thermo scientific

WHITE PAPER 65374

Delivering confidence for small molecule identification

Author
Tim Stratton
Manager, Library Technologies
Thermo Fisher Scientific

Keywords
Small molecule identification, spectral library, database, mzCloud, Compound Discoverer, Mass Frontier, mzLogic, mzVault

Introduction: Addressing the challenges in small molecule identification with mass spectral libraries

Laboratories performing LC-MS applications strive for certainty when identifying detected compounds. For many targeted LC-MS methods, positive identifications made using retention time, accurate mass data, database searches of chemical formula, or selected reaction monitoring (SRM) experiments typically provide sufficient confidence in identifications. Reference standards are usually available for most targeted LC-MS analyses, meaning positive identification does not require spectral library matching.

However for many LC-MS applications, laboratories require additional information, such as mass spectral information, to make satisfactory identifications. A spectral match against a reference spectrum, a fragment ion match, a ratio of matching fragment ions, or even an isotope pattern, may be needed to address the level of certainty required. Moreover, for applications that involve numerous unknowns, like metabolomics, a scientist may not have standards to use as references. In these situations a spectral library, particularly an MS^n spectral library, provides the extra information needed to confidently identify compounds.

ThermoFisher SCIENTIFIC

The online mzCloud™ spectral library and Thermo Scientific™ mzVault™ software with local spectral libraries offer a solution for both routine and research applications needing extra certainty when identifying compounds.

- Extensively curated, high-quality online library
- Multiple fragmentation energy levels represented
- Covers a wide chemical space to include compounds relevant to all applications
- For both research and targeted screening workflows
- Freely accessible to all online

mzCloud: The largest, most curated, mass spectral fragmentation library

Developed in collaboration with HighChem, mzCloud is an extensively curated, high-quality mass spectral fragmentation library with significant amounts of spectra and metadata per compound entry. Its spectral data is acquired using Thermo Scientific™ Orbitrap™ mass spectrometers. With a significant number of high-resolution accurate-mass (HRAM) spectra publicly available online (Figure 1), mzCloud includes exhaustive high-resolution MS/MS and multi-stage MS^n spectra.

