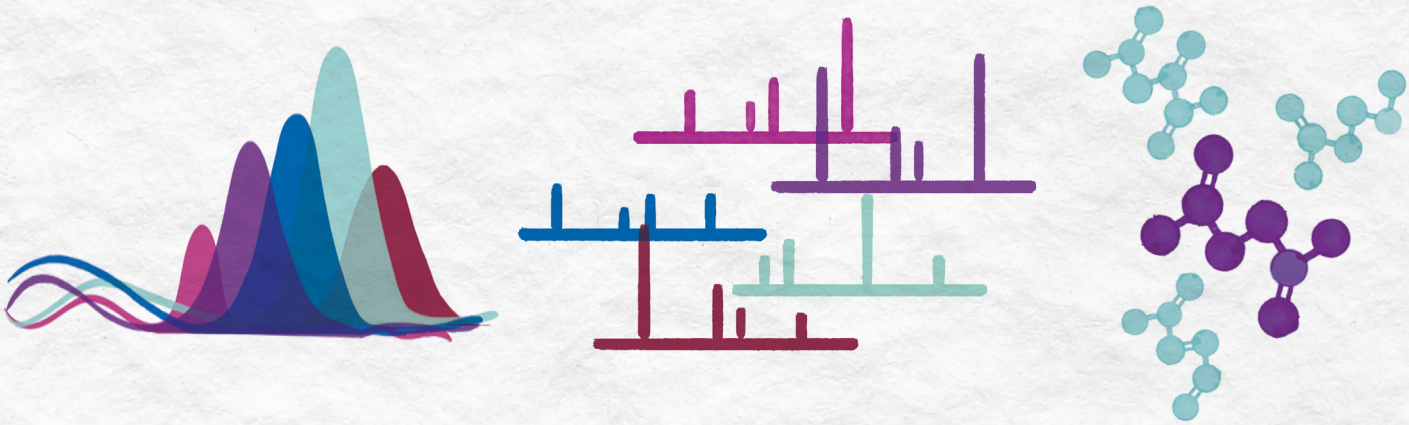
White paper | WP65374



# Delivering confidence for small molecule identification

**Authors**

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 liquid chromatography-mass spectrometry (LC-MS) applications strive for certainty when identifying detected compounds. For many targeted LC-MS methods, positive identifications made using a combination of 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.

However, for many LC-MS applications, laboratories require additional information, such as mass spectral fragmentation information, to make satisfactory identifications. A spectral match against a reference spectrum, 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, or standards may simply not be available. In these situations, a spectral library, particularly an $MS^n$ spectral library, provides the information needed to attempt to identify compounds.

The online Thermo Scientific™ 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.

## mzCloud library: The largest, most curated collection of mass spectral fragmentation data

Developed in collaboration with HighChem, mzCloud library is an extensively curated, high-quality mass spectral fragmentation collection of data 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 library includes exhaustive high-resolution MS/MS and multi-stage $MS^n$

- 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

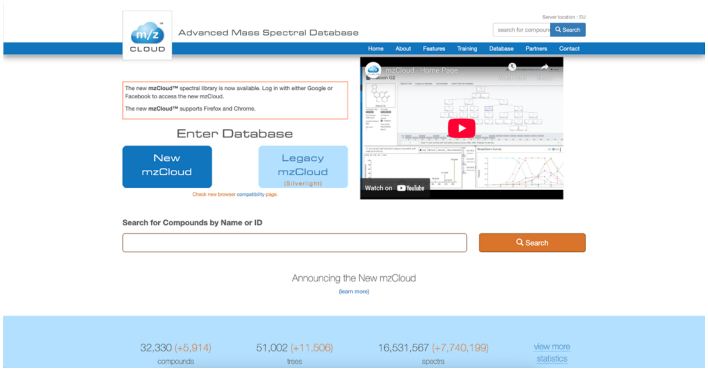



**Figure 1: mzCloud library is easily and freely accessed online.** Anyone can search, review, and use its extensive information. Note that the numbers above were correct at time of publication. For the most up-to-date numbers, visit **mzCloud.org**.

