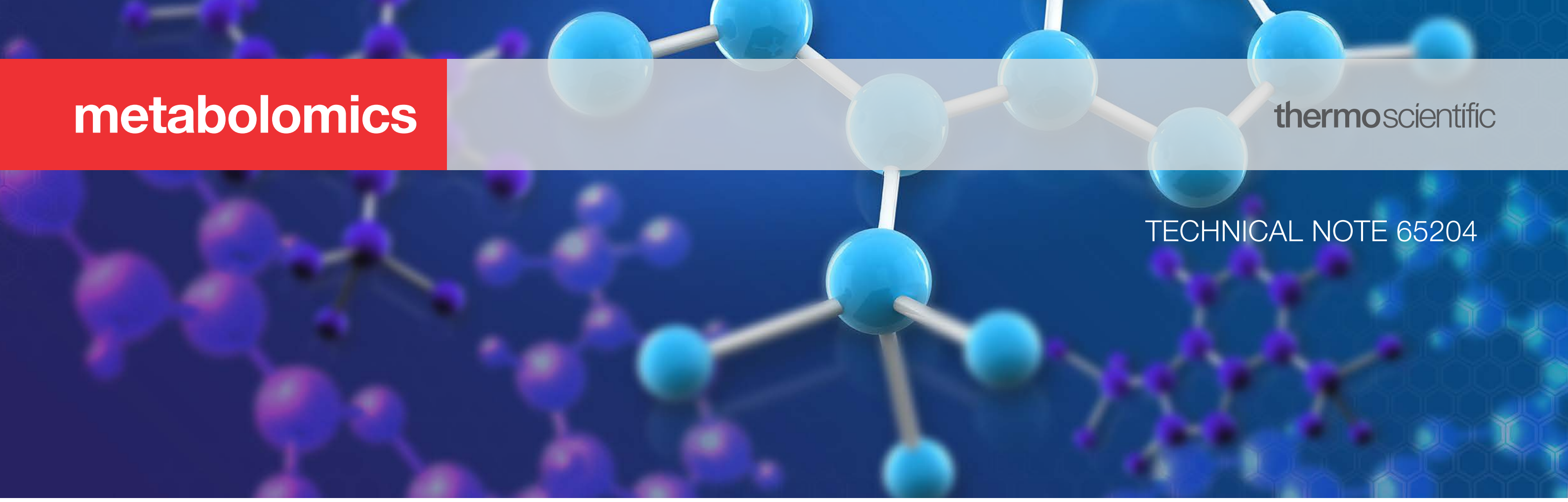


metabolomics

TECHNICAL NOTE 65204

Features or compounds? A data reduction strategy for untargeted metabolomics to generate meaningful data

Authors

Amanda Souza and Ralf Tautenhahn

Thermo Fisher Scientific

Keywords

Metabolomics, untargeted metabolomics, features, compounds, data reduction, Compound Discoverer software

Introduction

Untargeted metabolomics describes an unbiased approach that globally analyzes the metabolome in any given sample. Comprehensive metabolome coverage is obtained using analytical techniques such as liquid chromatography – mass spectrometry (LC-MS) by collecting all measurements indiscriminately. While a defined mass range is normally implemented with this approach, there is a fundamental notion to “collect everything and leave nothing behind.” At the same time, unbiased data collection results in an exhaustive list of spectral features or signals, making data analysis laborious and cumbersome. Through a data reduction process these features can be converted to a list of meaningful compounds by accounting for artifacts such as naturally occurring isotopes and chemical interactions such as adduct formations with metal ions. Likewise, background features unrelated to the experimental samples such as chemical noise, contaminants, or analytes resulting from sample handling can be eliminated from the working list. Furthermore, neglecting artifacts may lead to over interpretation of data, thus drawing incorrect conclusions and wasting time. Here we present how to differentiate a compound from a molecular feature using Thermo Scientific™ Compound Discoverer™ software to reduce redundancies and accelerate data analysis.