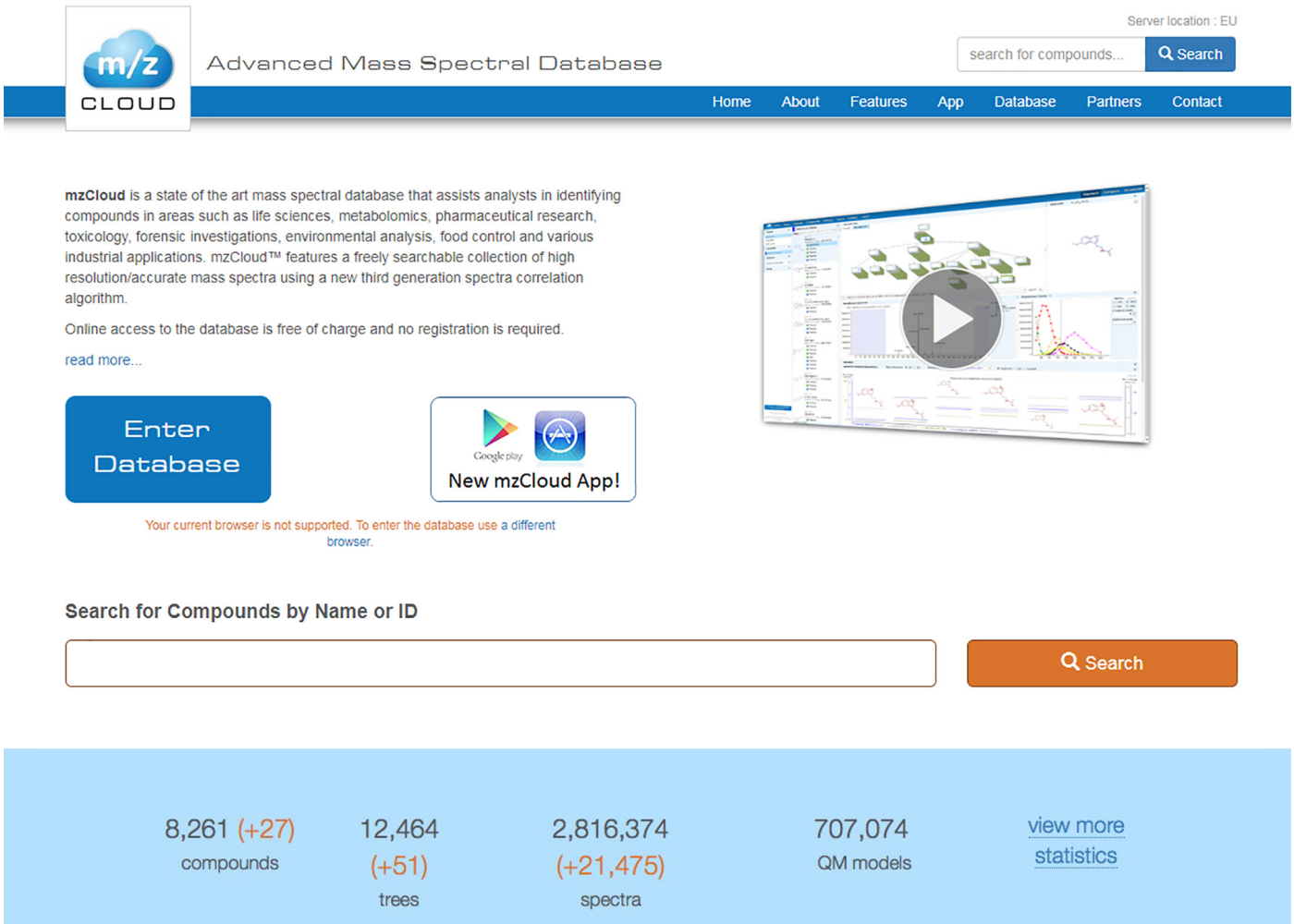

Figure 1. mzCloud is easily and freely accessed online. Anyone can search, review, and use its extensive information. Note that the numbers above were correct at time of publication. For the most up to date numbers visit mzCloud.org.

Each raw mass spectrum has been filtered, recalibrated, averaged, and annotated. Importantly, unlike other mass spectral libraries, an expert curator evaluates each spectrum before it is added to the library, ensuring absolute confidence in library contents and ultimately the results delivered to the user. In addition to experimental and processed mass spectral data, each mzCloud record contains the compound name with synonyms, chemical structure, computationally and manually annotated fragments, adduct and multiply charged ions, molecular formulas, predicted precursor structures, detailed experimental information, peak accuracies, mass resolution, InChi, InChiKey, and numerous other identifiers (Figure 2).

