse compounds are ionized and detected. Consequently, a single untargeted method is more limited in breadth of coverage, thus obviating the use of a single method for truly global metabolomics studies. Conversely, targeted methods designed for particular metabolite classes provide optimized conditions for separation and detection of molecules sharing related structures and ionizable functional groups, but with limited coverage.

Table 1. Comparison of GC-MS and LC-MS approaches to metabolomics.

Comparison	GC-MS	LC-MS
Compounds	<700 Da; volatile at 350 °C	Polar/nonpolar, thermally labile
Sample preparation	Derivatization required	
Chromatography	Excellent isomer resolution	RP/HILIC/IC depending on polarity
Ionization	EI, Pos CI, Neg CI	ESI, APCI
Matrix effects	EI insensitive to	ESI quite sensitive
MS acquisition	MS, low or high resolution	MS, MS/MS (MS^2) high resolution
Data analysis	MS deconvolution required	Feature grouping of adducts
Annotation	MS library search	Molecular formula, DB search
Mass spectral libraries	EI (NIST/EPA/NIH)	MS^2 (Wiley, HMDB, Metlin)

We need to distinguish at least three types of unknown compounds in metabolomics studies:

1. "known unknowns" are endogenous metabolites from known metabolic pathways and are found in existing databases or mass spectral libraries;
2. "unknown unknowns" are unexpected but real compounds of biological origin; and
3. unknowns from chemical background or experimentally unrelated sources.[6]

In GC-MS studies, identification is facilitated by searching large established libraries of EI (electron ionization) mass spectra. Given the mature state of GC-MS libraries (NIST/EPA/NIH 2017 contains 262,150 compounds), identification of all sufficiently volatile organic compounds including drugs and endogenous metabolites, and their silylated derivatives is relatively straightforward. This is because GC-MS with EI generates reproducible molecular fragmentation patterns, making it an essential tool for metabolite identification. The ionizing 70-eV electrons yield radical cations that dissociate readily to give many structurally informative fragment ions, thus providing rich structural details of the molecular species.

For example, Figure 2a shows the EI mass spectrum observed for Glutamic acid from human plasma derivatized with three trimethylsilyl (3TMS) groups. Figure 2b illustrates the molecular ion, and elemental composition by EI, and confirmation by methane chemical ionization (CI) of the protonated molecular ion.

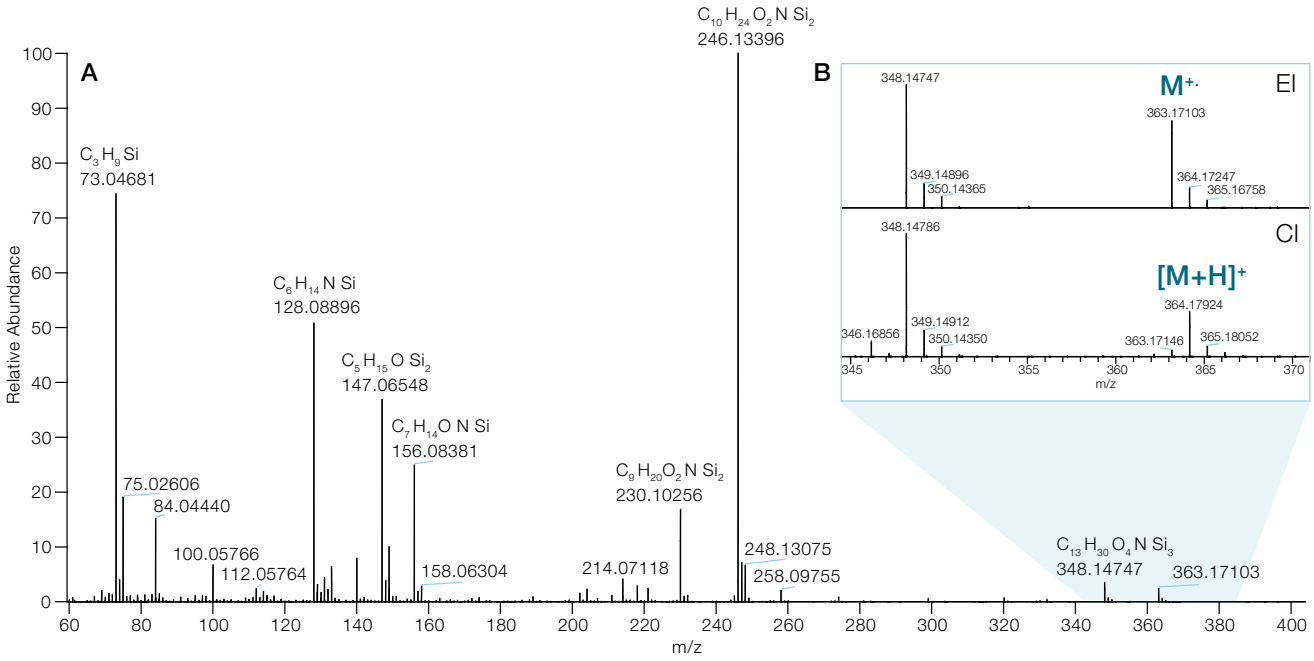


Figure 2. a) EI high resolution (Thermo Scientific™ Q Exactive™ GC Orbitrap™ GC-MS/MS system 60,000 FWHM at *m/z* 200) mass spectrum of Glutamic acid (3TMS, $C_{14}H_{33}O_4NSi_3$) with base peak *m/z* 246.13396, loss of $(CH_3)_3Si$-OC = O. radical from the $M^{+\cdot}$. and b) Region of the EI mass spectrum showing the molecular ion, $M^{+\cdot}$. (m/z 363.17103, -0.44 ppm), and the methane CI positive ion mass spectrum confirming the protonated molecular ion, $[M+H]^+$ (*m/z* 364.17924, 0.63 ppm).

In contrast, for LC-MS studies, putative annotation is typically accomplished by searching databases of known metabolites by precursor ion or elemental formula determined by accurate mass data. Subsequent LC-MS^2 experiments that generate fragmentation or product ions are then performed to determine the further confirmation of metabolites of interest by searching appropriate tandem mass spectral libraries. Since current MS^2 libraries are more limited in scope by the number of entries (Wiley 2017 MS/MS 13,808 compounds) and dependent on specific instrumental conditions, small molecule identification using LC-MS^2 is often a process that gives incomplete structural information. In some metabolomic peer reviewed publications, features (*m/z*) are reported without any annotation. Thus, the real crisis in LC-MS untargeted metabolomics is a frequent lack of identification with no real ability for follow up.

Identification of metabolites using GC-MS and LC-MS

The identification of potentially hundreds to thousands of metabolites in complex matrices such as plasma or urine is one of the most difficult challenges faced by metabolomics scientists. One of the main strategies employed in metabolomics is to reduce this complexity by applying several different chromatographic methods to separate very polar, moderately polar and hydrophobic (non-polar) metabolites prior to mass analysis. For example, capillary electrophoresis (CE), capillary gas chromatography (GC), hydrophilic interaction liquid chromatography (HILIC), ion chromatography (IC) or reversed phase HPLC separation methods may be employed. The various chromatographic techniques are typically interfaced to a high-resolution mass spectrometer via either a direct heated GC-MS inlet or by atmospheric pressure with an electrospray ionization (ESI) LC-MS interface.

GC-MS

Gas chromatography coupled to mass spectrometry (GC-MS) was developed prior to LC-MS techniques and is well-suited to analysis of sufficiently volatile and thermally stable organic compounds or their more volatile/stable derivatives. Ionization in GC-MS is typically performed in positive ion EI mode or alternatively with chemical ionization (CI) in positive ion or electron capture CI in the negative ion polarity. Recently, atmospheric pressure APCI has become another popular means for interfacing gas chromatography to LC-MS instrumentation.

One finite limitation of the GC-MS approach is that gas chromatography requires that analytes are volatilized at the temperature of the GC injector inlet (typically 350 °C) which limits the mass range of derivatized compounds to less than approximately 700 Da. Compounds with active hydrogens usually need to be derivatized with a combination of methoxime and trimethylsilyl (TMS) reagents prior to injection. Since the EI process forms mixtures of molecular ions ($M^{+\cdot}$) and fragment ions, a deconvolution process is required to group molecular ions, fragment ions and their isotopes arriving at the same time prior to a library search.

The main advantages of using high resolution accurate mass (HRAM) GC-MS for metabolomics analyses are high chromatographic resolution, reproducible retention times, robust quantitation, high specificity, high sensitivity and confident compound identification using existing commercial mass spectral libraries or dedicated HRAM libraries. Most GC-MS libraries were acquired in the past with low mass resolution quadrupole mass spectrometers. Although high resolution is not required to obtain spectral matches from EI GC-MS libraries, HRAM GC-MS data provides higher specificity and confidence in compound identification particularly for metabolomics applications in complex sample matrices such as human plasma.

LC-MS

In LC-MS with electrospray ionization molecular ions are formed as protonated or deprotonated singly-charged adducts, and more rarely doubly-charged species. In addition, mobile phase modifiers and salts in the biological extracts give rise to other adduct ions including ammonium, sodium, potassium, formate and chloride. Combinations of neutral molecules and adducts may also give rise to dimer adduct species. In some cases, labile molecules may eliminate water or other neutral species to form fragment ions in the ion source. For every adduct ion, there are potentially multiple isotopic peaks for the naturally-occurring stable isotopes such as ^{13}C, ^{2}H, ^{18}O, ^{15}N, and ^{34}S. Given this complexity, when uncontrolled or unaccounted for, the probability for false positives due to unassigned or improperly annotated adducts increases particularly for poorly ionizing species.