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, helping ensure 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 multiple charged ions, molecular formulas, predicted precursor structures, detailed experimental information, InChi, InChiKey, and numerous other identifiers (Figure 2).

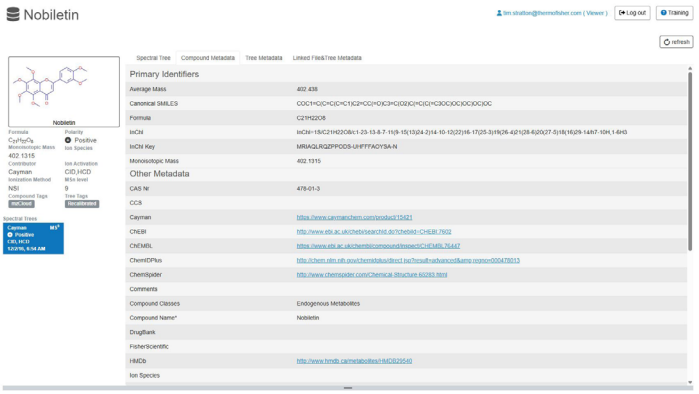



**Figure 2: Compound information for nobiletin including multiple links to external data sources.** Similar metadata is available for the spectral tree.

## Multiple fragmentation energy levels included

Unlike many other libraries containing spectra acquired at only one or very few energy levels, the $MS^n$ fragmentation spectra in mzCloud library are acquired at various collision energies, for multiple adducts, using ion trap resonance excitation 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 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 5,000 spectra for each compound, with many exceeding $MS^4$ fragmentation steps (Figure 3).

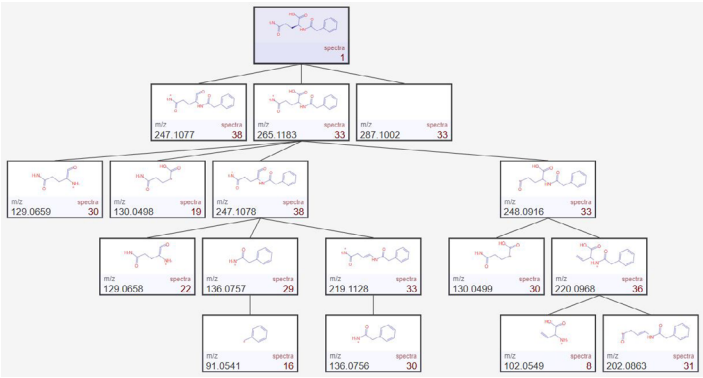



**Figure 3: Extensive $MS^n$ fragment spectral tree for phenylacetylglutamine.**

Users can acquire data using a broad range of parameters and still achieve high confidence matches with 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 library to provide a confident match.

mzCloud library includes interactive breakdown curves for every molecule, which allow users to review compound fragments for each energy level (Figure 4). This capability can be used to quickly determine the best acquisition energy to use for a particular compound, streamlining method development.

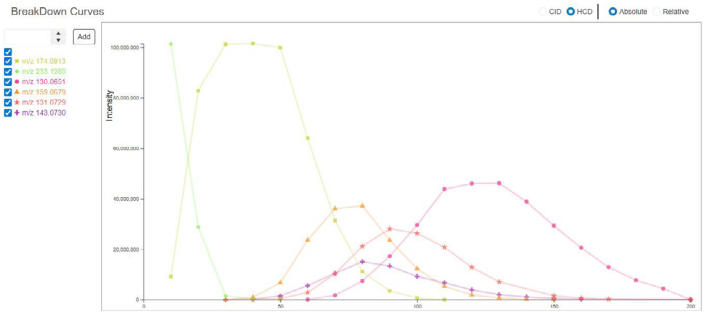



**Figure 4: HCD Breakdown curve for melatonin, top 6 fragment ions.**

## Extensively curated for highest quality

Experts curate all data provided in mzCloud library. Why? All measurements have an amount of error; this includes both the query data and the data used to build a library. By curating and recalibrating the library data, a portion of the overall error in any potential library match is reduced. The curation process helps to avoid these errors becoming additive, making it easier to make identifications with high match scores. The curation process significantly improves quality of matches, and, therefore, confidence in compound identifications. For the mzCloud Reference library, this curation involves the manual review by a curation expert to tailor the curation parameters. Collecting the information over the manual curation of thousands of spectral trees, an expert-tuned curation method is applied to generate the mzCloud Autoprocessed library.

