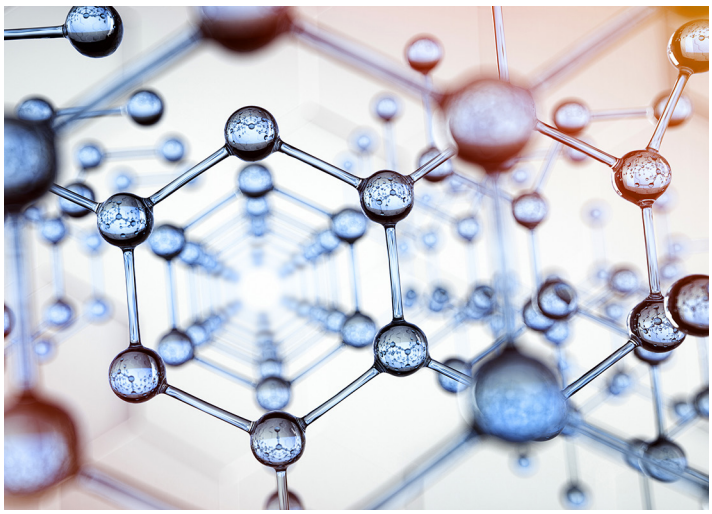
TECHNICAL NOTE No. 65602

Building curated and annotated HRAM MS^n spectral libraries to aid in unknown structure elucidation

Authors: Caroline Ding, Kate Comstock, Seema Sharma, Mark Sanders, Michal Raab

Thermo Fisher Scientific, San Jose, CA

Keywords: Orbitrap ID-X, Mass Frontier, mzLogic, small molecule structure elucidation, library searching, sildenafil



Goal

Validate new library searching algorithms and structure ranking algorithm Thermo Scientific™ mzLogic for local proprietary MS^n spectral libraries to aid in structure elucidation of unknown drug metabolites, impurities or degradants.

Introduction

Small molecule structure elucidation is a very challenging and time-consuming task. A mass spectral library with extensive MS^n spectral tree and substructural information is a valuable tool for rapid identification of small molecule unknowns and unknown structure characterization. The objective of this work was to demonstrate a complete workflow from building a local version of Thermo Scientific mzCloud-HRAM MS^n spectral library using the Thermo Scientific™ Orbitrap ID-X™ Tribrid™ mass spectrometer for in-house proprietary compounds. Automated curation and structural annotations were performed using Thermo Scientific™ Mass Frontier™ structure identification software, version 8.0. The MS^n spectral library with structure annotations not only enables library searching to quickly identify unknown compounds, but also enables substructure matching for unknown compounds not in the library. The new mzLogic structure ranking algorithm uses spectral knowledge within the library to rank possible structure candidates based on a combination of spectral similarity and common substructure overlapping; this results in quick and confident structural identification of unknown drug metabolites, impurities, or degradants.

Experimental

Data acquisition

An Orbitrap ID-X Tribrid mass spectrometer coupled with a Thermo Scientific™ Vanquish™ UHPLC system was used for data acquisition. As proof of concept, a library was created for ten sildenafil drug analog standards (Figure 1).

Compounds were infused by electrospray ionization, and MS^n data were collected with varying collision energies for multiple fragmentation activation types. Four sildenafil standards with the same molecular mass (two of them were in the library, and two of them were not in the library, Figure 2) were mixed and analyzed by LC/MS^n.

Methods

The library compounds were acquired on the Thermo Scientific Orbitrap Tribrid ID-X mass spectrometer with Thermo Scientific™ Tune 3.1 software, which includes library builder templates for infusion and LC/MS. The library template automatically acquires multiple collision energies, multiple spectra for each MS^n (up to MS^4) stage in both HCD and CID. It uses assisted collision energy (15, 30, 45, 60, 75) for the branch leading to CID MS^3. The purpose is to capture comprehensive fragmentation data at multiple collision energy levels and fragmentation modes for each compound in the library to allow confident matching with LC/MS fragmentation data.

The four LC/MS standard compounds were first run individually to obtain the retention times. They were then mixed into one vial and run with the same LC/MS conditions. Fragmentation data for the four *m/z* 489 compounds included up to MS^4: MS^2 at HCD stepped collision energies 40, 60, 90; MS^3 at HCD 40; and MS^4 at CID 40.