High resolution and accurate mass spectral information is acquired in order to enable automated and confident analysis of the LC-MS data. One key advantage of obtaining an accurate mass measurement is the ability to assign elemental composition to the monoisotopic species based on a specified mass tolerance. The accurate mass information is enhanced by sufficiently high mass resolution, or the ability to separate isobaric ions of similar mass to charge.

Accurate mass spectral information may be interpreted to correctly identify the elemental composition of one or more metabolites that elute within the same chromatographic peak.[7,8] For example, the exact mass difference between a protonated species, $[M+H]^+$ and the sodium adduct ion, $[M+Na]^+$ is 21.98194 amu and the mass difference between the mono-isotopic ^{12}C and ^{13}C species is 1.00335 amu.[9] Figure 3 shows the mass spectrum of Tryptophan from human plasma acquired at 240,000 resolution with the protonated molecular ion, other adducts and dimer ions assigned by their accurate mass differences.

In high resolution LC-MS analysis, accurate mass is obtained to establish the identity of metabolites. For compounds measured with *sufficiently* high resolution the accurate mass and isotopic fine structure can be used to establish unequivocally elemental composition. Figure 4 shows the region around the protonated molecular ion of Methionine (*m/z* 205.05818) from human plasma acquired at 240,000 mass resolution (at *m/z* 200) and the isotopic fine structure present at the A1 and A2 peaks. The isotopic fine structure reveals the presence of C, H, N, O, and S. The ratios of the isotopomers present in the mass spectrum support the unambiguous assignment of elemental formula, $C_5H_{11}NO_2S$.

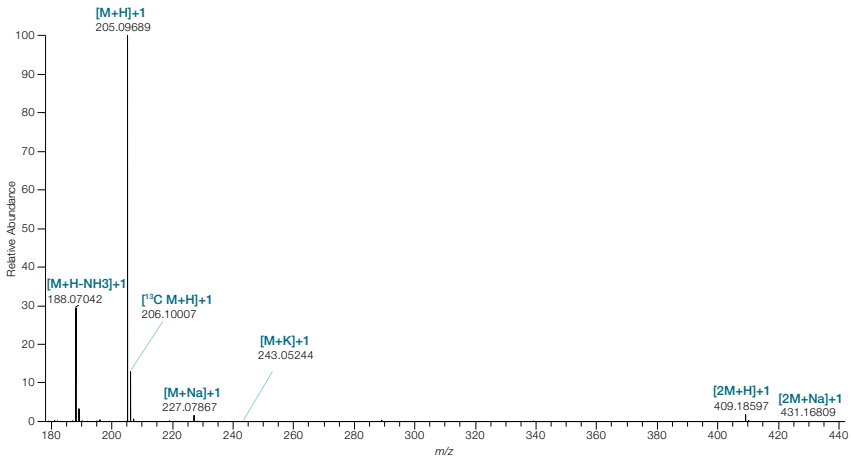


Figure 3. ESI positive ion mass spectrum (Res. = 240K, Thermo Scientific™ Q Exactive™ HF hybrid quadrupole-Orbitrap mass spectrometer) of Tryptophan from NIST SRM1950 human plasma, showing the protonated molecular ion and associated adducts and dimer ions.

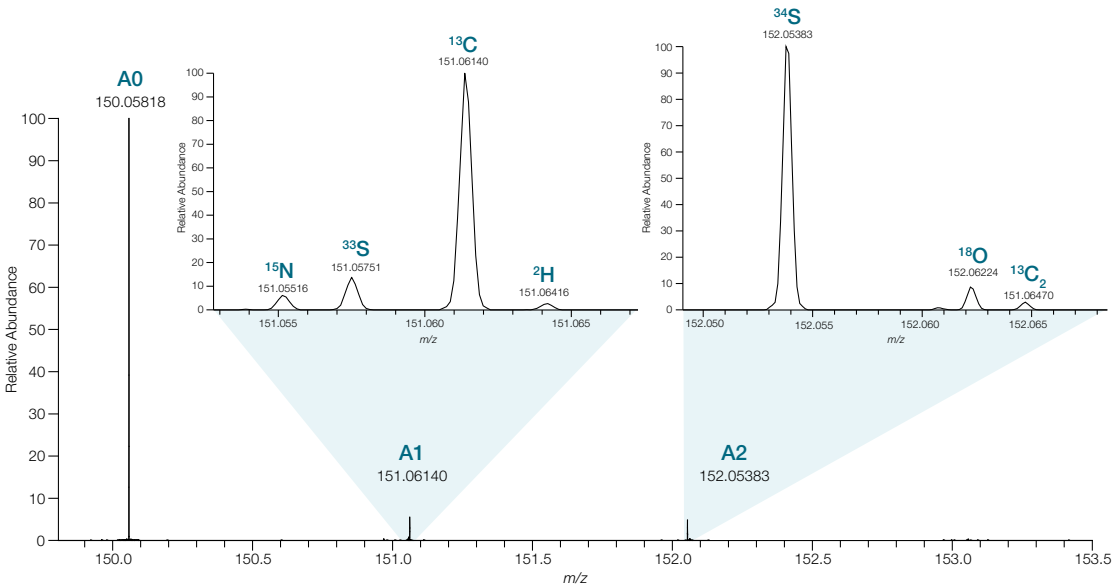


Figure 4. Positive ion mass spectrum (Res. = 240K, Q Exactive HF mass spectrometer) of Methionine showing the M+H ion and isotopic fine structure. The ratios of the isotopomers present in the A1 and A2 regions is entirely consistent with the elemental composition of $C_5H_{11}NO_2S$.

Compound identification reporting guidelines

One of the key issues currently being addressed by the metabolomics community is the lack of consistent standards for reporting compound annotation in the literature. The European Union (EU) has already established more stringent criteria related to small molecule identification.

We briefly review the current status here to make sure the reader is aware of these efforts.

Metabolomics Society guidelines

The metabolomics community is continuing work to establish overall guidelines concerning how to document identification levels achieved in metabolomics studies. The Metabolomics Standards Initiative[10] (MSI) chemical analysis working group proposed guidelines[11] describing four different levels of identification in 2007. Briefly, MSI levels 1–4 are described in Table 2.

MSI proposed update

At the Metabolomics Society conference in 2017 the MSI proposed adding an additional level of confidence. The additional compound level "0" was proposed to account for unambiguous compound identification using a combination of MS, NMR and comparison to synthetic standards or reference compounds (Table 2).

Proposal for combining EU and MSI guidelines

Bertrand Rochat's proposal to combine environmental food safety and metabolomics guidelines was recently published.[12] This proposal recognizes that at a given "level" of identification there may be fundamental differences in the quality of the data used and proposes combining "General ID category" with "Chromatography class", "Identification points" and "Identification confidence". This proposal gives more flexibility to assigning level 1 with or without available standards, depending on the confidence in the identification.

Table 2. Proposed minimum reporting standards for chemical analysis.

Level	Name	Minimum requirements
0	Unambiguous compound identification	Full structure elucidation using a combination of MS and 2D NMR (correlation of ^{1}H and ^{13}C nuclei) and other methods including compound isolation and purification; chemical synthesis; reference compounds
1	Identified compounds	At least 2 independent and orthogonal data relative to an authentic compound analyzed under identical conditions (retention time and mass spectrum, accurate mass and MS^2, accurate mass and isotopic pattern)
2	Putatively annotated compounds	Similar to level 1, but based on literature values reported for authentic samples by other laboratories
3	Putatively characterized compound classes	Based upon characteristic physicochemical properties of a chemical class of compounds or by spectral similarity to known compounds of a chemical class
4	Unknown compounds	These metabolites are differentiated based upon their mass spectral data

High resolution mass spectrometry for more confident compound annotation

Global metabolomics had its genesis in NMR and nominal mass GC-MS profiling of metabolites. With the advent of atmospheric pressure interfaces, triple quadrupole systems were developed for targeted GC-MS^2 and LC-MS^2 analysis of metabolites. However, effective methods for untargeted metabolomics require higher resolution instruments such as TOF, QTOF, FT-ICR and Orbitrap FTMS instruments.[9,13] Without the higher specificity and low ppm mass accuracy afforded by these instruments it is much more difficult to confidently annotate small molecule metabolites.

The Orbitrap mass spectrometer was first introduced in 2005 as a hybrid linear ion trap Orbitrap system.[14,15] Subsequently, the Thermo Scientific™ Q Exactive™ system was introduced in 2011[16], the high field Thermo Scientific™ Orbitrap Fusion™ MS system in 2013[17] and Thermo Scientific™ Q Exactive GC-MS system was introduced in 2015[18,19]. The combination of high resolution, high dynamic range, high sensitivity and low ppm mass accuracy provides a unique combination of attributes (Table 3) in a single high-resolution instrument.[20]

One key advantage for Orbitrap mass spectrometers is that a single MS or MS^2 experiment provides high mass accuracy from very low to very high precursor ion abundances.[21] The number of ions in the Orbitrap analyzer is detected via measuring their charge and mass accuracy is determined by precisely measuring the ions frequency.[22] Fourier Transform mass spectrometry is a fundamentally different detection process than the ion counting process in triple quadrupole or time-of-flight (TOF) instruments.[23] Mass accuracy in TOF MS requires summing individual ion counts in order to obtain a peak profile and accurate mass centroid. At low ion abundance, many spectra may need to be summed to obtain good mass measurement accuracy which limits the quality of data dependent MS^2 analysis.