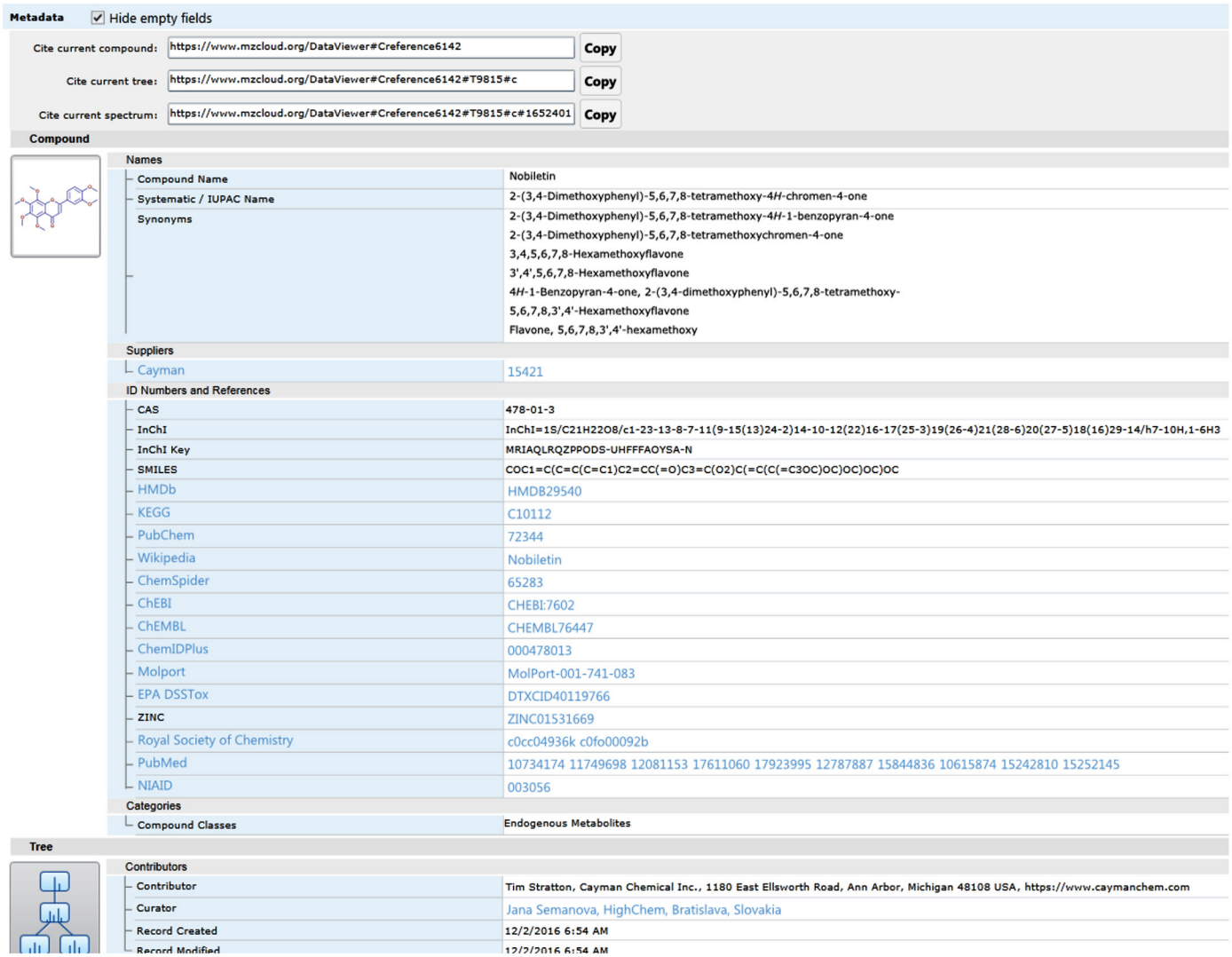

Figure 2. mzCloud includes a wealth of useful information for every compound, including structures annotated onto the fragmentation spectra, chemical synonyms, and links to other data sources. Information stored with the metadata includes instrument and acquisition parameters, and who performed the curation.

Multiple fragmentation energy levels included

Unlike many other libraries containing spectra acquired at only one or two or three energy levels, the MS^n fragmentation spectra in mzCloud are acquired at various collision energies, for multiple adducts, using collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD). Approximately 10–20 HCD energies, which produce triple-quadrupole-like spectra, are included for every compound. Ion trap-like CID spectra are collected using dynamic optimization. The multi-stage, multi-dimensional MS^n spectra are arranged into spectral trees, which are both extensive and comprehensive, typically containing more than 5000 spectra for each compound, with many exceeding MS^4 fragmentation steps (Table 1C and 1D).

Users can acquire data using a broad range of parameters and still achieve high-confidence matches due to the extensive spectral data available for each compound. As a result, labs can run their instrument using the parameters that best suit their applications and still rely on mzCloud to provide a confident match.

mzCloud includes interactive breakdown curves for every molecule, which allow users to review compound fragments for each energy level (Figure 3). This capability can be used to quickly determine the best acquisition energy to use for a particular compound, streamlining method development.