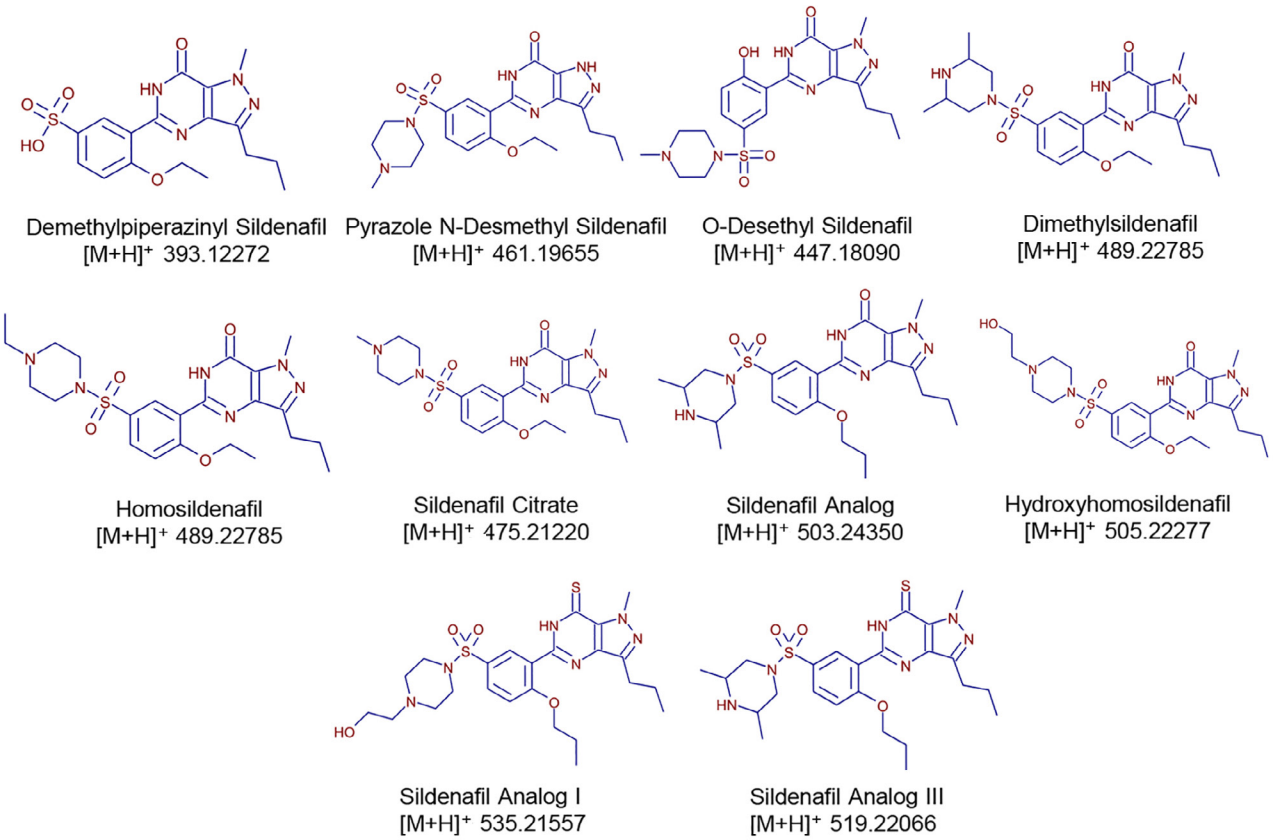


Figure 1. The ten sildenafil drug standards for the library

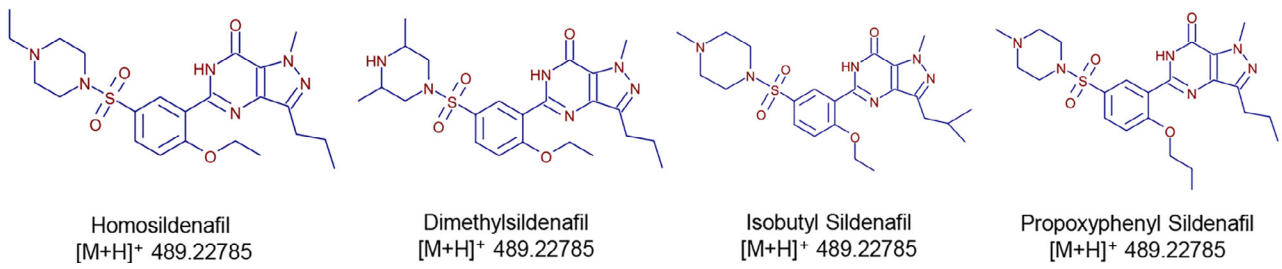


Figure 2. The four sildenafil compound standards in the LC/MS mixture

Data analysis

Mass Frontier 8.0 software was used to build the curated and annotated MS^n spectral library. The DICD (direct infusion component detection) algorithm automatically compiles HCD, CID spectra for each precursor ions at each MS^n stage into a spectral tree. The MS^n spectral trees were curated using the Curator module (Figure 3) in an automated fashion for each compound and saved into a local library (Figure 4).