Unbiased peak detection with high resolution accurate mass (HRAM) LC-MS

Untargeted metabolomics aims to capture comprehensive coverage of the metabolome utilizing an all-encompassing approach. The application can be applied to a variety of samples including biofluids, animal or plant tissue, and intracellular components. These samples of biological origin are commonly regarded as complex matrices due to the presence of numerous and diverse endogenous components such as biogenic amines, amino acids, organic acids, and lipid species. These complex mixtures benefit from separation techniques such as liquid chromatography to resolve components in the time dimension based on compound polarity. Indeed the data generated is highly dense and complex in nature. Mass spectrometers measure the mass-to-charge ratio (*m/z*) of chemical species. Combined with signal intensity, a three dimensional array of components representing molecules from the extracted complex matrix is generated using unbiased detection, thus producing an untargeted output (Figure 1).

HRAM mass spectrometers operated in full scan mode detect all ions in a complex matrix where spectral resolution pertains to the spectrometer's ability to differentiate adjacent mass peaks in the spectrum. High resolution mass spectrometers such as those using Thermo Scientific™ Orbitrap™ technology offer ultrahigh resolving power, up to 500,000 FWHM, and recently available is the option for one million FWHM at *m/z* 200. In most cases, data acquisition of untargeted metabolomics applications using Orbitrap mass spectrometers generally start with full scan mode analysis using resolving power from 60,000 to 140,000 FWHM at *m/z* 200 depending on the instrument. A wide mass range, for instance 70–800 *m/z*, is interrogated and allows for quantification of precursor (parent) ions in either positive or negative ion polarity. One analyte can give rise to multiple adducts, fragments, and even cluster ions in addition to their respective isotopic peaks. As a result, detection of thousands of features, also termed signals or peaks, occurs when analyzing a wide mass range in full scan mode. This list of thousands of features in untargeted analysis can be misleading because not all features independently represent a singular biomolecular, endogenous metabolite.

Accounting for naturally occurring isotopes

The small molecules of interest in metabolomics studies are predominantly composed of chemical elements like carbon, hydrogen, nitrogen, sulfur, and oxygen. For example, the endogenous small molecule creatine