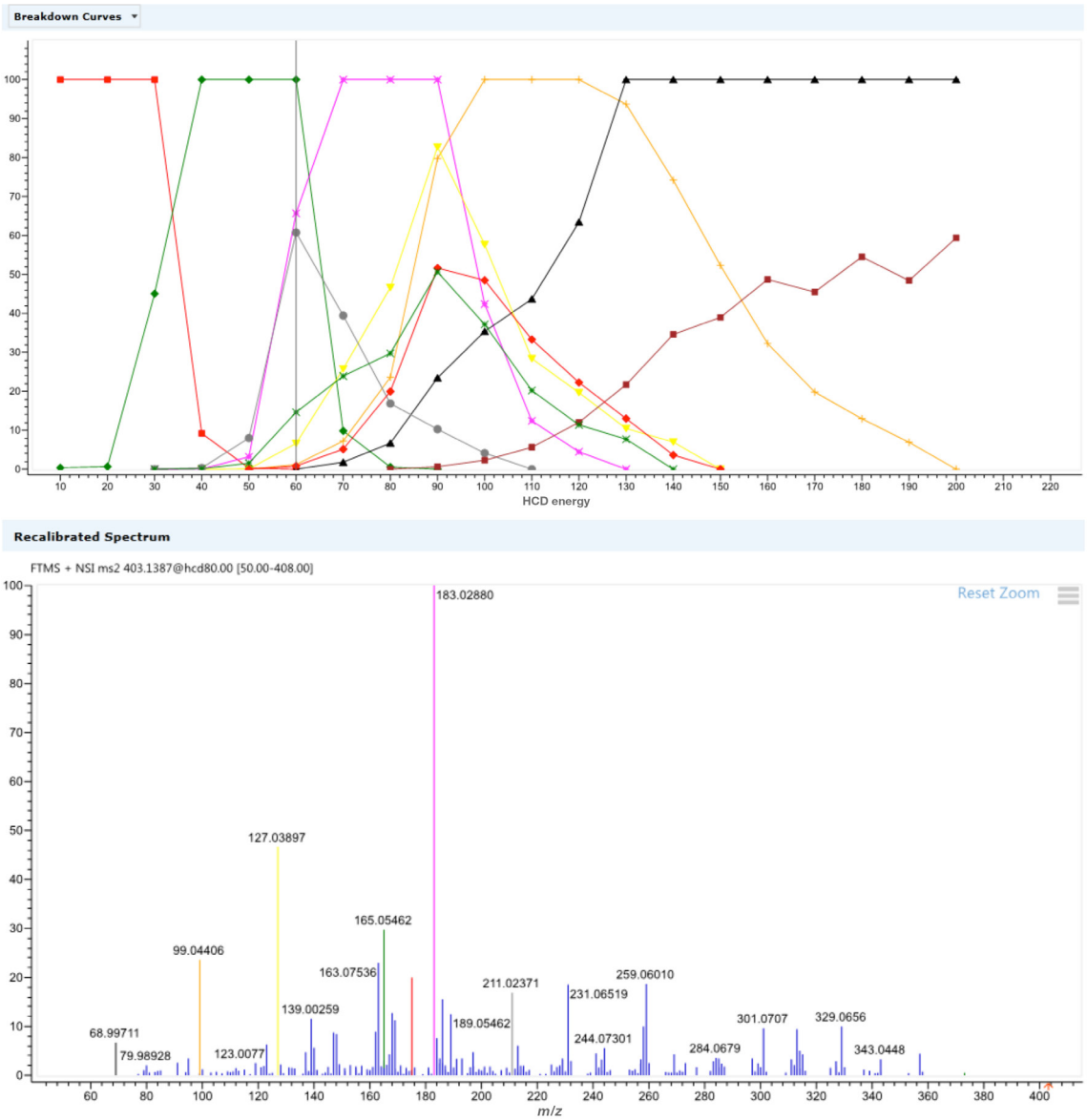

Figure 3. Example of a breakdown curve and corresponding fragmentation spectrum. Being able to review data from multiple fragmentation types and collision energies for each compound facilitates rapid method development for optimal fragmentation acquisition energies.

Extensively curated for highest quality

Experts curate all of the data provided in mzCloud. Why? The curation process significantly improves quality of matches, and therefore confidence in compound identifications. Though Orbitrap mass spectrometers have less than 3-ppm mass error, there is always a slight amount of inaccuracy in library query data. If the reference spectra in the library also contain errors, these errors become additive, making it harder to make identifications with high match scores.

Curation eliminates the error and variability in reference spectra by averaging multiple individually acquired spectra, removing mass spectral noise, and recalibrating mass spectra based on predicted formulas. Because the precursor is known, fragment ion elemental composition and structures can be predicted. During curation, the process is performed for every energy and MS^n level, for each fragmentation method. Due to the curation process, the spectral data in mzCloud is of extremely high quality (Figure 4).

Curated spectral libraries provide compelling advantages compared to crowd-sourced libraries, the most important of which is that crowd-sourced libraries may have inconsistent, inaccurate, and uncontrolled data. As a result, searching crowd-sourced libraries can produce inconclusive or incorrect matches which could result in decreased productivity as users reanalyze data.

Wide chemical space includes compounds relevant to all applications

At the time of this publication, mzCloud includes over 2.7 million spectra (Tables 1A–1H); with new compounds being added daily. Visit mzcloud.org for the latest updates.

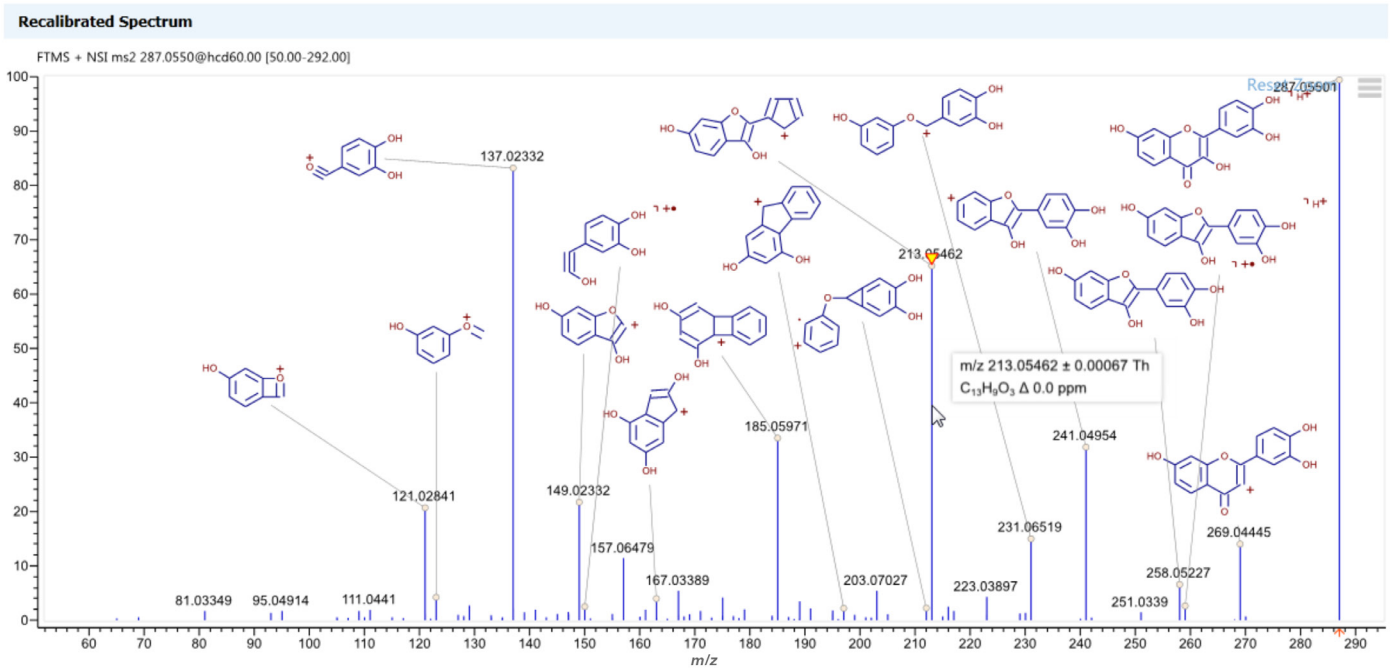

Figure 4. Highly curated data are noise filtered, recalibrated and include annotation of structure, formula, and neutral loss. Manual data curation ensures high quality of all library contents, enabling scientists to achieve greater confidence when identifying unknowns.