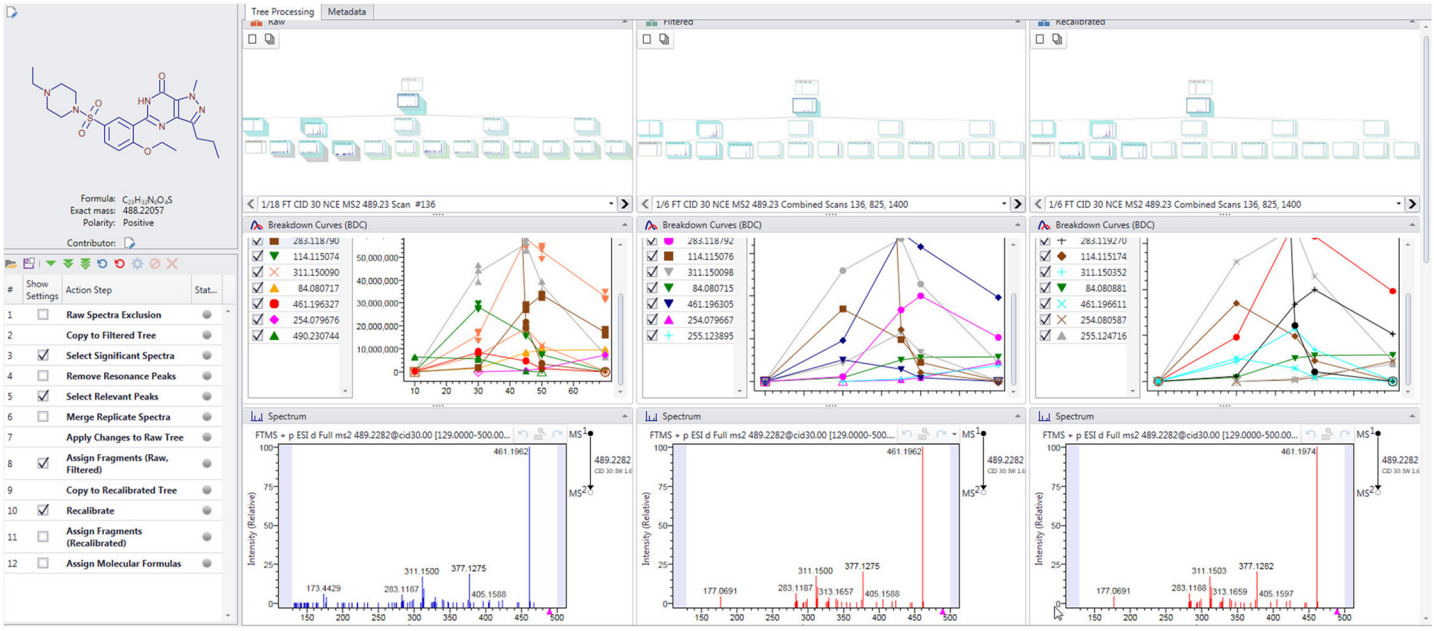


Figure 3. The Curator module in Mass Frontier 8.0 software automatically removes noise and bad quality spectra, keeps only relevant peaks, adds fragment annotations, and recalibrates spectra. The energy dependence of fragment ion formation is plotted in the breakdown curves (BDC) representing the absolute or relative intensity of the ions versus the specific energies used.