In the Orbitrap analyzer, mass resolution increases with decreasing *m/z* giving higher resolution for small molecules and their fragment ions. Higher resolution does not come at a cost in sensitivity such as in other instruments. Higher resolution analysis simply requires more time in the Orbitrap for mass analysis. For example, in a "high field" Orbitrap mass analysis at 30K, 60K, 120K, and 240K resolution (full width half maximum at *m/z* 200) requires a 64, 128, 256, and 512 millisecond transient, respectively. Note that the number of scans across a chromatographic peak is adequate even at more than 100,000 resolution for high quality relative and absolute quantitation. This is not the case for FT-ICR instruments in general.

Table 3. Orbitrap instrument developments for enhanced structure elucidation.

Year	Instrument advance	Structure elucidation improvements
2005	Thermo Scientific™ LTQ Orbitrap XL™ Hybrid Ion Trap-Orbitrap mass spectrometer: 100K Resolution at *m/z* 400; >5000:1 dynamic range in mass accuracy (5ppm)	Ion trap CID and MS^n capabilities for determining compound structure; Better mass accuracy for all ions within a single MS or MS^2 scan; Higher resolution needed for separation of isobaric interferences.
2007	Collision cell HCD MS^2	Higher energy (HCD) MS^2 for improved small molecule characterization and identification
2011	Q Exactive mass spectrometer: 140K resolution, HCD MS^2	Routine accurate mass LC-MS and MS^2 for improved metabolite profiling and identification
2013	Orbitrap Fusion Tribrid mass spectrometer: 500K resolution	High speed MS^n with excellent speed and data quality for structure elucidation; Isotope fine structure reveals the elements present and their ratios (elemental composition)
2014	Q Exactive GC mass spectrometer: 100K resolution and 1ppm mass accuracy	Routine accurate mass GC-MS (EI, CI) and MS^2 for improved metabolite profiling and identification
2017	Thermo Scientific™ Orbitrap Fusion™ Lumos™ Tribrid™ mass spectrometer with 1M resolution and UV photodissociation (PD)	Higher resolution needed for determining isotopomers in labeled metabolites and lipids; UVPD for obtaining unique structural details (location of double bonds)

For difficult structure elucidation problems, the Orbitrap Fusion Tribrid mass spectrometer (combining quadrupole, ion trap and Orbitrap mass analysis) provides a combination of higher-energy collision cell fragmentation (HCD) and lower-energy linear ion trap collision induced dissociation (CID) with sequential MS^n capabilities. This instrument provides real time high-resolution accurate mass LC-MS and MS^n analyses for determining the elemental composition of small molecules and their sub-structural product ions. Ultrahigh-resolution analysis at 240K–500K (Orbitrap Fusion mass spectrometer) and 1 M (million) resolution on Orbitrap Fusion Lumos mass spectrometer[24] facilitates separating ^2H, ^{13}C, ^{15}N and ^{18}O isotopically labeled species thus, enabling further mechanistic structural investigations.

An alternative dissociation method for small molecules, 213 nm UV-laser photo dissociation (UVPD), is an available option on the advanced Orbitrap Fusion Lumos MS system.

Data processing workflows

Data processing workflows in unknown metabolomics analysis involves the following general steps including:

1. Chromatographic alignment
2. Unknown peak detection
3. Feature grouping/deconvolution
4. Background removal
5. Statistical analysis
6. Database/library searches and
7. Compound annotation

The specific steps required depends on the analytical technique and the acquisition method (i.e. Full Scan mode or data dependent acquisition). For example, GC-MS EI data processing starts with peak picking followed by deconvolution of the MS peaks based on chromatographic profiles for each extracted ion (Figure 5a). This step is required to provide a mass spectrum with a molecular ion (if present) and fragment ions related to the same compound. This step is similar to grouping related features into a feature group in the corresponding LC-MS workflow (Figure 5b). Background removal is often performed to remove chemical noise followed by either a database search (LC-MS, MS^2) or library search (GC-MS, $LC\text{-}MS^2$) for identification. Compounds are then annotated based on the most confident database and/or library search results. Thermo Scientific™ LipidSearch™ software relies on matching the predicted lipid fragmentation to MS^2 data for annotation.

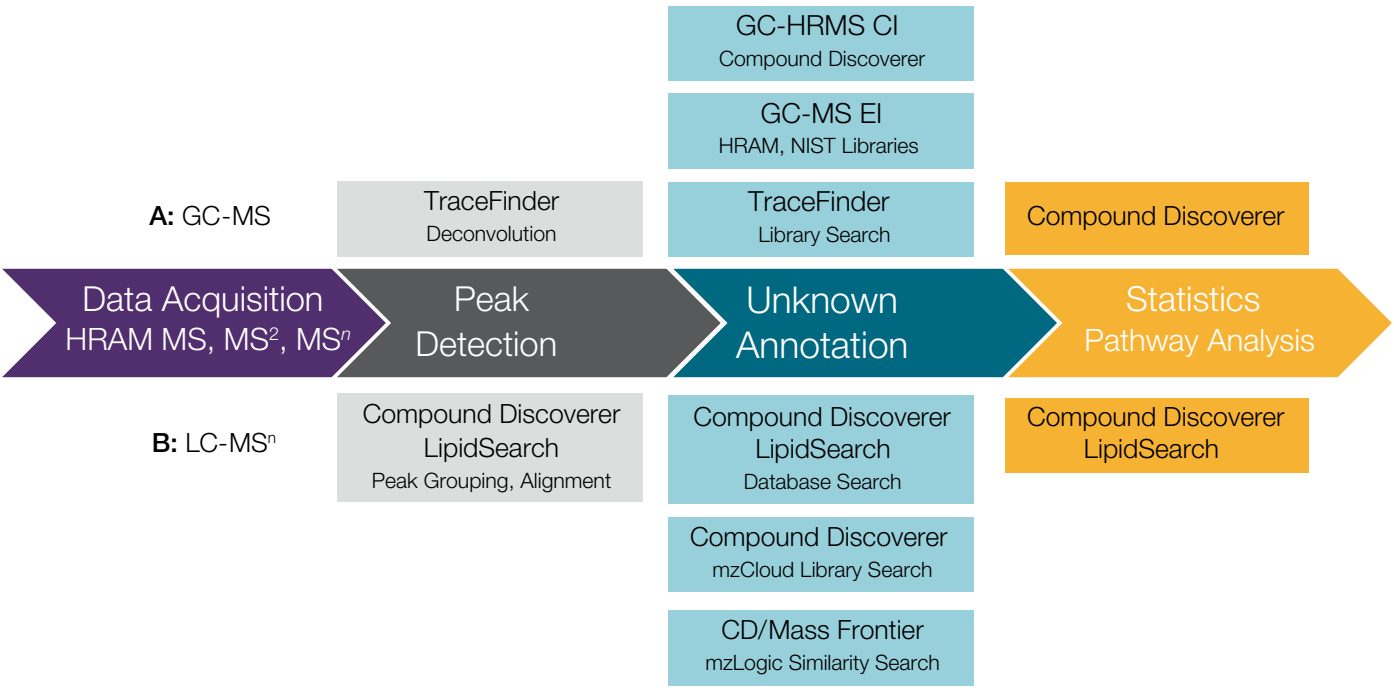


Figure 5. High resolution accurate mass workflows for: a) GC-MS using Thermo Scientific™ TraceFinder™ software for deconvolution and library searches, and Thermo Scientific™ Compound Discoverer™ software for statistics, and b) LC-MS and MS^2 using Compound Discoverer software for metabolomics data processing and LipidSearch software for $LC\text{-}MS^2$ and MS^3 lipidomics data processing. Compound Discoverer software is used for automated searching of the mzCloud mass spectral library including similarity searches, and Thermo Scientific™ Mass Frontier™ spectral interpretation software is used for metabolite identification and $LC\text{-}MS^n$ spectral tree searches.

Databases

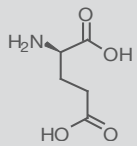
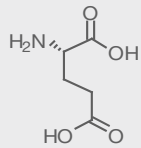
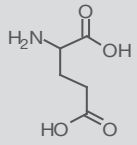
A database is a repository of chemical structures and information for individual molecules including chemical name, identifiers, elemental composition, molecular mass, intrinsic and calculated chemical properties, and often contains cross references to other small molecule and metabolite databases.

Some databases are very general while others are designed for specific applications. ChemSpider is a public database with 251 different sources that provides search access to over 67 million structures. Other compound databases include CAS (Chemical Abstracts Services) curated from thousands of journals, DNP (dictionary of natural products), FooDB (food) and PubChem (NIH, published chemicals). Highly-specialized metabolite databases include HMDB (human metabolome database) and KEGG (Kyoto encyclopedia of genes and genomes) which maps metabolites to biologically-relevant enzymatic pathways.

The first step of GC-MS or LC-MS compound identification usually proceeds by a molecular formula search of known compound databases. Formula searches yield potential candidate molecules but need to be restricted by one or more methods including: choosing databases relevant to the analytical samples, mass measurement tolerance, isotope pattern and isotope fine structure (very high resolution) and by using MS^2 data.

The partial database entries for Glutamic acid from three different sources are illustrated in Table 4. Accurate mass GC-MS (Figure 2) provides the underivatized formula of $C_5H_9NO_4$ which is matched to DL Glutamic acid [HMDB0060475, KEGG C00302, CSID 591]. Although L-Glutamic acid is the most common candidate from human plasma, racemic Glutamic acid is reported since the GC-MS chromatographic method was not capable of resolving the enantiomers.