Table 1A. mzCloud statistics for number of compounds (8,234) and number of spectral trees (12,413)

	Raw	Filtered	Recalibrated
Positive	8,980	8,980	8,980
Negative	3,433	3,433	3,433

Table 1B. mzCloud statistics for number of spectra (2,794,899)

	Raw	Filtered	Recalibrated
Positive	1,294,267	479,038	479,048
Negative	309,114	116,716	116,716

Table 1C. mzCloud statistics for number of spectra in positive ESI (2,252,353)

	Raw	Filtered	Recalibrated
MS^1	56,388	28,828	28,828
MS^2	523,300	188,831	188,841
MS^3	223,365	78,606	78,606
MS^4	283,006	104,749	104,749
MS^5	163,817	61,387	61,387
MS^6	36,765	13,774	13,774
MS^7	6,705	2,511	2,511
MS^8	843	322	322
MS^9	78	30	30

Table 1D. mzCloud statistics for number of spectra in negative ESI (542,546)

	Raw	Filtered	Recalibrated
MS^1	22,097	10,694	10,694
MS^2	187,266	68,721	68,721
MS^3	46,019	17,124	17,124
MS^4	43,351	16,207	16,207
MS^5	9,927	3,792	3,792
MS^6	454	178	178

Table 1E. mzCloud statistics for number of precursors (121,748)

	Positive	Negative
MS^1	43,169	17,144
MS^2	14,869	6,802
MS^3	17,579	6,118
MS^4	11,268	1,707
MS^5	2,493	43
MS^6	484	0
MS^7	65	0
MS^8	7	0

Table 1F. mzCloud statistics for number of peak annotations (9,487,424)

Formulae	Structures
8,423,172	1,064,252

Table 1G. mzCloud statistics for number of peak annotations (9,487,424)

PM6 level	BLYP/6-31$^+$G* level	BLYP/6-311^{++}G** level
578,286	101,761	13,844

Table 1H. mzCloud statistics for compound classes

	Compounds	Trees	Spectra
Therapeutics/ Prescription Drugs	1,047	2,031	299,127
Drugs of Abuse/ Illegal Drugs	924	1,350	162,738
Sports Doping Drugs	214	334	37,186
Steroids/ Vitamins/ Hormones	21	49	10,268
Endogenous Metabolites	1,203	2,174	911,004
Natural Products/ Medicines	79	143	119,736
Natural Toxins	64	99	132,882
Counterfeit Drug (Therapeutic)	71	116	11,297
Extractables/ Leachables	349	583	124,349
Pesticides/ Herbicides	536	996	108,073
Excipients/ Additives/ Colorants	124	216	81,912
Illegal Additives	53	98	9,204
Personal Care Products/ Cosmetics	75	126	37,350
Textile Chemicals/ Auxiliary/Dyes	45	69	9,811
Industrial Chemicals	303	527	62,661
Perfluorinated Hydrocarbons	20	34	2,066

Library searches can be compound-class specific, or for complete unknowns, across all compound classes. For targeted metabolomics workflows, the library includes MS/MS spectra for every compound in the five most important pathways. For clinical toxicology research and forensics, the library attempts to cover 100% of the currently known (that is as soon as they are identified) psychoactive designer drug compounds, facilitating rapid sample turn-around times for these challenging analyses where understanding which compounds have been used can be of critical importance. Sub-structure search capabilities enable identification of known common moieties in a new and unknown designer drug, which can aid in putative identification of the drug compound.

Definitions: database versus library

- **Chemical database:** A collection of chemical information with, at minimum, a name and a formula. It may also include structures, elemental composition, retention time information, and additional biological or chemical metadata.
- **Accurate-mass (AM) database:** An AM database contains accurate mass formulas.
- **Accurate-mass retention-time (AMRT) database:** An AMRT database contains accurate mass and retention time(s) for specific analytical methods.
- **Library:** A library is a database (AM or AMRT) that in addition to the above also contains reference mass spectral data, MS/MS, and/or MS^n data on compounds.

Software for searching, curating and storing local spectral libraries

mzVault is the Thermo Scientific software application for searching and storing your high-quality custom in-house (offline) MS/MS spectral libraries. Because mzVault includes the MS^2-level spectral data provided in mzCloud, it is useful when the data system used for data processing is not, or cannot be, connected to the Internet.