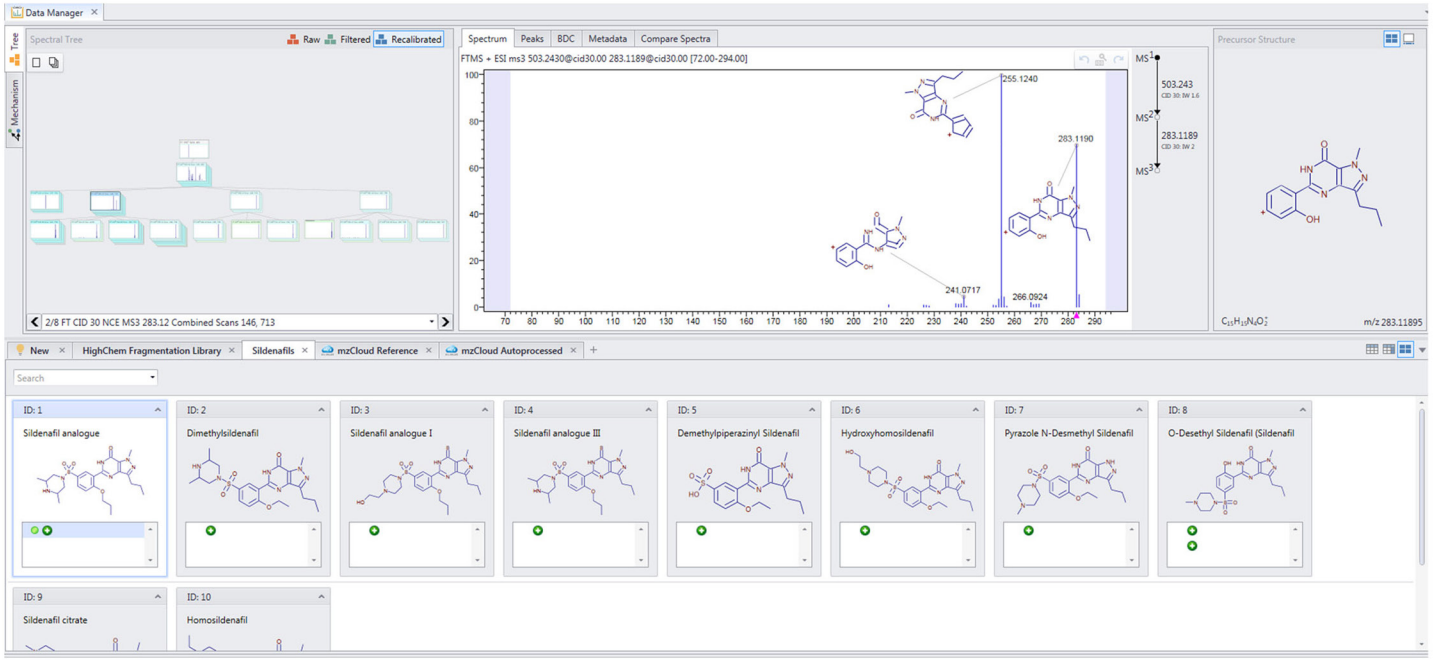


Figure 4. The local MS^n spectral library in Data Manager with precursor structures and fragment annotations at each MS^n stage

LC/MS data of the four standards were processed in Mass Frontier software by Joint Component Detection (JCD) (Figure 5) and spectral tree deconvolution (Figure 6). Joint Component Detection (JCD) is a component detection algorithm that is based on the statistical analysis of ion profile in *m/z* and retention time. It extracts individual mass spectral peak abundance profiles to produce "purified" spectrum or spectral trees, and generates the peak shape of a representative component.

Results and discussion

Compound identification with library searching

The four components with spectral trees were searched against the local sildenafil compound library in Mass Frontier 8.0 Chromatogram Processor module. Mass Frontier 8.0 software includes new library searching types, as well as new and improved matching algorithms that are trained on the extensive mzCloud spectral library with real fragmentation data.

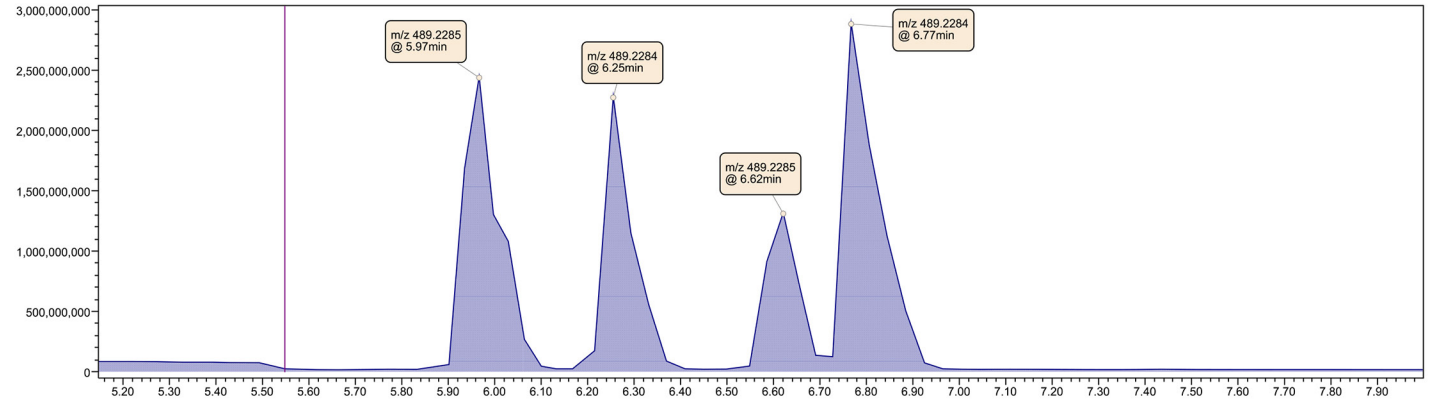


Figure 5. JCD detected four components with *m/z* of 489.2285, which corresponded to the four sildenafil compounds in the LC/MS mixture.