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 and fragment structures. 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. Because of the curation process, the spectral data in mzCloud library is of extremely high quality (Figure 5).

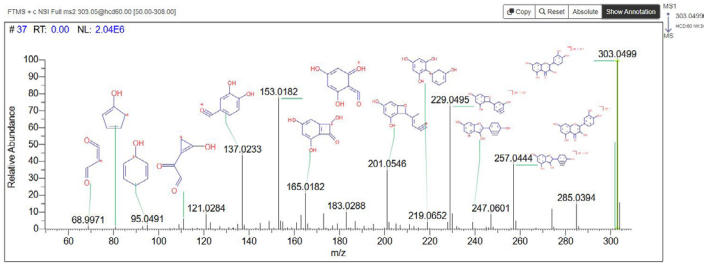



**Figure 5: Highly curated data are averaged, noise filtered, recalibrated, and include annotation of structure.** Data curation helps ensure high quality for all library contents, enabling scientists to achieve greater confidence when identifying unknowns.

Curated spectral libraries created from known standards 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 may 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 library includes over 16 million spectra (Table 1); with new compounds being added daily. Visit **mzcloud.org** for the latest updates.

**Table 1. Compound classes and data in mzCloud library.**

| Class | Compounds | Spectra |
|---|---|---|
| Counterfeit drug (therapeutic) | 97 | 19,580 |
| Drugs of abuse/illegal drugs | 1,377 | 87,783 |
| Endogenous metabolites | 13,818 | 9,434,583 |
| Excipients/additives/colorants | 483 | 144,830 |
| Extractables/leachables | 553 | 49,790 |
| Illegal additives | 65 | 5,824 |
| Industrial chemicals | 527 | 49,416 |
| Natural products/medicines | 1,498 | 797,565 |
| Natural toxins | 156 | 63,009 |
| Perfluorinated hydrocarbons | 163 | 17,910 |
| Personal care products/cosmetics | 465 | 125,750 |
| Pesticides/herbicides | 1,611 | 584,148 |
| Sports doping drugs | 1,325 | 373,036 |
| Steroids/vitamins/hormones | 228 | 174,478 |
| Textile chemicals/auxiliary/dyes | 295 | 27,653 |
| Therapeutics/prescription drugs | 3,252 | 1,603,428 |
| Veterinary medicines | 483 | 389,480 |

Library searches can be compound-class specific, or for a more robust unknown analysis, 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 all of the currently known (that is, as soon as they are identified) psychoactive designer drug compounds, facilitating rapid sample turnaround times for these challenging analyses where understanding which compounds have been used can be of critical importance. Sub-structure and spectral similarity 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.

## Software for searching, curating, and storing local spectral libraries

mzVault software is an application for searching and storing your high-quality custom in-house (offline) MS/MS spectral libraries. Because mzVault software includes the $MS^2$-level spectral data provided in mzCloud library, 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 library, 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 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 software covers fewer compounds than mzCloud library, because it is only a snapshot of the information contained within it. An updated snapshot is available to Thermo Scientific™ TraceFinder™ software (revision 5.0 and above) customers free of charge from the Thermo Scientific™ Software Portal, which is typically updated annually.

For users wishing to curate their own in-house full $MS^n$ libraries, the ability to curate high-quality mzCloud spectral libraries is available as part of Thermo Scientific™ Mass Frontier™ spectral interpretation software. In-house library creation is particularly useful for organizations performing proprietary analyses, or for those wishing to catalog 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 6). All steps help ensure the highest confidence in the spectral library created.