Even when complex samples give rise to complete unknowns that do not match any compound in mzCloud, or where there are too few identified fragments to make a putative identification, the ability to store this information within a custom library allows users to then search against this to see if any 'unknown' has previously been observed, thereby building up a repository of useful information. It is important to note that the included local spectral library accessed by mzVault covers fewer compounds than mzCloud, because mzVault is only a snapshot of the information contained within mzCloud. An updated snapshot is available to Thermo Scientific™ TraceFinder™ software (revision 5.0 and above) customers free of charge from the Flexera Thermo Scientific Software Portal, and is typically updated twice a year.

For users wishing to curate their own in-house libraries, the ability to curate high-quality mzCloud spectral libraries is available as part Thermo Scientific™ Mass Frontier™ spectral interpretation software. In-house library creation is particularly useful for organizations performing proprietary analyses, or for those wishing to catalogue unknowns and the frequency of their occurrence, as well as those needing to share collected knowledge.

To speed curation, the software can automate curation steps, or curation can be performed manually, with simple step-by-step reviewing. As already described, the curation steps include spectral averaging, noise and artifact removal, annotation of fragments with elemental compositions and structures, and breakdown-curve creation if the spectra have been acquired at multiple energy levels (Figure 5). All of the steps ensure the highest confidence in the spectral library created.

Powerful research solution: Compound Discoverer software for unknown identification

Because small molecule researchers don't necessarily know the compounds they will encounter, they need an up-to-date library comprised of compounds with a broad chemical diversity. With comprehensive, integrated libraries, databases, and statistical analysis tools able to link in customizable workflows, Thermo Scientific™ Compound Discoverer™ software provides a complete solution for small-molecule research that includes unknown identification, determination of real differences between samples, and elucidation of biological pathways.

Compound Discover software allows users to realize the power of both mzCloud and mzVault spectral libraries, within a single software application and data processing workflow, for individual samples or for batch analyses (Figure 6). In addition, spectra of interest can be directly exported from Compound Discover software to mzVault to create special interest or proprietary libraries. Curation tools can be used to ensure that the spectral data added are of highest quality and therefore utility. Using the mzLogic algorithm, putative search results can be ranked using the extensive fragmentation spectra in mzCloud, streamlining the unknown identification process.

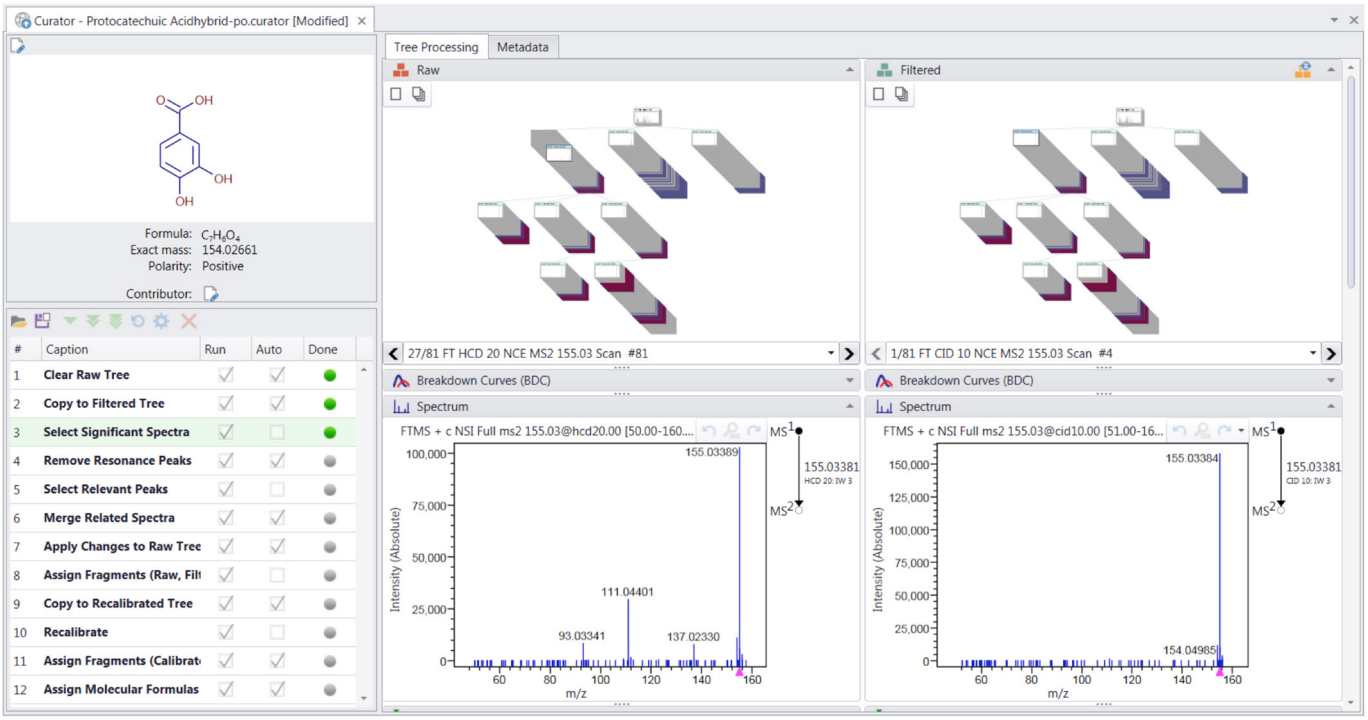

Figure 5. Spectral library curation tools provided by Mass Frontier software. Curator, part of Mass Frontier software, allows you to create your own high-quality spectral libraries; the left panels show the raw and uncurated data, and the right panel shows the filtered, curated data.