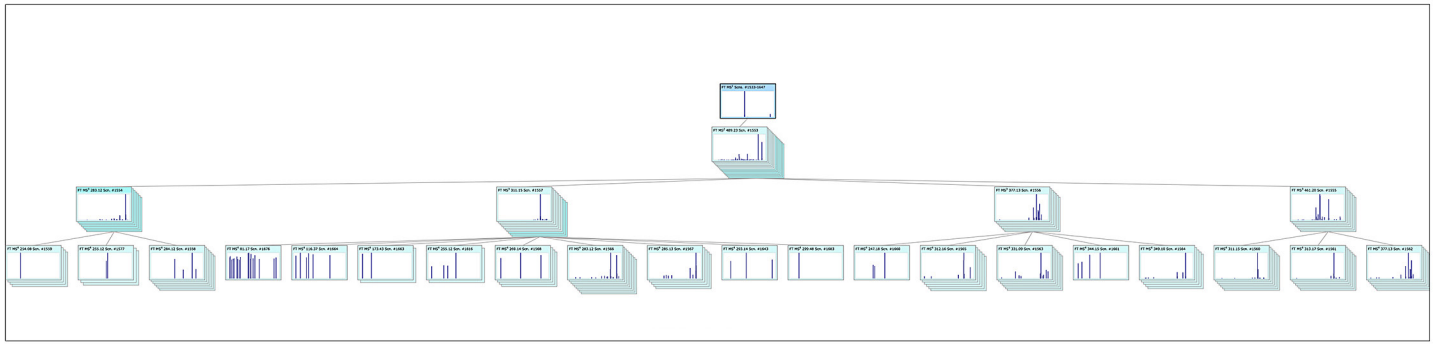


Figure 6. Deconvoluted spectral tree in Mass Frontier software for each of the 489 compounds with up to MS^4 fragmentation spectra from the Orbitrap ID-X system

Table 1 is a summary of the library search types, used stages and constraints.

Compound identification

For the two *m/z* 489 compounds (RT 5.97 min and 6.25 min) that are in the local sildenafil library, Mass Frontier software returned the right compound with confidence match scores of 99 for both identity search (Figure 7) and tree search (Figure 8).

MS^n tree search was able to separate the two library compound hits more in match scores when compared to identity search, which is MS^2 vs. MS^2 only. A comparison of identity search vs. tree search for component #1 is shown in Figures 9 and 10.

The MS^n tree search was able to differentiate the two isomers due to the matching of MS^3 spectra of 461.1962 between query and library. For the example shown

Table 1. Library search types in Mass Frontier 8.0 software

Search types	Used stages and constraints	Use
Identity	• Compares the MS^2 library spectra against the MS^2 query spectra; uses best confidence match score calculation. • The MS^2 precursor ions must match.	Compound Identification
Identity Substructure	• Compares any MS^n library spectra against any MS^n query spectra; uses best confidence match score calculation. • The precursor ions at any MS^n stage must match.	Substructure Identification
Similarity (Forward and Reverse)	• Compares the MS^2 library spectra against the MS^2 query spectra, uses best confidence match score calculation. • The MS^2 precursor ions do not have to match.	Identify structurally similar compounds
Tree Search	• Compares any MS^n library spectra against any MS^n query spectra; considers the whole MS^n hierarchy; uses aggregated tree match score calculation. • The MS^2 precursors for the query spectrum and the library spectrum must match.	Compound Identification **with increased specificity**
Subtree Search	• Compares any MS^n library spectra against any MS^n query spectra; considers the MS^n subtree hierarchy; uses aggregated subtree match score calculation. • The precursor ions at any MS^n stage must match.	Substructure Identification **with increased sensitivity**

Name	Scan No.	Precursor m/z	Match	MSn	t_R (min)	Abundance	Annotation Sources
Components							
Component 1	1569	489.2285	99 Homosildenafil	4	5.966	2,205,419,520	Identity
Component 2	1732	489.2284	99 Dimethylsildenafil	4	6.254	1,925,246,848	Identity
Component 4	1937	489.2285	57 Dimethylsildenafil	4	6.620	794,834,816	Identity
Component 6	2024	489.2284	58 Dimethylsildenafil	4	6.766	1,999,293,184	Identity

Figure 7. Compound matches with identity search

Name	Scan No.	Precursor m/z	Match	MSn	t_R (min)	Abundance	Annotation Sources
Components							
Component 1	1569	489.2285	99 Homosildenafil	4	5.966	2,205,419,520	Tree Search
Component 2	1732	489.2284	99 Dimethylsildenafil	4	6.254	1,925,246,848	Tree Search
Component 4	1937	489.2285	57 Dimethylsildenafil	4	6.620	794,834,816	Tree Search
Component 6	2024	489.2284	58 Dimethylsildenafil	4	6.766	1,999,293,184	Tree Search

Figure 8. Compound matches with tree search

in Figure 9 and Figure 10, the homosildenafil library compound has a signature fragment of 461.1962 in both HCD MS^2 (less intense) and CID MS^2 (very intense); whereas the dimethylsidenafil library compound shows no 461.1962 in HCD MS^2 (at multiple collision energies) and a very low intensity peak in CID MS^2. The spectral tree of component #1 matched MS^2 of 489 from both homosildenafil and dimethylsidenafil with good confidence scores (98.8 and 92.3, respectively). Its MS^3 of 461.1962 matched only with MS^3 of 461.1962 from homosildenafil (with good confidence score of 83.7) (Figure 11), but no match of 461 at MS^3 with dimethylsidenafil. This is a good example of why multiple collision energies and activation types for the library compounds must be acquired on and why MS^n improves confidence of the identification.