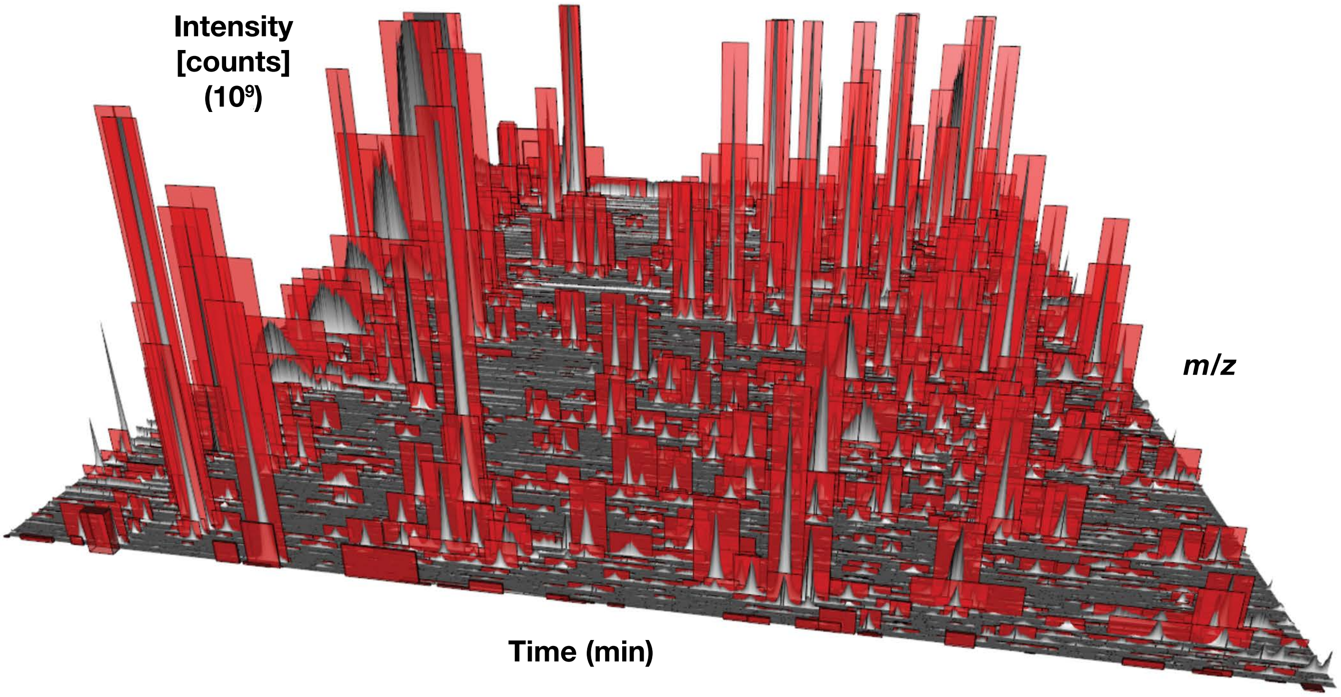


Figure 1. Three dimensional array of spectral features using LC-MS for untargeted metabolomics analysis. The first dimension, X axis, represents chromatography separation by time. The second dimension, Z axis, represents mass detection based on a mass-to-charge ratio (*m/z*). The final dimension is signal intensity or the response of the measurement based on ion current.

has an elemental composition of $C_4H_9N_3O_2$. Further, these chemical elements contain naturally occurring stable isotopes. For instance, ^{14}N and ^{15}N are naturally abundant isotopes of nitrogen. So for compounds observed in a mass spectrum, there are multiple isotopic peaks representing different naturally occurring isotopes. For instance, carbon has one isotope, ^{13}C, of detectable abundance with a mass difference of 1.0034 Daltons from the monoisotopic peak, ^{12}C. Undoubtedly, these two spectral peaks for any given compound are baseline resolved with high resolution mass spectrometers. Yet, other elements have isotopes of one mass unit from the monoisotopic peak, thus creating an A1 cluster of detectable ions.[1] Consider the amino acid methionine with an elemental composition of $C_5H_{11}NO_2S$. The theoretical *m/z* of the monoisotopic peak in positive ion polarity is 150.05833 *m/z*. The calculated *m/z* for the ^{13}C and ^{15}N isotope peaks of methionine is 151.06168 and 151.05536, respectively. Thus for small molecules containing nitrogen along with carbon, the ^{15}N isotope peak, with a mass difference of 0.997 Daltons from the monoisotopic peak, can be resolved from the ^{13}C isotope peak, a mere spectral distance of 0.0063 Daltons. Higher resolution settings can sufficiently resolve these closely related ions within an isotope cluster as shown with methionine in a metabolomics extract from the human plasma reference material, NIST 1950 (Figure 2).

Resolution can further be extended to the A2 cluster or elements possessing an isotope differing by two mass units from the monoisotopic peak. Ultimately, higher resolving power reveals a fine isotope structure of detected peaks where several spectral peaks represent the same molecule.

Confidence in spectral peak assignments is generated by high mass accuracy. Orbitrap mass spectrometers are capable of obtaining less than three parts per million (ppm) mass accuracy with an external calibration and less than one ppm mass accuracy with internal calibration. Exact mass measurements are then compared to theoretical or expected masses based on elemental composition as described above. With high mass accuracy, low mass tolerance criteria can be applied when determining spectral peaks. Lower tolerances, in turn, allow for higher confidence in mass assignments. Correct mass assignments are critical in unknown metabolomics analysis given the exhaustive list of unknown spectral peaks present.