Table 4. HMDB, KEGG and ChemSpider database entries for Glutamic acid.

HMDB identifier	CAS number	KEGG identifier	ChemSpider identifier	Names	Structure	Formula	Monoisotopic mass
HMDB0000148	56-86-0	C00025 D00007	30572	L-Glutamic acid, (2S)-2-Amino-pentanedioic acid		$C_5H_9NO_4$	147.053157781
HMDB0003339	6893-26-1	C00217		D-Glutamic acid, (2R)-2-Amino-pentanedioic acid		$C_5H_9NO_4$	147.053157781
HMDB0060475	617-65-2	C00302 D04341	591	DL-Glutamic acid, 4-Amino-1,5-pentandioic acid		$C_5H_9NO_4$	147.053157781

Mass spectral libraries

Libraries are a collection of full mass spectra of metabolites obtained under well-defined experimental conditions (LC-MS ESI, GC-MS EI) including tandem (nano-infusion, flow injection (FIA) or LC-MS^2) mass spectra typically acquired under one or more collision energies. The quality of the mass spectral library depends on the individual contributors as well as curation performed by the library owner. Many libraries contain nominal mass and high resolution accurate mass spectra depending on the instrument type.

Mass spectral libraries for EI spectra are published by Wiley and the latest update to the NIST/EPA/NIH library contains EI spectra (262,150 compounds), MS^2 (13,600 compounds) and a compilation of GC retention indexes for 99,400 compounds. Other mass spectral databases include Fiehn GC-MS (~700 compounds), HMDB (800 MS^2 and 260 GC-MS compounds), MassBank (1,900 compounds) and METLIN (~7,000 compounds). However, MS^2 library search results using these databases may vary widely since the library quality can differ because of significant differences in acquired mass resolution, collision gas, collision energies and instrument types. In addition, some libraries such as METLIN contain both mass spectra and predicted fragmentation typically found in compound databases.

mzCloud™ (mzcloud.org) is a completely new type of mass spectral library (>8,141 compounds; 2,773,741 spectra) based on MS^n spectral trees[25] acquired at more than 100K resolution on hybrid Orbitrap mass spectrometers. A mass spectral tree includes HCD MS^2 and CID MS^n spectra obtained under a full range of collision energies as shown in Figure 6 for the flavonoid Rutin. Better mass spectral matches are consistently achieved by matching the library to the experimental conditions instead of the traditional approach of matching the experimental conditions to the library.

Finally, in order to confirm the compound(s) identified from a mass spectral search the experimental results must be compared to authentic MS and MS^2 spectra, and the retention time of a standard run under identical experimental conditions. However, in some cases where the reference compound is not available the researcher must use caution in not over-annotating metabolite data.

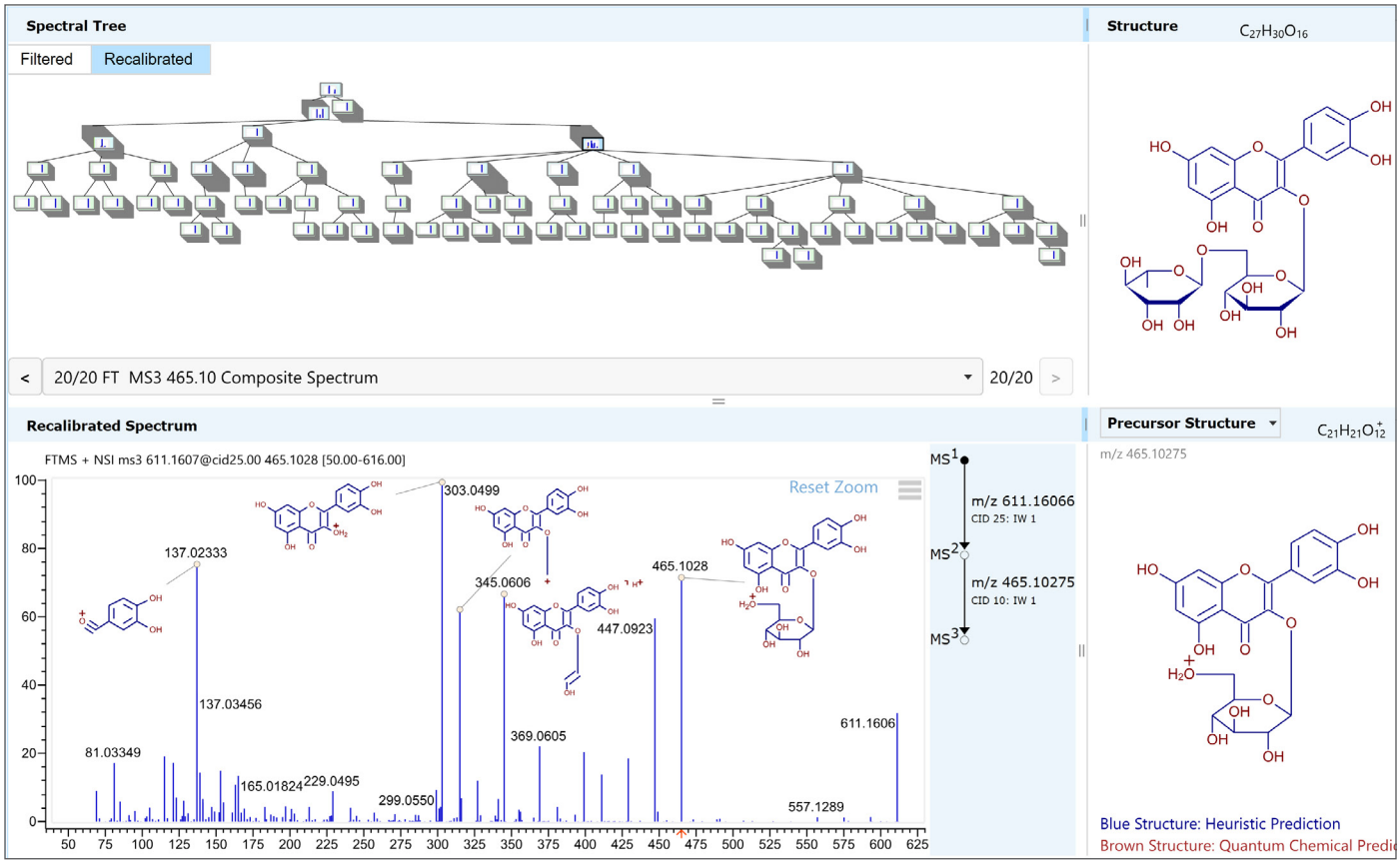


Figure 6. mzCloud reference library entry for a flavonoid—Rutin positive ion MS^n mass spectral tree.

Structure determination tools

Compounds with no library search results (unknown unknowns) will require further manual identification including additional targeted experiments such as obtaining chemical ionization in GC-MS to verify the molecular ion and elemental formula. Other experiments such as H/D exchange and chemical derivatization may need to be performed to help elucidate labile hydrogens and to characterize the presence of functional groups. Structure elucidation of unknowns typically also requires additional spectroscopic evidence such as FT-IR, 2D NMR and photodiode UV spectra.

Other tools available for LC-MS-based structure elucidation include obtaining MS^n mass spectral trees, in-silico fragmentation and precursor ion fingerprinting.[26] Additional LC-MS^n experiments are then performed to obtain a more complete mass spectral fingerprint of the unknown compound. By performing an mzCloud similarity search sub-structural elements of known structures can be compared to substructures present in an unknown compound. The precursor ion fingerprinting approach allows matching of unknown structures to related compounds and by the difference in molecular formulas may lead directly to a rational for proposing possible structures for the unknown.[13]

Once a hypothesis is formed for the unknown compound structure, advanced in-silico prediction may be used to rationalize the product ions expected in the proposed structure(s) and the observed mass spectral fragmentation tree. Mass Frontier software was developed specifically for predicting EI and MS^2 fragmentation via a complete rule set based on the published mass spectral literature.

Metabolite identification tools

Software workflows are an essential element for automation of structural characterization and elucidation. Although there are many commercial software applications and open access computational tools available, no single solution offers a complete workflow for structure elucidation. Flexible pipelines for automating data reduction offer the ability to use open access or customized in-house programs.[27] We focus here on the workflows developed to process high resolution Orbitrap metabolomics data utilizing TraceFinder software for high resolution GC-MS data analysis (Figure 5a) and Compound Discoverer software for high resolution LC-MS and MS^2 data analysis, identification of primary and secondary metabolites, statistical analysis and metabolic pathway mapping (Figure 5b). For lipid analysis using LC and data dependent MS^2, LipidSearch software is used for lipid annotation.

GC-MS

Since GC-MS library searching is an already well-established method for identification, most commercial software packages for GC-MS including TraceFinder software use a mass spectral deconvolution algorithm and library search software for automated identification. The excellent mass accuracy of the Q Exactive GC-MS system also enables a new high-resolution scoring algorithm (HRF) used together with the spectral library match factor to provide higher confidence in GC-MS library searches. A typical GC-MS extracted ion chromatogram (XIC) and EI mass spectrum for an unknown metabolite from human plasma is shown in Figure 7a. The mass spectrum was searched against a HRAM Metabolomics library and the compound with the best library match, Glutamic acid (3TMS) is shown in Figure 7b.

In addition to the library match score (SI = 914), the HRF score (99.7) ensures that the proposed compound structure must fit the elemental composition of the molecular ion and all of the fragment ions present in the EI mass spectrum. Combining the traditional spectral match and HRF score gives very high confidence in an identification using a GC-MS mass spectral library. Another measure of increased confidence for the Glutamic acid (3TMS) annotation is the retention time index (RI = 1613, calculated from the experimental retention time) vs. the library (RI = 1614).