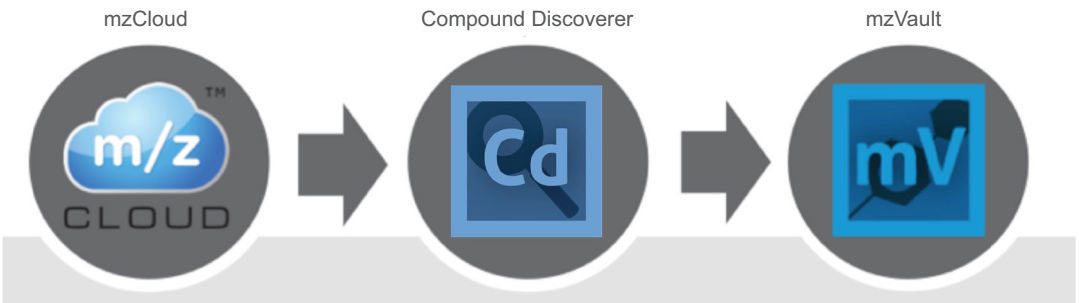

Figure 6. Connectivity of mzCloud and mzVault to Compound Discoverer. To identify unknowns, both unknown and similarity searches can be performed directly from Compound Discoverer software. In addition, scientists can create and curate customized libraries using compounds identified with Compound Discoverer software.

When no good matches result from a spectral library search of mzCloud, structural similarity matches can provide structural information useful for unknown identification (Figure 7). Similarity searches of mzCloud spectra yield compounds that may be structurally related to an unknown compound of interest. The results of a similarity search can be combined with the results of orthogonal chemical database searches to assist with putative structure selection. Orthogonal chemical databases typically also include additional compound information to aid in making identifications.

Integrated solutions for routine applications

mzCloud and mzVault supply important compound content for Thermo Fisher Scientific solutions for routine LC-MS applications (Figure 8).