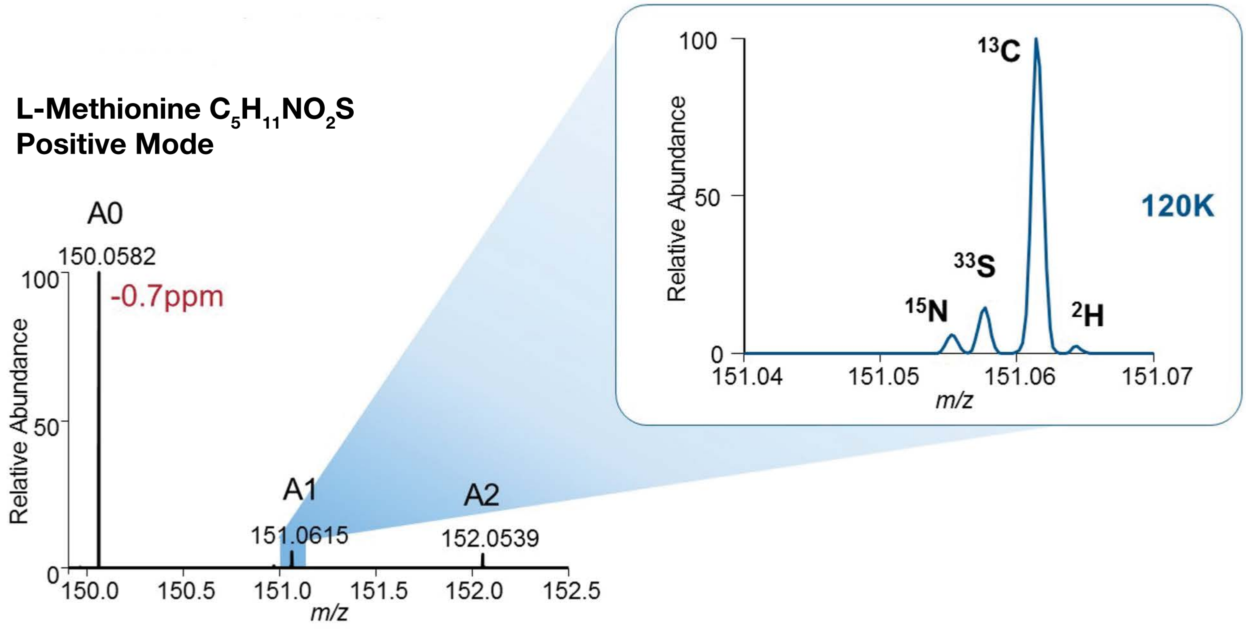


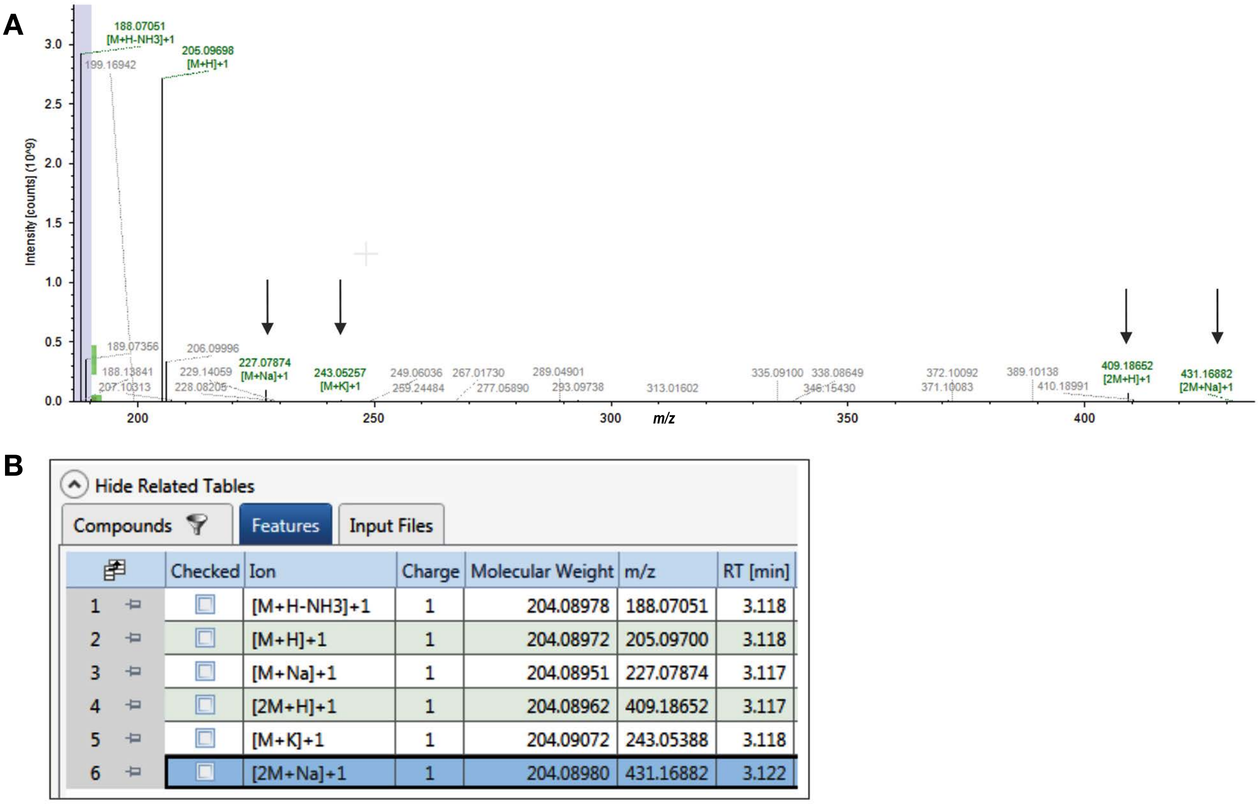
Figure 2. Example of a fine isotopic distribution for the A1 ion cluster of methionine. Data collection was on an organic extraction of human plasma reference material, NIST SRM 1950, using a Thermo Scientific™ Q Exactive™ HF MS with a resolution setting of 120,000 FWHM @ 200 *m/z*. Elemental composition for the amino acid is $C_5H_{12}N_2O_2$. Multiple isotope peaks are resolved in the A1 cluster.