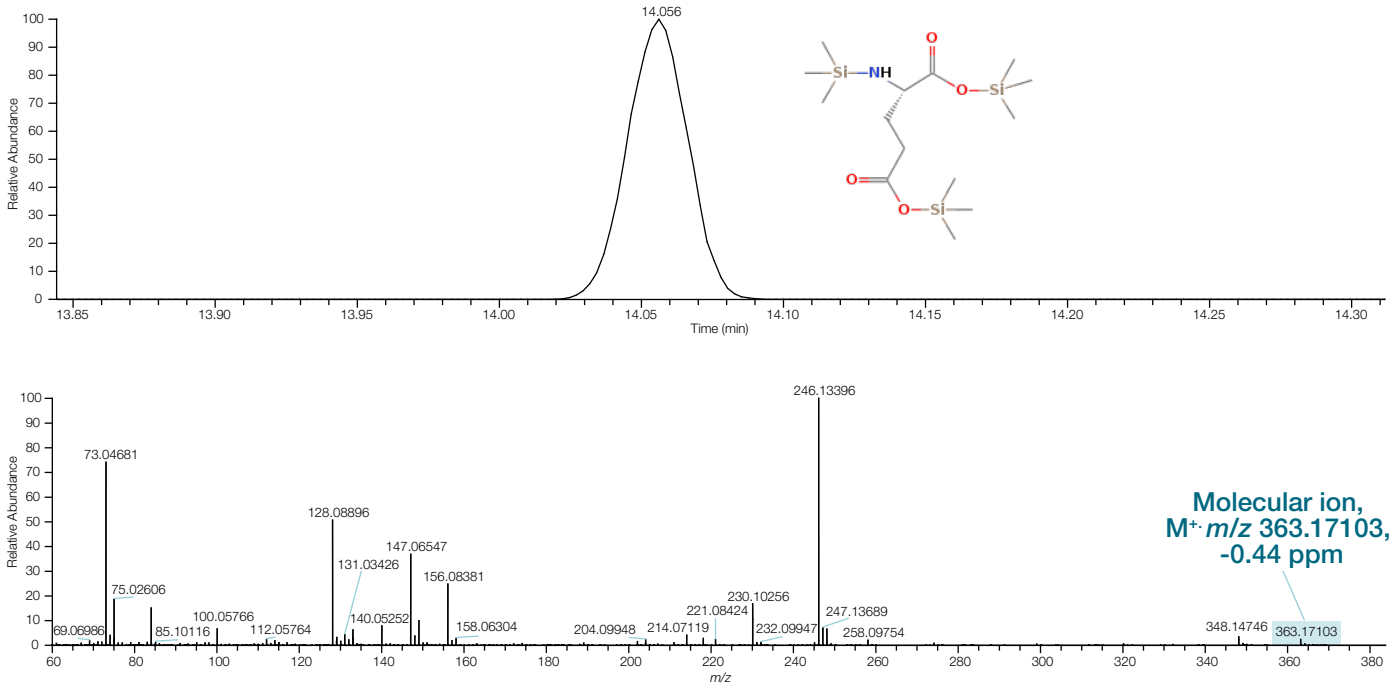


Figure 7a. High resolution Q Exactive GC-MS EI mass chromatogram (*m/z* 363.1710, $C_{14}H_{33}O_4NSi_3$) and mass spectrum at 14.056 minutes.

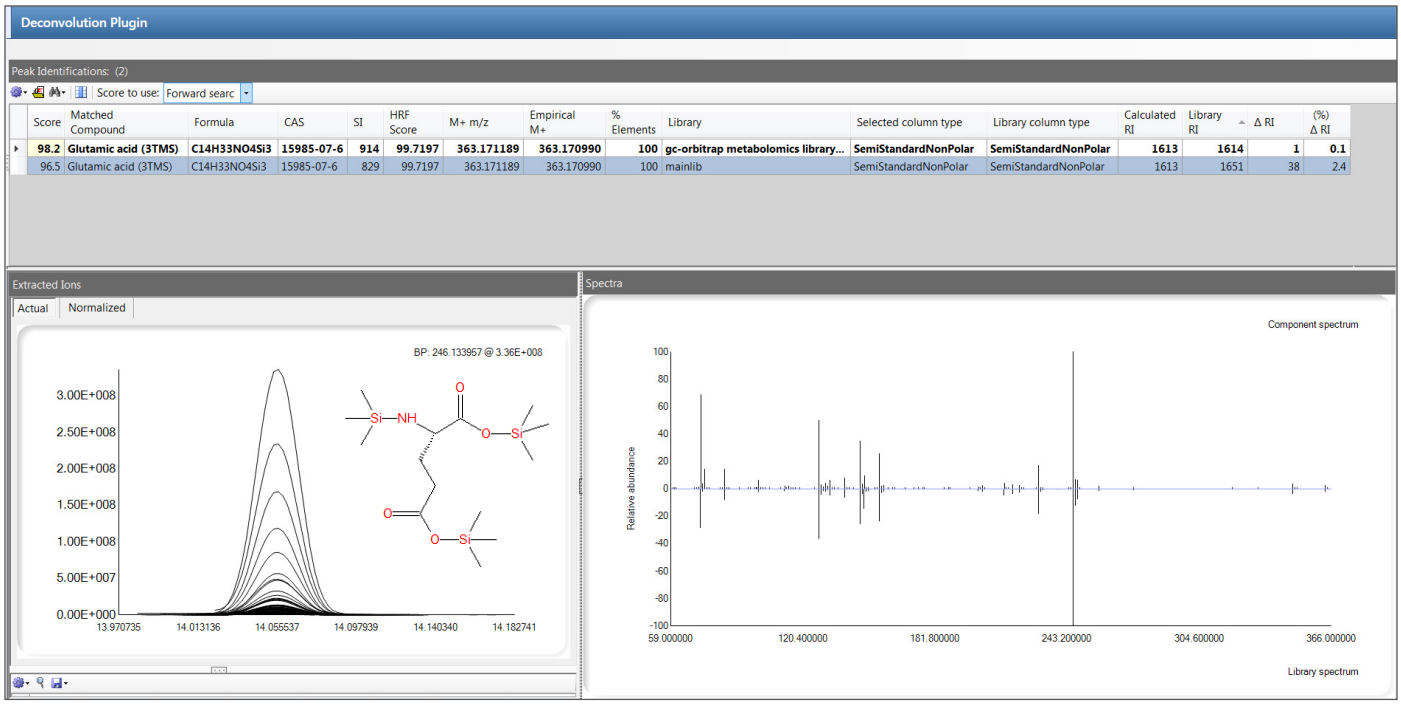


Figure 7b. High confidence GC-MS library match, high resolution formula score and retention time index match for Glutamic acid, 3TMS.

LC-MS

Compound Discoverer software is an integrated package for small molecule analysis that supports workflows including metabolism and metabolomics. The workflow is based on a node-based pipeline and provides for flexibility including node development by third parties and researchers who have developed their own informatics tools. Compound Discoverer software is a unique framework for further development of structure elucidation and other workflows for small molecule characterization. The nodes can be arranged according to a particular research need, and only nodes with the correct input and output may be connected together.

Identification in Compound Discoverer software is implemented as a series of logical data reduction and processing steps.

1. After feature grouping, elemental compositions are predicted using the measured accurate mass, isotopic pattern and isotopic fine structure—at very high resolution the individual isotopomers containing ^{13}C, ^{18}O, ^{15}N and ^{34}S are separated into their constituent peaks. In addition, MS^2 fragmentation data are also used to determine elemental formula.
2. In parallel, a ChemSpider database search is performed on the neutral monoisotopic mass, and elemental formulas determined for each compound group, and
3. An mzCloud MS^n library search is performed on the LC-MS^2 data.

Results from all of the data sources are analyzed and the compound name and formula are automatically assigned. The software reports how each compound was identified and the level of consensus between the annotation sources. When mass spectral library and database annotations are not found, similarity searches are performed looking for matches to fragment ions present in other molecules having different molecular mass from the unknown compound.

The high resolution (240,000 FWHM at *m/z* 200) LC-MS analysis is illustrated in Figure 8a for an unknown metabolite at *m/z* 205.0969 and retention of 4.75 min.

The measured mass of the $[M+H]^+$ ion, isotope pattern and isotopic fine structure provides unequivocal evidence of the molecular formula ($C_{11}H_{12}N_2O_2$). The green bars (Figure 8b) in the mass spectrum inset represent results matching the isotopic fine structure within the expected error in mass and intensity for the A+1 and A+2 isotopomers. A molecular formula search was performed of the ChemSpider database and 12 isomeric compounds (Figure 8c) were found fitting the molecular formula.

The mzCloud high resolution library search of the *m/z* 205.0969 MS^2 spectrum (Figure 9a) returned three Tryptophan isomers as the only hits with DL Tryptophan giving the best match score (96.8). The mirror plot (Figure 9b) shows the MS^2 spectrum from a single Orbitrap HCD scan matched to the mzCloud reference library. The three hits from mzCloud were found in the top four database hits (ranked by number of references). Thus, there is excellent evidence for annotation of Tryptophan and consensus from all of the sources including predicted elemental composition, isotopic fine structure, mzCloud library and ChemSpider database matches. The mass spectrometric evidence and library tools allows identification to a level of confidence as a probable structure (level 2 in Table 3). Additional spectroscopic evidence and/or comparison with a known reference standard are required for a confident structure assignment (level 1).