**Definitions: database versus library**

- Chemical database: A collection of chemical information with a name and a formula, at a minimum. 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.

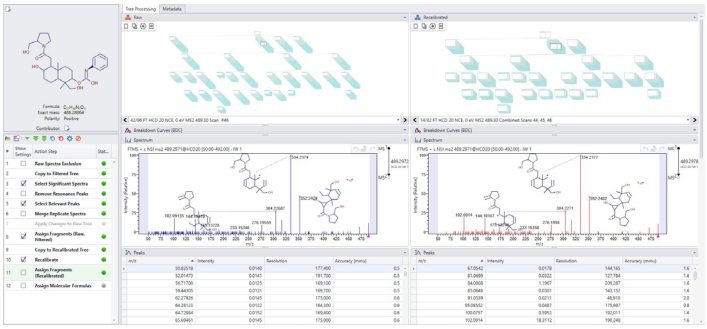


**Figure 6: Spectral library curation tools provided by Mass Frontier software.** Curator, an optional part of Mass Frontier software, allows you to create your own high-quality spectral libraries; the left panel shows the raw and uncurated data, and the right panel shows the filtered, curated, and recalibrated data.

## 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 7).

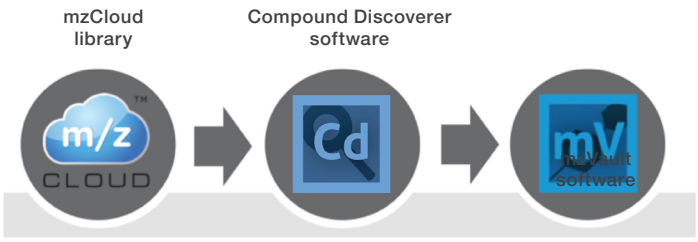



**Figure 7: Connectivity of mzCloud library and mzVault software to Compound Discoverer software.** To identify unknowns, both unknown and similarity searches can be performed directly from Compound Discoverer software. In addition, scientists can create and curate customized libraries using compounds identified with Compound Discoverer software.

When no good matches result from a spectral search of mzCloud library, other options are available. Structural similarity matches can provide structural information useful for unknown identification. These searches can be performed on either the $MS^2$ or $MS^n$ level within Compound Discoverer software to provide potentially informative information when selecting putative candidates (Figure 8). Similarity searches of mzCloud spectra yield compounds that may be structurally related to an unknown compound of interest, while $MS^n$ matches can help to expose possible substructures in unknowns. The Thermo Scientific™ mzLogic™ algorithm takes the use of spectral similarity data a step further. It can help to rank putative chemical database candidates by leveraging fragmentation similarity to map substructures from mzCloud library to candidates' structures. Orthogonal chemical databases typically also include additional compound information to aid in making identifications.

**Figure 8: An example of substructure ID through $MS^n$ search matching portions of an unknown against the library compound stercobilin.**

## Integrated solutions for routine applications

mzCloud library and mzVault software supply important compound content for Thermo Fisher Scientific solutions for routine LC-MS applications (Figure 9).

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

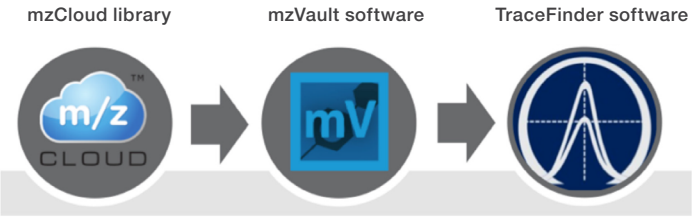



**Figure 9: mzCloud spectral libraries provide content for mzVault and TraceFinder workflows.**

## 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 library helps assure 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 algorithm uses all of the available data to aid in ranking putative identifications, streamlining the small molecule 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 proprietary libraries.

Both 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 offline libraries and mzVault software can be used with TraceFinder software for targeted screening and unknown identification.

## Learn more

- mzcloud.org

## What are you waiting for? Evaluate for free today

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

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