Multiple ion species for one molecule in electrospray ionization

Electrospray ionization (ESI) is widely implemented in untargeted metabolomics LC-MS analysis. While the exact mechanism of how this ionization technique works have yet to be fully determined, the fundamental principle states that molecules in liquid phase are transferred to gas phase by applying an electric potential to the analyte solution to produce a fine mist of charged droplets subjected to solvent evaporation and subsequent ion ejection.[2] ESI is considered a soft ionization process because little to no fragmentation of the charged molecule occurs. Positively and negatively charged ions form in solution by the presence of acidic or basic functional groups on organic molecules and charge separation of inorganic species.[3] Ionization can be impacted by the choice of solvents, additives, and salt buffers for chromatographic mobile phases. Further, analytes with adduct formation to cationic (sodium, potassium, ammonium) and anionic species (chloride) may be observed.

Given the possibilities for chemical interactions, one molecule may be represented by several ionic forms.[4] When analyzing mass spectral data, the protonated or deprotonated form of the molecule is typically searched. However, it is possible and likely that other forms of the molecule could be present like those arising from adduct formation, dimerization, or interaction with molecules of the mobile phase. For example, the sodium adduct [M+Na] should be searched for data acquired in the positive ion polarity. In this simple case of considering sodium, two features reflect the same molecule. To add to this, these two monoisotopic features have related naturally occurring isotopes attributing to a greater number of features. Taken together, the potential that multiple features exist representing one molecule should not be underestimated. Figure 3 shows multiple features detected representing tryptophan in a metabolomics extract of rat plasma.

Failure to recognize and account for these chemical interactions is time consuming and can lead to misinterpretation of results. Analyzing data at the features level is a formidable and inefficient task given the labeled



Hide Related Tables

Compounds | Features | Input Files

	Checked	Ion	Charge	Molecular Weight	m/z	RT [min]
1	□	[M+H-NH3]+1	1	204.08978	188.07051	3.118
2	□	[M+H]+1	1	204.08972	205.09700	3.118
3	□	[M+Na]+1	1	204.08951	227.07874	3.117
4	□	[2M+H]+1	1	204.08962	409.18652	3.117
5	□	[M+K]+1	1	204.09072	243.05388	3.118
6	□	[2M+Na]+1	1	204.08980	431.16882	3.122