LC-MS^n

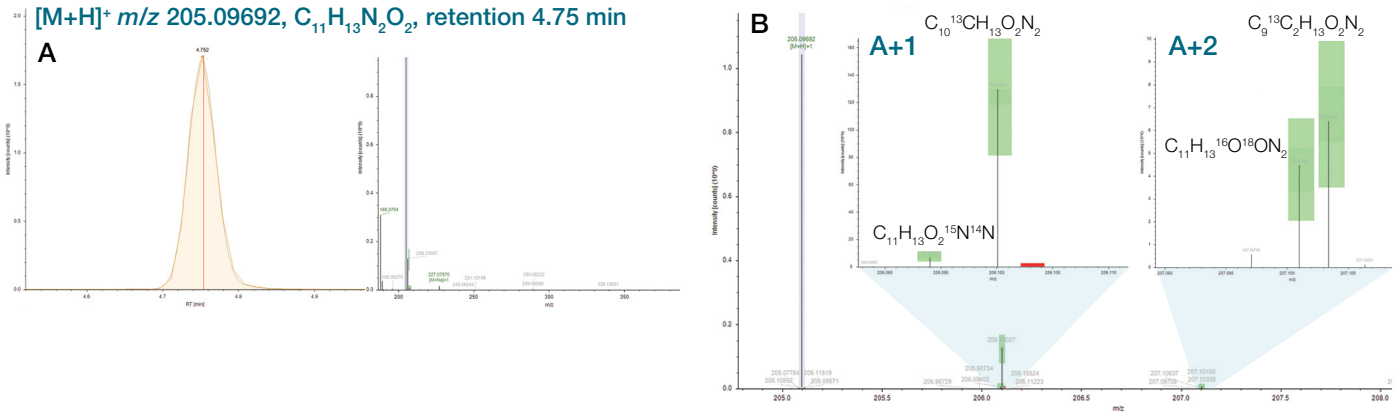
For lipid identification, a standard HRAM workflow is the LC-data dependent MS^2 approach along with LipidSearch software for structure annotation (Figure 5). This provides simultaneous untargeted profiling and identification for lipid from cells, plasma and tissues. The analysis of complex lipid extracts from insect larvae requires a more sophisticated approach to distinguish coeluting isomeric lipid species.[28] Targeted CID LC-MS^2 or LC-MS^3 experiments (Figure 10) are used to selectively characterize specific lipids during a data dependent LC-MS^2 run using the Thermo Scientific™ Orbitrap ID-X™ Tribrid™ mass spectrometer and LipidSearch software for lipid annotation.[29]

Figure 11 shows the MS^2 spectrum of a triglyceride (TG 48:1, *m/z* 822.7534) ammonium adduct found in corn rootworm larvae lipid extracts. Three product ions are observed corresponding to neutral loss of 18:1, 16:0, and 14:0 fatty acids. During a single scan cycle, the neutral losses of fatty acid were automatically detected and three additional CID MS^3 scans were performed. The MS^3 spectrum corresponding to loss of 18:1 fatty acid (Figure 11a) produces 14:0 and 16:0 acyl ions giving the assignment TG 18:1-14:0-16:0 (*isomer 1*). Similarly, the

MS^3 spectrum corresponding to 14:0 loss (Figure 11c) produces 16:0 and 18:1 acyl ions giving the same assignment (*isomer 1*). However, the MS^3 spectrum from 16:0 loss (Figure 11b) is a mixture of product ions consisting mainly of *isomer 1* (14:0 and 18:1 acyl ions) and a lesser amount of *isomer 2*, TG 16:0-16:1-16:0, giving rise to fatty acyl ions 16:0 and 16:1. This example illustrates the power of LC-MS^n for elucidating the structure of isomeric mixtures.

Conclusions

Small molecule characterization and unknown identification are greatly enhanced by the very high quality accurate mass MS and MS^n high resolution data obtained from modern Orbitrap mass spectrometers. Structural data obtained from either GC-MS or LC-MS Orbitrap instruments provides the highest probability of matching library spectra and confirming the presence of known compounds. The possibility of matching unknowns with sub-structural features of a known compound or classes is increasing as the software tools and MS^n libraries are expanding with available reference compounds. The level of compound annotation, first generated automatically by software and then confirmed by further expert analysis, needs to be reported in a manner consistent with the guidelines proposed by the Metabolomics Society and other regulatory agencies. For truly novel compound identification, the tools for obtaining structural characterization by mass spectrometry need to be coupled with chemical synthesis and other spectroscopic methods such as IR, UV and NMR for unambiguous identification.



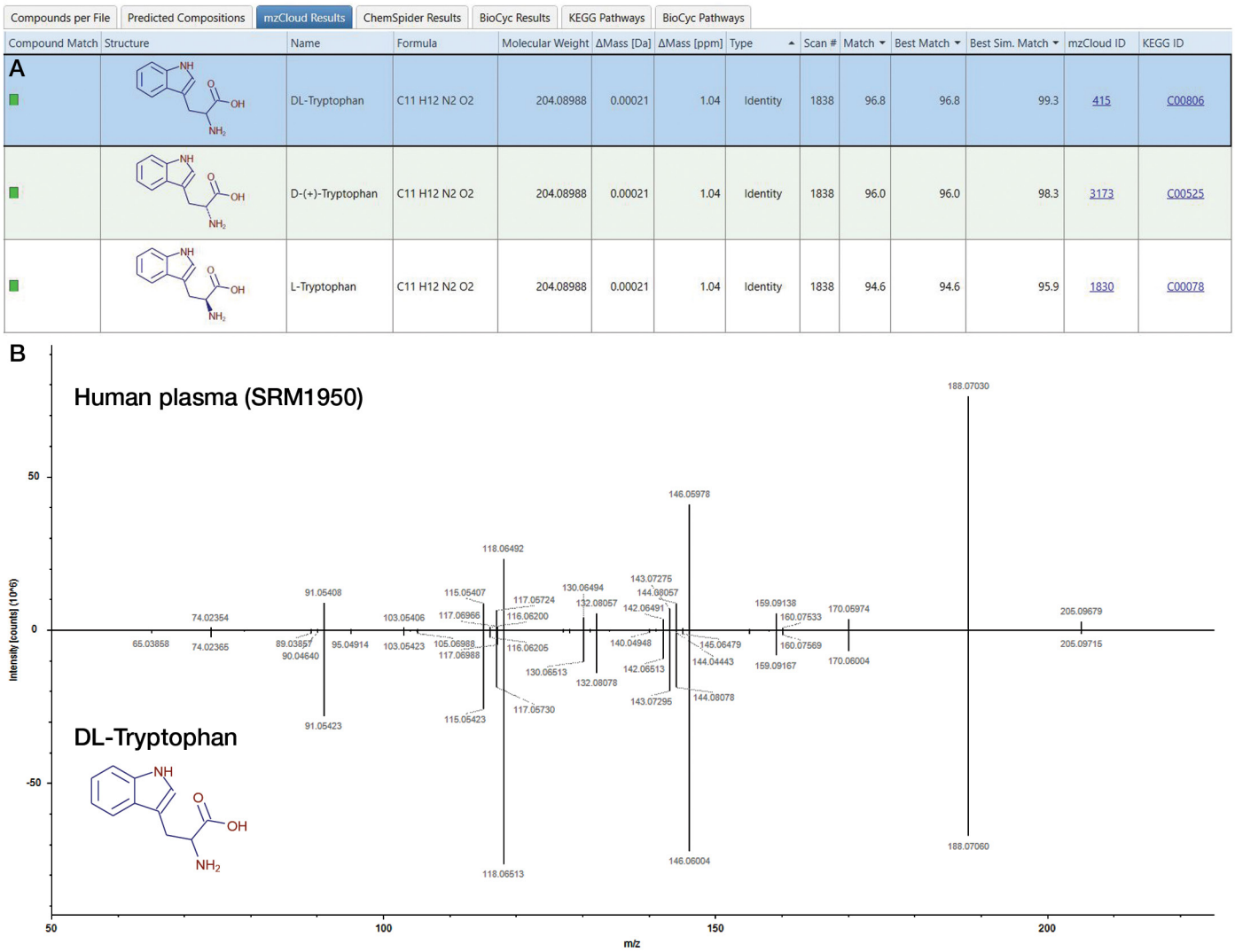
Compounds per File | Predicted Compositions | mzCloud Results | ChemSpider Results | BioCyc Results | KEGG Pathways | BioCyc Pathways

Compound Match	Formula	Molecular Weight	ΔMass [Da]	ΔMass [ppm]	RDBE	H/C	Rank	# Matched Iso.	# Missed Iso.	# Matched Frag.	SFit [%]	Pattern Cov. [%]	MS Cov. [%]	MSMS Cov. [%]
	C11 H12 N2 O2	204.08988	-0.00021	-1.04	7.0	1.1	1	5	1	38	57	99.87	99.95	99.61

Compounds per File | Predicted Compositions | mzCloud Results | ChemSpider Results | BioCyc Results | KEGG Pathways | BioCyc Pathways

Compound Match	Structure	Name	Formula	Molecular Weight	ΔMass [Da]	ΔMass [ppm]	CSID	mzLogic Score	# References
C		DL-Tryptophan	C11 H12 N2 O2	204.08987	0.00021	1.02	1116	280	22239
		L-Tryptophan	C11 H12 N2 O2	204.08987	0.00021	1.02	6066	280	743
		D-(+)-tryptophan	C11 H12 N2 O2	204.08987	0.00021	1.02	8707	280	205
		Nirvanol	C11 H12 N2 O2	204.08987	0.00021	1.02	82605	157	151

Figure 8. Confident assignment of elemental composition using accurate mass, isotope fine structure and ChemSpider database match from Compound Discoverer software.