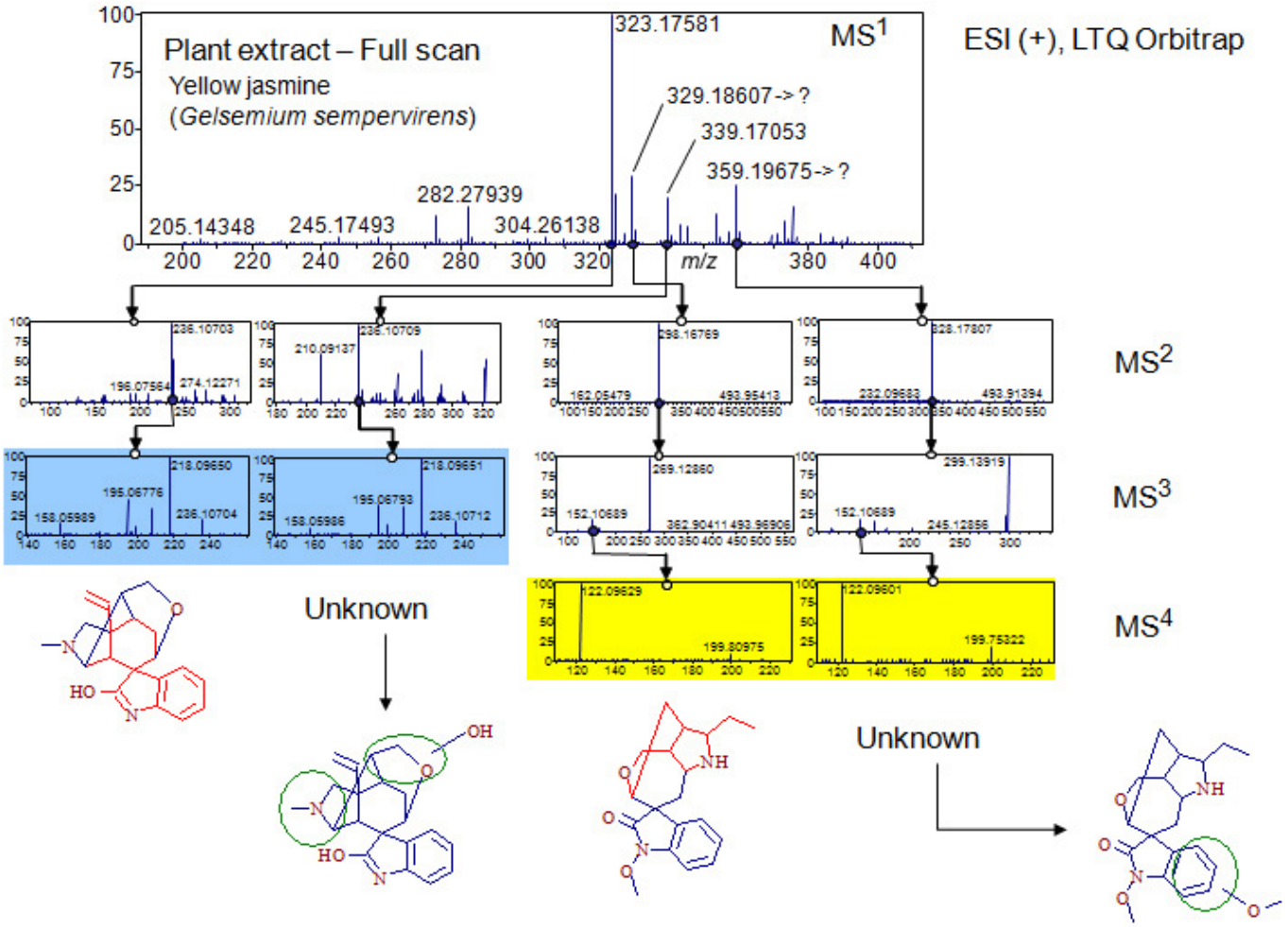

Figure 7. An example of how structural similarity, or sub-structure searching can aid in compound identification. Even when an unknown yields no search results, comparing experimental fragmentation spectra to the extensive fragmentation information in mzCloud can yield sub-structural matches that aid in putative identification, with mzLogic using this information to rank proposals, aiding the identification process.

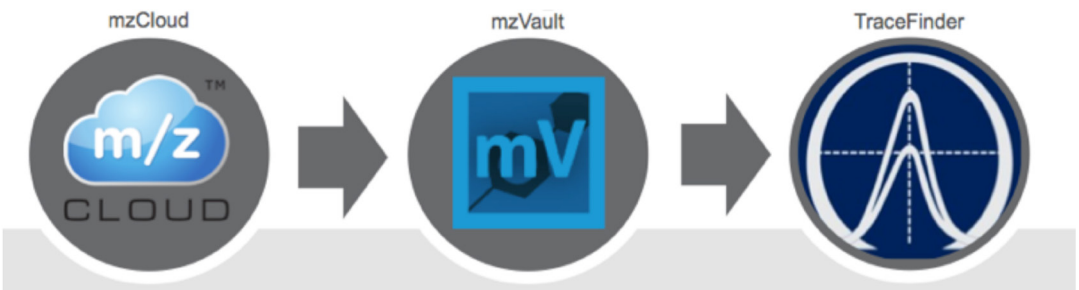

Figure 8. mzCloud spectral libraries provide content for mzVault and TraceFinder workflows.

thermo scientific

With broad compound coverage, the mzCloud and mzVault libraries contain extensive information required for routine targeted screening using TraceFinder software. With the ability to create and edit local libraries and compound databases, mzVault users can easily add compounds to their TraceFinder software targeted workflows. If needed, the libraries can also be used for unknown identification.

In addition, the mzCloud library is the comprehensive source of the application-specific spectral libraries included with the Thermo Scientific™ Pesticide Explorer Collection and Thermo Scientific™ VetDrugs Explorer Collection. These collections provide pre-tested, validated, start-to-finish LC-MS workflows tailored for multiclass analysis of pesticides or veterinary drugs in complex matrices. The libraries included with these collections are updated as new, relevant compounds are added to mzCloud.

Conclusion

mzCloud spectral libraries and mzVault software are designed to address the challenges of small molecule identification for routine and research applications. The online mzCloud spectral library is an extensively curated online library that contains high-quality, fully annotated Orbitrap based MS^n spectral data. mzCloud assures unprecedented certainty in compound identification through its comprehensive and diverse coverage of chemical space, extensive manual curation, and multiple energy levels of fragmentation provided. mzLogic uses all of the available data to aid in ranking putative identifications, streamlining the small moilecule unknown identification workflow. mzVault software offers a solution for laboratories that wish to perform offline data processing, or the power to build and search their own propriety libraries.

> The use of these libraries allows users to rapidly progress from sample analysis to actionable knowledge with a high degree of confidence.

Both the mzCloud spectral library and mzVault software integrate with Compound Discoverer software to create flexible small-molecule data-processing workflows with an enormous degree of functionality and customizability. For routine applications, mzCloud and mzVault can be used with TraceFinder software for targeted screening and unknown identification. In addition, the mzCloud spectral library provides the high-quality application-specific spectral libraries that are part of the Pesticide Explorer Collection and VetDrugs Explorer Collection, end-to-end LC-MS workflows.

Learn more

- www.mzcloud.org
- Pesticide Explorer Collection
- VetDrugs Explorer Collection

What are you waiting for? Evaluate for free today:

- thermofisher.com/CompoundDiscoverer
- thermofisher.com/TraceFinder
- thermofisher.com/MassFrontier

For Research Use Only. Not for use in diagnostic procedures.

Learn more at **thermofisher.com/Orbitrap**

ThermoFisher SCIENTIFIC