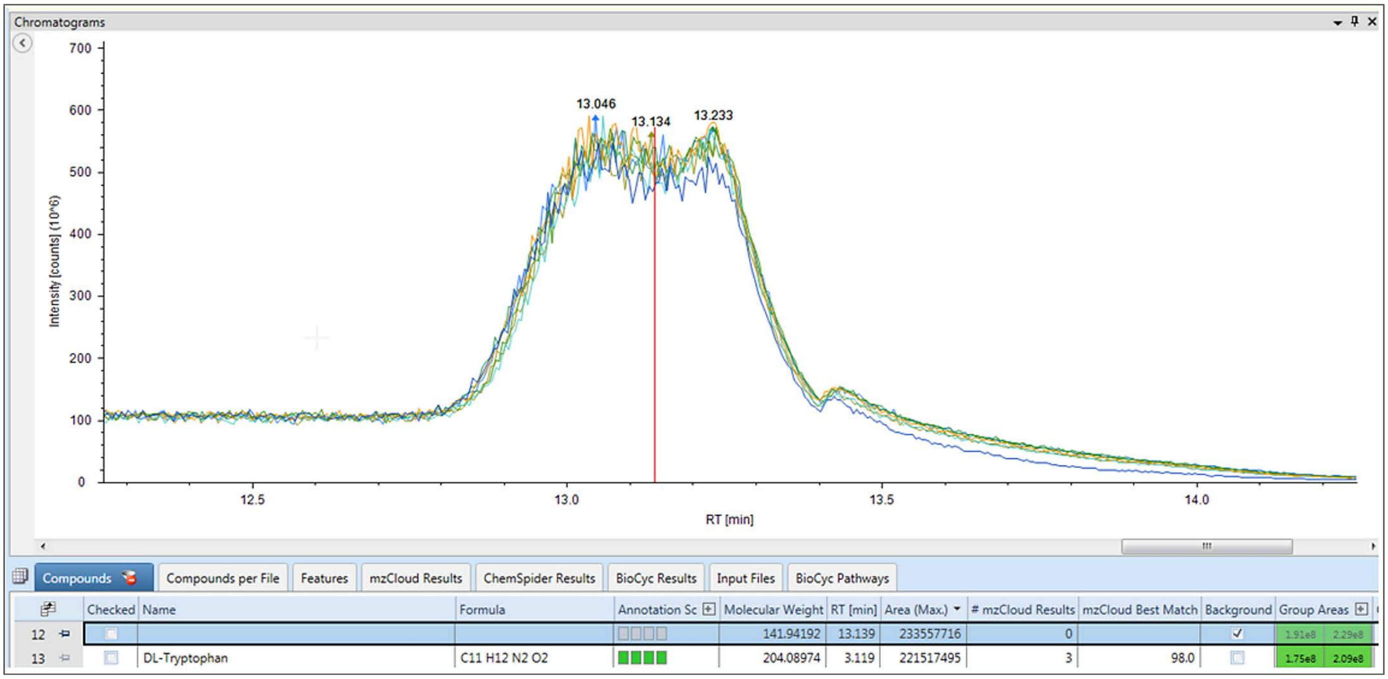
Figure 3. MS1 mass spectrum in positive ion polarity and related table from Compound Discoverer software for tryptophan. A) The mass spectrum shows the most intense peak highlighted in lavender while peaks marked with green are associated ions for tryptophan. B) Table displaying ion assembly list for tryptophan.

redundancies from isotopes and adducts already noted. Besides this, the risk of carrying out analyses on false positives is much greater. Consequently, with the recognition that untargeted metabolomics is typically processed for statistical significance by differential analysis, the statistics can simply be inflated leading researchers to incorrect conclusions. These pitfalls can be avoided when handling the data appropriately.

Background ions contribute to molecular features

Data output from an untargeted metabolomics LC-MS analysis is comprised of not only biologically relevant small molecules but also components unrelated to the experiment which are derived from external sources. The experimental sample, whether biofluid or tissue, goes through a series of processing steps before reaching the detector of the mass spectrometer. The sample preparation process often entails quenching and extracting target compounds with organic solvents and possibly reconstitution with chromatographic mobile phases all the while being subjected to consumables like microcentrifuge tubes and pipette tips. This process creates a favorable environment to attract unsuspecting molecules such as plasticizers like phthalates and polymers like polysiloxanes.[5] Additional sources of unrelated compounds include solvents not only for sample extraction but also chromatographic mobile phases. Further, molecules from within the laboratory environment may trickle in. Fortunately, these unrelated components can be controlled for during data processing with proper experimental design.