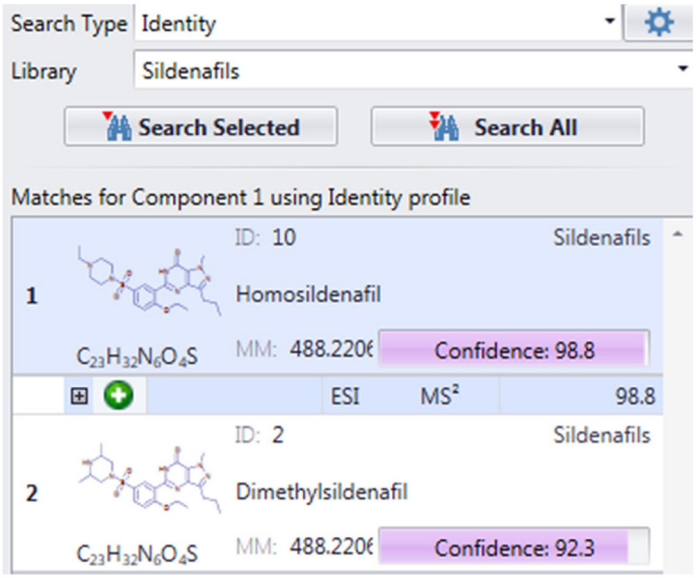


Figure 9. The two library hits and match scores for component #1 with identity search (MS^2 vs MS^2 only)

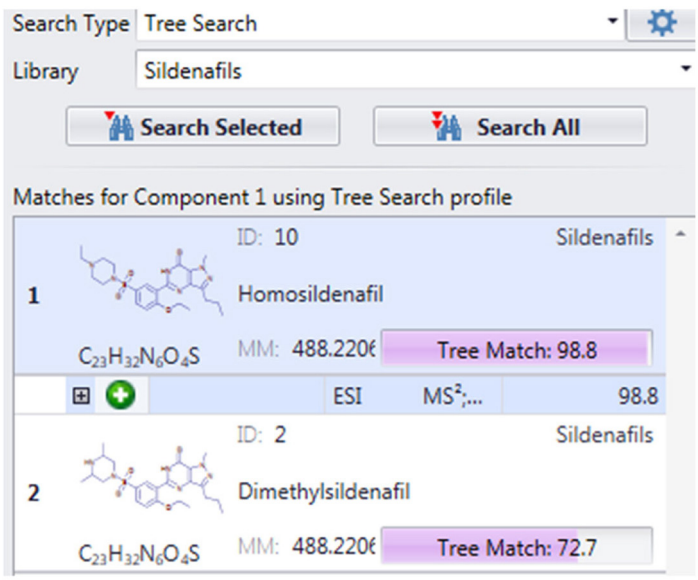


Figure 10. The two library hits and tree match scores for component #1 with MS^n tree search (MS^n vs MS^n)