Compound Match	Structure	Name	Formula	Molecular Weight	ΔMass [Da]	ΔMass [ppm]	Type	Scan #	Match	Best Match	Best Sim. Match	mzCloud ID	KEGG ID
A		DL-Tryptophan	C11 H12 N2 O2	204.08988	0.00021	1.04	Identity	1838	96.8	96.8	99.3	415	C00806
		D-(+)-Tryptophan	C11 H12 N2 O2	204.08988	0.00021	1.04	Identity	1838	96.0	96.0	98.3	3173	C00525
		L-Tryptophan	C11 H12 N2 O2	204.08988	0.00021	1.04	Identity	1838	94.6	94.6	95.9	1830	C00078

Figure 9. High confidence LC-MS² library match of Tryptophan from mzCloud and displayed in Compound Discoverer software.

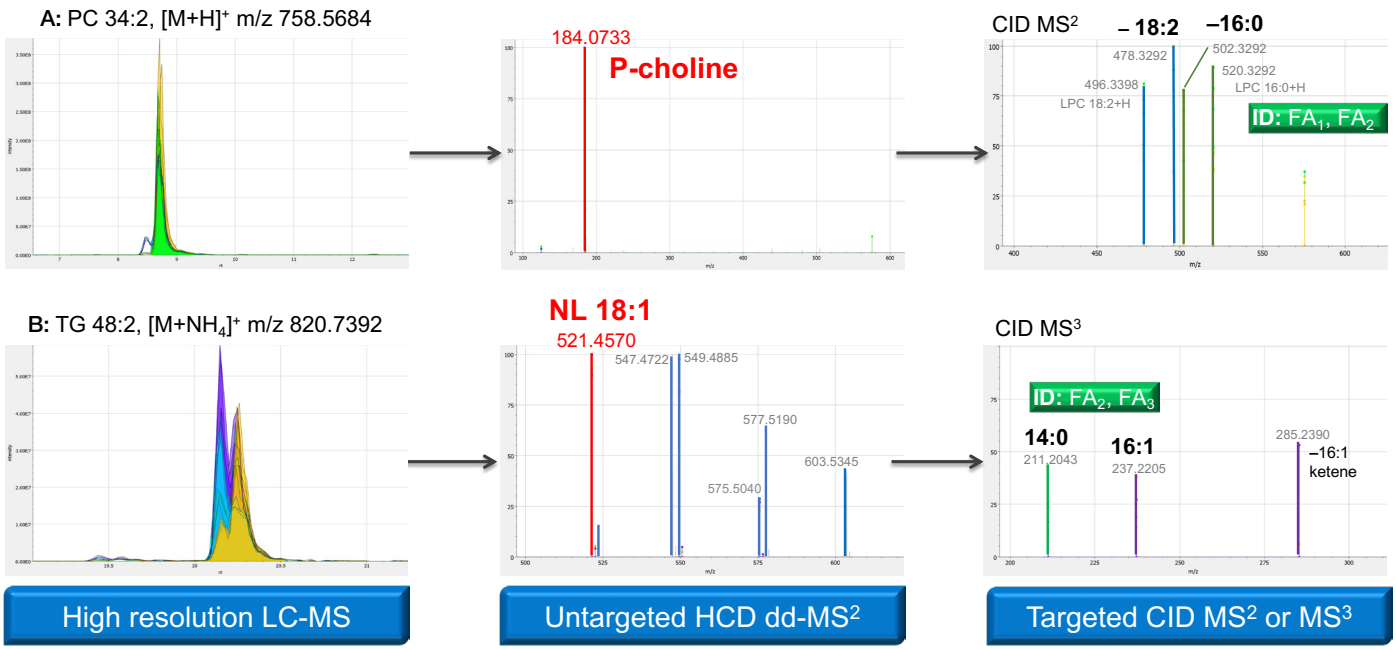


Figure 10. Combined data-dependent LC-MS² and class-specific targeted MS²/MS³ workflows for more confident lipid characterization. a) PC 34:2: HCD MS² of 758.5684 gives *m/z* 184.0733 product ion but, no fatty acyl information; *m/z* 184 targeted CID MS² provides identification of PC 16:0_18:2. b) TG 48:2: HCD MS² of 820.7392 gives seven different fatty acid neutral losses; loss of 18:1 (*m/z* 521) targeted CID MS³ provides positive identification of TG 14:0_18:1_16:1.

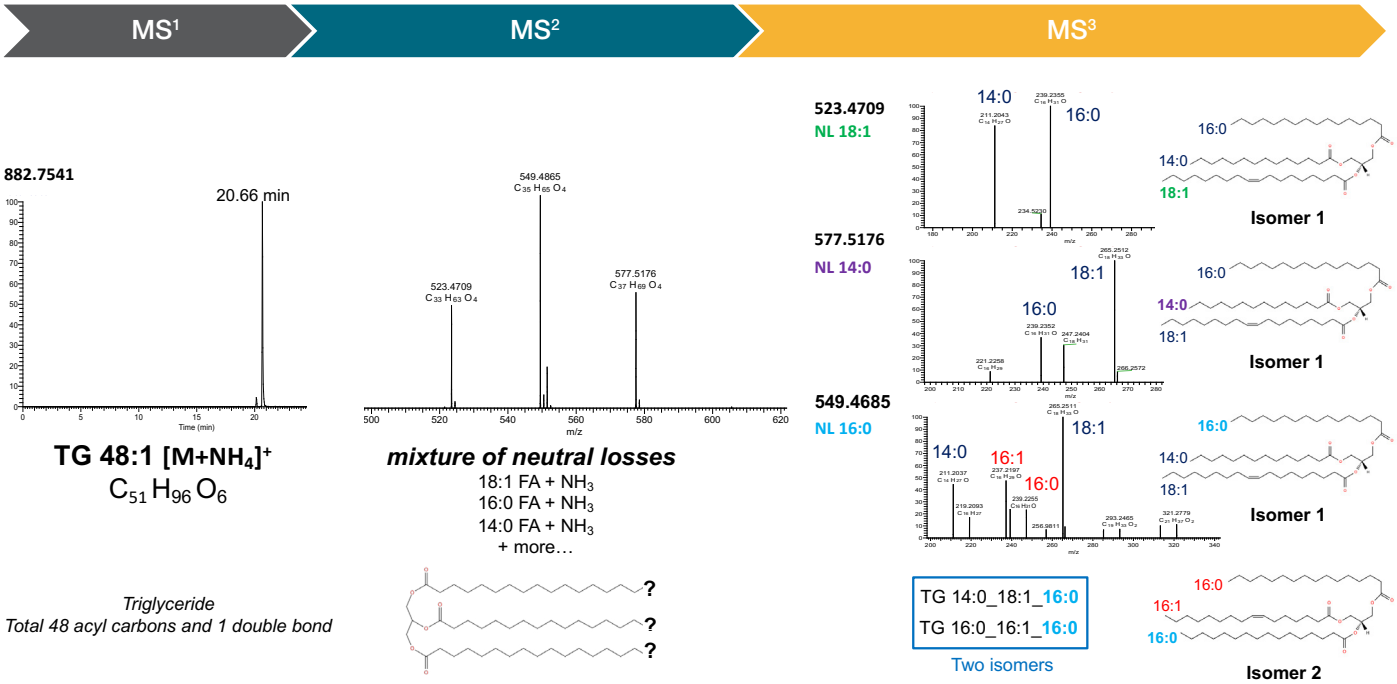


Figure 11. LC/MS^3 identification of a mixture of two TG 48:1 isomers from western corn rootworm larvae: (isomer 1) 14:0-18:1-16:0 and (isomer 2) 16:0-16:1-16:0. The TG 48:1 precursor at *m/z* 882 and 20.66 min gives a mixture of neutral losses of fatty acids and ammonia. The three main losses (*m/z* 523, 577 and 549) are automatically targeted for CID MS3; NL of 18:1, 14:0 and 16:0 generate annotations for isomer 1, TG 14:0_18:1_16:0; NL of 16:0 also gives isomer 2, TG 16:0_16:1_16:0.