Prior to the start of the experimental analysis, a “blank” sample should be analyzed. In the context of untargeted metabolomics analysis, a blank sample can be either a solvent blank or an experimental blank. The solvent blank often mirrors the sample injection medium, which tends to be the starting chromatographic conditions. For instance, using reversed-phase chromatography, this will likely be a high percentage of water. With hydrophilic interaction chromatography, the blank will consist of mostly organic solvent compositions. As mentioned above, a sample is exposed to external factors like consumables. Same as the sample, the selected solvent can be prepared with the same procedure to mimic the sample extraction. This blank in turn is referred to as the experimental blank. Either way, the components originating from these samples can then be referenced to the true experimental samples to determine unrelated background molecules, also termed background noise (Figure 4). Incorporating a solvent blank or experimental blank enables determining spectral features originating from external sources.



	Checked	Name	Formula	Annotation Sc	Molecular Weight	RT [min]	Area (Max.)	# mzCloud Results	mzCloud Best Match	Background	Group Areas	
12	□			□□□□	141.94192	13.139	233557716	0		✓	1.91e8	2.29e8
13	□	DL-Tryptophan	C11 H12 N2 O2	🟩🟩🟩🟩	204.08974	3.119	221517495	3	98.0	□	1.75e8	2.09e8

Figure 4. Extracted ion chromatogram (XIC) overlay from Compound Discoverer software showing an unknown background compound detected in all samples. This unknown compound was consistently detected in the experimental samples as well as the solvent blank injection. The XICs for all samples are overlaid showing the same peak shape and intensity. Compounds of this nature are marked as such and can easily be hidden from the results list or recalled up when necessary. The red vertical line represents the apex for post chromatographic alignment.

Meaningful compounds extracted from molecular features using Compound Discoverer software

Distinguishing a feature from a compound using Compound Discoverer software provides perspective into the data reduction strategy in unknown analysis toward generating a working list of meaningful results. The discussion thus far focuses on molecular features and the origin of these features in order to appreciate how compound lists using Compound Discoverer software are generated. Specifically, a feature is an *m/z* at a given retention time. The term feature is synonymous with chromatographic peak or signal. As described above, multiple features can represent a single molecule via naturally occurring isotopes or chemical interactions. HRAM MS resolves closely related masses and even fine isotopic structure to efficiently detect these constituents. In contrast, a compound within Compound Discoverer software reflects a group of associated features. To this end, the feature assembly process in Compound Discoverer software was applied to a rat plasma extract analyzed in full scan at 70,000 resolving power, FWHM at 200 *m/z*, using a Thermo Scientific™ Q Exactive™ mass spectrometer. Using typical detection parameter settings for untargeted metabolomics studies, a total number of 54,000 features was reduced to 20,000 unique isotope groups, and further converted to 2,467 compounds.

Beyond detecting experimental artifacts like naturally occurring isotopes and ion formations by chemical interactions, Compound Discoverer software can mark background compounds from blank samples and further refine the list of compounds (Figure 5).

When analyzed using the same LC-MS conditions, the solvent or experimental blank contains ions unrelated to the experimental samples. Blank sample injections can be placed at the beginning of the injection sequence or at various points within throughout the analysis run. Multiple blank injections can be submitted to Compound Discoverer software for consideration. Based on a user-defined threshold of blank samples compared to experimental samples, background compounds are marked and can be hidden from the working results table. Accounting for unrelated background ions can eliminate unrelated molecules from the data output while substantially reducing the number of compounds for analysis.