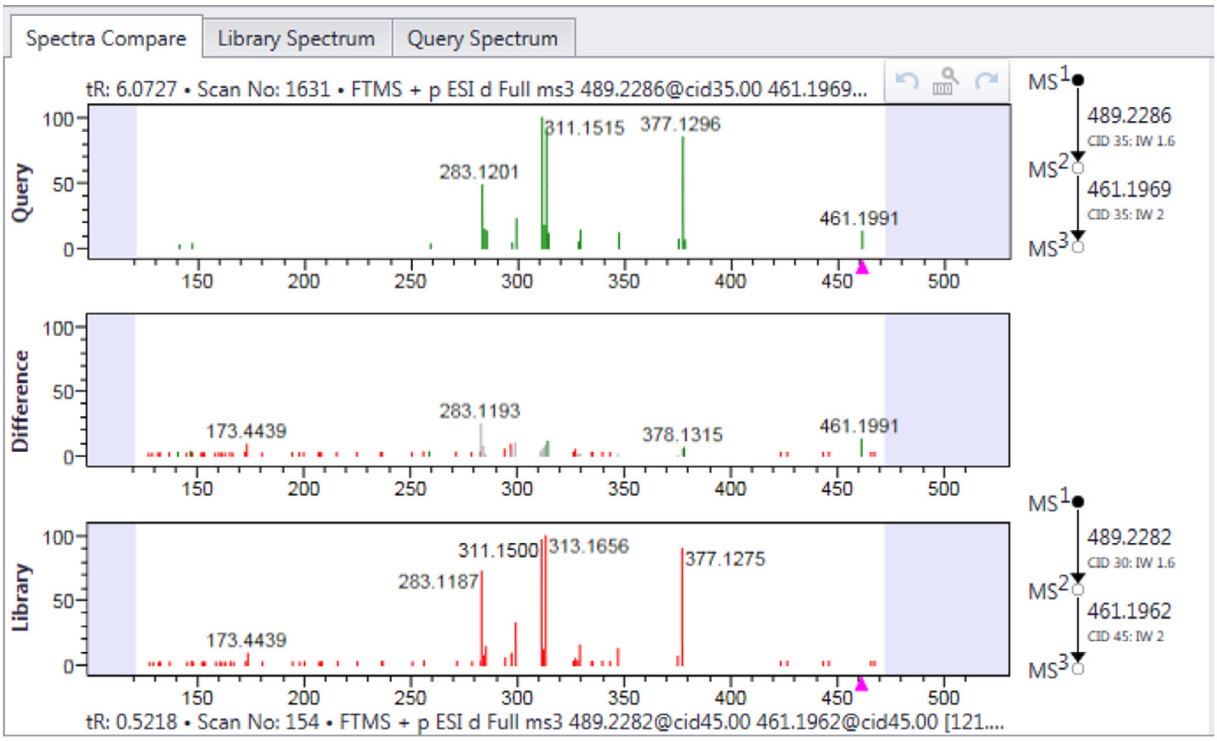


Figure 11. Spectral comparison of MS^3 at 461.1969 of component #1 vs MS^3 at 461.1962 of homosidenafil library compound

Similarity and substructure matching

For the other two *m/z* 489 compounds (RT 6.62 min and 6.77 min) that are not in the sildenafil library, identity search or tree search gave confidence scores of 57 or 58 due to no exact compound matches in the sildenafil library, which is expected. The next step was to try the similarity, identity substructure, and subtree searches to identify substructure matches in the library.

With similarity search, confidence scores of 38 for both *m/z* 489 compounds were obtained (Figure 12). Since similarity search compares query MS^2 vs. library MS^2 only, it did not provide conclusive substructure matches from the library.

Conversely, identity substructure and subtree searches that match MS^n vs. MS^n both gave good substructure matches for the *m/z* 489 compounds, matching the substructure of O-desethyl sildenafil in the library (Figures 13 and 14). The substructure search results are consistent with the common substructure overlap between these *m/z* 489 compounds (isobutyl sildenafil at RT 6.62 min and propoxyphenyl sildenafil at 6.77 min) and O-desethyl sildenafil in the library (Figure 15).

Name	Scan No.	Precursor m/z	Match	MSn	t_R (min)	Abundance	Annotation Sources
Components							
Component 1	1569	489.2285	99 Homosildenafil	4	5.966	2,205,419,520	Tree Search
Component 2	1732	489.2284	99 Dimethylsildenafil	4	6.254	1,925,246,848	Tree Search
Component 4	1937	489.2285	38 Dimethylsildenafil	4	6.620	794,834,816	Similarity Forward
Component 6	2024	489.2284	38 Dimethylsildenafil	4	6.766	1,999,293,184	Similarity Forward

Figure 12. Similarity forward search results for the two *m/z* 489 compounds at RT 6.62 min and 6.77 min

Name	Scan No.	Precursor m/z	Match	MSn	t_R (min)	Abundance	Annotation Sources
Components							
Component 1	1569	489.2285	99 Homosildenafil	4	5.966	2,205,419,520	Tree Search
Component 2	1732	489.2284	99 Dimethylsildenafil	4	6.254	1,925,246,848	Tree Search
Component 4	1937	489.2285	90 O Desethyl Sildenaf	4	6.620	794,834,816	Identity Substructure
Component 6	2024	489.2284	90 O Desethyl Sildenaf	4	6.766	1,999,293,184	Identity Substructure

Figure 13. Identity substructure search results for the two *m/z* 489 compounds at RT 6.62 min and 6.77 min

Name	Scan No.	Precursor m/z	Match	MSn	t_R (min)	Abundance	Annotation Sources
Components							
Component 1	1569	489.2285	99 Homosildenafil	4	5.966	2,205,419,520	Tree Search
Component 2	1732	489.2284	99 Dimethylsildenafil	4	6.254	1,925,246,848	Tree Search
Component 4	1937	489.2285	90 O Desethyl Sildenaf	4	6.620	794,834,816	Subtree Search
Component 6	2024	489.2284	89 O Desethyl Sildenaf	4	6.766	1,999,293,184	Subtree Search

Figure 14. Subtree search results for the two *m/z* 489 compounds at RT 6.62 min and 6.77 min