Definitions and abbreviations used in this paper

Item	Description
Annotation	Compound identifier based on level of information available
APCI	Atmospheric pressure chemical ionization creates ions via a corona discharge
CAS	Chemical Abstracts Service database of compounds reported in the literature
CD	Compound Discoverer software for metabolomics
CE	Capillary electrophoresis; method for electro-kinetic separations
CID	Collision induced dissociation (stepwise collisions in a linear ion trap)
COSY	2D NMR homonuclear correlation spectroscopy
Database	Compound database including structure, formula, accurate mass and product ions
EI	Electron Ionization ion source generates 70-eV electrons typically used in GC-MS
Elucidation	Process of characterizing synthetic compounds by spectroscopic and MS methods
ESI	Electrospray ionization creates ions for sampling by atmospheric pressure interfaces
Feature	An ion with a measured mass-to-charge (*m/z*) and retention time (Rt)
FIA	Flow injection analysis, method of sample introduction without a LC column
FiehnLib	GC-MS metabolite library from GC/TOF MS at UC Davis (Oliver Fiehn)
FoodDB	Food metabolite database
FT-IR	Fourier transform infrared spectroscopy
FT-NMR	Fourier transform nuclear magnetic resonance spectroscopy
GC-MS	Gas chromatography mass spectrometry
HILIC	Hydrophilic interaction liquid chromatography; method for polar analytes
HMDB	Human metabolome metabolite database from the Wishart lab (www.hmdb.ca)
IC/MS	Ion chromatography mass spectrometry; method of separating anions or cations
Identification	Annotation structure proposal confirmed by comparison with a reference standard
Infusion	Continual sample introduction by a syringe or nano-infusion chip device
HCD	Higher-energy collisional dissociation (refers to collision in a high-pressure gas cell)
HSQC	2D NMR heteronuclear single quantum correlation experiment
LC-MS	Liquid chromatography mass spectrometry
MassBank	Mass spectral data repository (www.massbank.jp)
METLIN	Metabolite MS^2 database from QTOF instruments (metlin.scripps.edu)
NIST/EPA/NIH	National Institute of Science and Technology, Environmental Protection Agency National Inst. of Health mass spectral library (v. 17) with search program (v. 2.3)
Library	A compendium of reference mass spectra acquired under specific conditions
Mass Frontier	Structure elucidation software (HighChem)
MS/MS	Tandem mass spectrometry (MS^2) obtained with precursor mass selection
MS^n	Multiple steps of CID in a linear ion trap or other device suitable for MS^n
mzCloud	Advanced high-resolution MS^n mass spectral library (www.mzcloud.org)
Orbitrap	A new type of FT mass spectrometer introduced in 2005
PIF	Precursor ion fingerprinting; method for metabolite identification
PubChem	NIH sponsored database of published compounds (pubchem.ncbi.nlm.nih.gov)
Resolution	Mass resolution (m/Δm) defined at peak full-width-half-maximum (FWHM)
TOCSY	2D NMR total correlation spectroscopy experiment
TOF	Time-of-flight mass spectrometer
UVPD	Ultra-violet photodissociation
UV-VIS	Ultra-violet and visible spectrophotometry

thermo scientific

References

1. S G Oliver, M K Winson, D B Kell and R Baganz, **Systematic functional analysis of the yeast genome**. *Trends Biotechnol.* **1998**, 16, 373–378.
2. Fiehn O. **(2002) Metabolomics — the link between genotypes and phenotypes**. In: Town C. (ed.) Functional Genomics. Springer, Dordrecht, 155–171.
3. A Soriano Lopes, E Castañeda Santa Cruz, A Sussulini and A Klassen, **Chapter 4: Metabolomic Strategies Involving Mass Spectrometry Combined with Liquid and Gas Chromatography** in *Metabolomics: From Fundamentals to Clinical Applications, Advances in Experimental Medicine and Biology*, A Sussulini, ed., Springer International Publishing AG, **2017**, p. 77–98.
4. W Lu, X Su, MS Klein, IA Lewis, O Fiehn and JD Rabinowitz, **Metabolite Measurement: Pitfalls to Avoid and Practices to Follow**, *Annu Rev Biochem* **2017**, 86: 277–304.
5. J-L Wolfender, G Marti, A Thomas and S Bertran **Review: Current approaches and challenges for the metabolite profiling of complex natural extracts**, *J Chrom A* **2015**, 1382, 136–164.
6. NG Mahieu and GJ Patti "**Systems-Level Annotation of a Metabolomics Data Set Reduces 25,000 Features to Fewer than 1,000 Unique Metabolites**" *Anal Chem*, **2017**, 89, 10397–10406.
7. T Kind and O Fiehn, **Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm**, *BMC Bioinformatics* **2006**, 7: 234.
8. T Kind and O Fiehn, **Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry**, *BMC Bioinformatics* **2007**, 8: 105.
9. M Brown, WB Dunn, P Dobson, Y Patel, CL Winder, S Francis-McIntyre, P Begley, K Carroll, D Broadhurst, A Tseng, N Swainston, I Spasic, R Goodacre and DB Kell, **Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics**, *Analyst* **2009**, 134, 1322–1332.
10. O Fiehn, D Robertson, J Griffin, M van der Werf, B Nikolau, N Morrison, LW Sumner, R Goodacre, NW Hardy, C Taylor, J Fostel, B Kristal, R Kaddurah-Daouk, P Mendes, B van Ommen, JC Lindon and S-A Sansone, **The metabolomics standards initiative (MSI)**, *Metabolomics* **2007**, 3, 175–178.
11. LW Sumner, A Amberg, D Barrett, MH Beale, R Beger, CA Daykin, T W-M Fan, O Fiehn, R Goodacre, JL Griffin, T Hankemeier, N Hardy, J Harnly, R Higashi, J Kopka, AN Lane, JC Lindon, P Marriott, AW Nicholls, MD Reily, JJ Thaden and MR Viant, **Proposed minimum reporting standards for chemical analysis**, *Metabolomics* **2007**, 3, 231–241.
12. B Rochat, **Proposed Confidence Scale and ID Score in the Identification of Known-Unknown Compounds Using High Resolution MS Data**, *J Am Soc Mass Spectrom* **2017**, 28, 709–723.
13. M Ghaste, R Mistrik and V Shulaev, "**Applications of Fourier Transform Ion Cyclotron Resonance (FT-ICR) and Orbitrap Based High Resolution Mass Spectrometry in Metabolomics and Lipidomics**" *Int J Mol Sci* **2016**, 17, 816.
14. Q Hu, RJ Noll, H Li, A Makarov, M Hardman and RG Cooks, **The Orbitrap: a new mass spectrometer**, *J Mass Spectrom.* **2005**, 40: 430–443.
15. A Makarov, E Denisov, A Kholomeev, W Balschun, O Lange, K Strupat and S Horning, **Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer**, *Anal Chem* **2006**, 78, 2113–2120.
16. A Michalski, E Damoc, J-P Hauschild, O Lange, A Wieghaus, A Makarov, N Nagaraj, J Cox, M Mann, and S Horning, **Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer**, *Mol Cell Proteomics* **2011**, 10(9): M111.011015.
17. A Makarov, E Denisov and O Lange, **Performance Evaluation of a High-field Orbitrap Mass Analyzer**, *J Am Soc Mass Spectrom* **2009**, 20, 1391–1396.
18. AC Peterson, J-P Hauschild, ST Quarmby, D Krumwiede, O Lange, RAS Lemke, F Grosse-Coosmann, S Horning, TJ Donohue, MS Westphall, JJ Coon and J Griep-Raming, **Development of a GC/Quadrupole-Orbitrap Mass Spectrometer, Part I: Design and Characterization**, *Anal Chem* **2014**, 86, 10036–10043.
19. AC Peterson, AJ Balloon, MS Westphall and JJ Coon, **Development of a GC/Quadrupole-Orbitrap Mass Spectrometer, Part II: New Approaches for Discovery Metabolomics**, *Anal Chem* **2014**, 86, 10044–10051.
20. S Eliuk and A Makarov, **Evolution of Orbitrap Mass Spectrometry Instrumentation**, *Annu Rev Anal Chem* **2015**, 8: 61–80.
21. A Makarov, E Denisov, O Lange and S Horning, **Dynamic Range of Mass Accuracy in LTQ Orbitrap Hybrid Mass Spectrometer**, *J Am Soc Mass Spectrom* **2006**, 17, 977–982.
22. M Scigelova, M Hornshaw, A Giannakopulos and Alexander Makarov, **Fourier Transform Mass Spectrometry**, *Mol Cell Proteomics*, **2011**, 10(7): M111.009431.
23. IV Chernushevich, AV Loboda and BA Thomson, **An introduction to quadrupole–time-of-flight mass Spectrometry**, *J Mass Spectrom* **2001**, 36: 849–865.
24. E Denisov, E Damoc, O Lange and A Makarov, **Orbitrap mass spectrometry with resolving powers above 1,000,000**, *Int J Mass Spectrom* **2012**, 325:327, 80-85.
25. PT Kasper, M Rojas-Chertó, R Mistrik, T Reijmers, T Hankemeier and RJ Vreeken, **Fragmentation trees for the structural characterization of metabolites**, *Rapid Commun Mass Spectrom* **2012**, 26, 2275–2286.
26. MT Sheldon, R Mistrik and TR Croley "**Determination of Ion Structures in Structurally Related Compounds Using Precursor Ion Fingerprinting**" *J Am Soc Mass Spectrom*, **2009**, 20, 370–376.
27. RJM Weber, TN Lawson, RM Salek, TMD Ebbels, RC Glen, R Goodacre, JL Griffin, K Haug, A Koulman, P Moreno, M Ralser, C Steinbeck, WB Dunn, MR Viant, **Computational tools and workflows in metabolomics: An international survey highlights the opportunity for harmonisation through Galaxy**, *Metabolomics* **2017**, 13, 12.
28. D Gachotte, Y Adelfinskaya, J Gilbert, R Kiyonami, D Peake and Y Yokoi, **Increased Depth and Confidence of Lipidome Analysis from Insect Tissues using Chromatography Based Methods with High-resolution Orbitrap MS^n**, Proceedings of the 66th ASMS Conference on Mass Spectrometry and Allied Topics, San Diego, California, June 3–7, **2018**.
29. DA Peake, R Kiyonami, D Gachotte, GE Reid, Y Yokoi and A Hühmer, **Software Utilizing Positive and Negative Ion MS^2/MS^3 HCD and CID Spectra for Improved MS^n Lipid Identification**, Proceedings of the 66th ASMS Conference on Mass Spectrometry and Allied Topics, San Diego, California, June 3–7, **2018**.

Find out more at **thermofisher.com/metabolomics**

ThermoFisher SCIENTIFIC

WP65356-EN 0918M