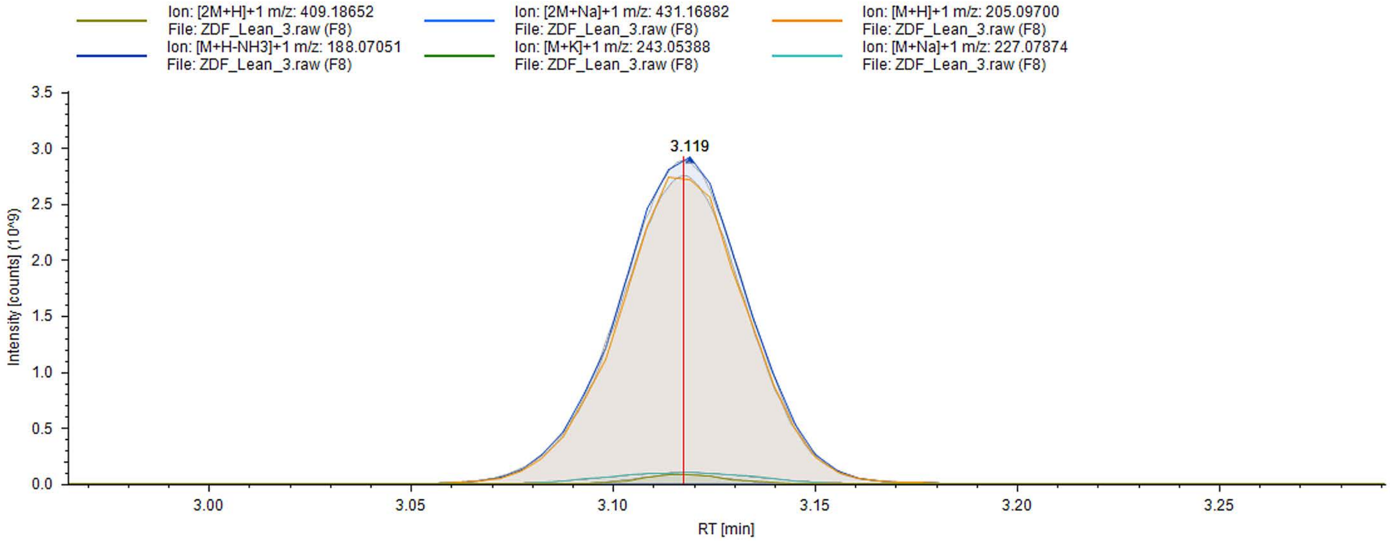


Figure 5. XIC overlay of tryptophan from Compound Discoverer software showing associated ion species. Several ions detected in positive polarity at the same retention time are assembled to reflect one compound within one sample injection of rat plasma extract. Associated ions include $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M+H-NH_3]^+$, $[2M+H]^+$, $[2M+Na]^+$.

Summary

HRAM LC-MS is a powerful tool for untargeted metabolomics analysis of complex biological matrices. High resolution full scan MS detects all ionic species present in analyzed samples generating an exhaustive results list of molecular features. This list of features is comprised of monoisotopic species, corresponding natural isotopes, associated adducts, and unsuspecting background ions due to the sample handling process and chromatographic mobile phase. Taking all of these experimental artifacts into consideration during data processing enables a proactive approach to eliminating redundancies and unrelated ions. Moving from features to compounds using Compound Discoverer software can provide a list of meaningful results for accelerated analysis.

References

1. McLafferty, F.W. (1980). *Interpretation of mass spectra, third edition.* University science books, Mill Valley, California.
2. Konermann, L., Ahadi, E., Rodriguez, A.D. and Vahidi, S. (2013). Unraveling the mechanisms of electrospray ionization. *Analytical Chemistry,* 85, 2–9.
3. Cech, N.B., and Enke, C.G. (2001). Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrometry Reviews,* 20, 362–387.
4. Mahieu, N.G. and Patti, G.J. (2017). Systems-level annotation of a metabolomics data set reduces 25,000 features to fewer than 1000 unique metabolites. A*nalytical Chemistry*, 89, 10397–10406.
5. Keller, B.O., Sui, J., Young, A.B., and Whittal, R.M. (2008). Interferences and contaminants encountered in modern mass spectrometry. *Analytica Chimica Acta,* 672, 1–81.

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