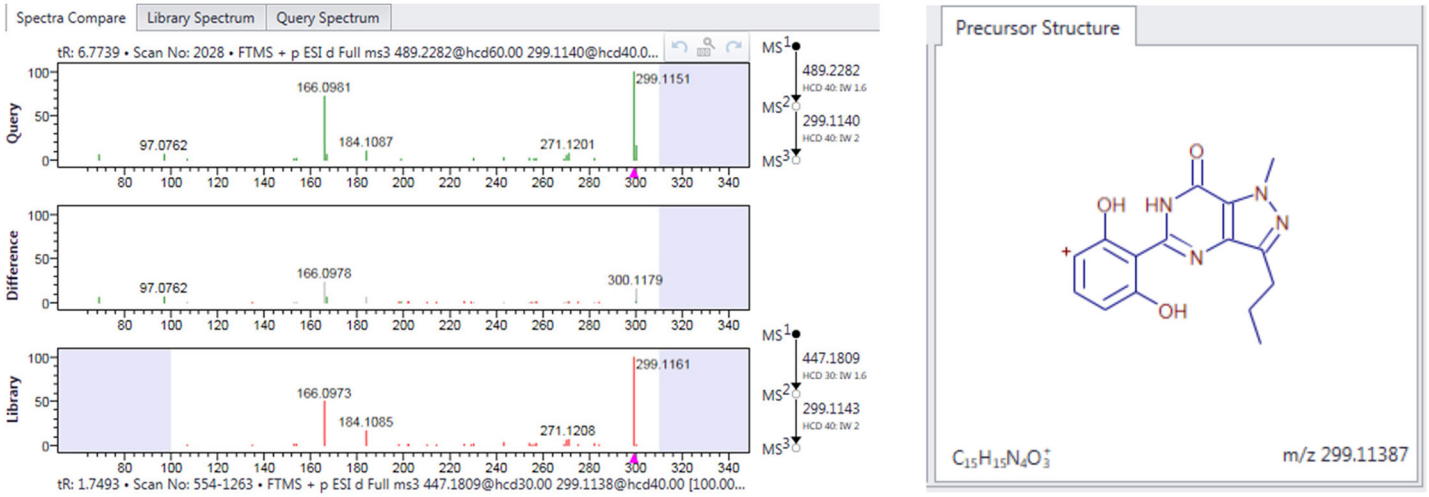


Figure 15. Example of substructure match on the MS³ level, indicating the *m/z* 489 compounds (component #4 and #6) share this common substructure from the library

mzLogic structure ranking

The mzLogic search algorithm is a novel structure ranking algorithm that combines spectral similarity search against spectral libraries and maximum common substructure overlap. The structure candidates can come from user proposed structures and public databases such as ChemSpider, PubChem®, KEGG, and DrugBank, which are seamlessly integrated within the Mass Frontier 8.0 application. For this study, we proposed nine structural isomers for mzLogic ranking for *m/z* 489 compounds at 6.62 min and at 6.77 min. The correct structures were ranked among the top scored candidates (Figures 16 and 17).

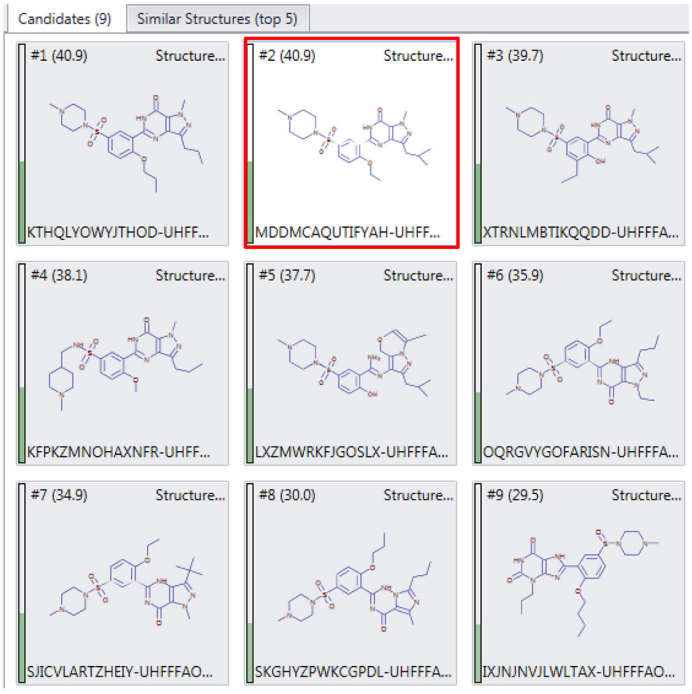


Figure 16. For the *m/z* 489 compound at 6.62 min, the correct structure was ranked #2 and had the same mzLogic score as rank #1.

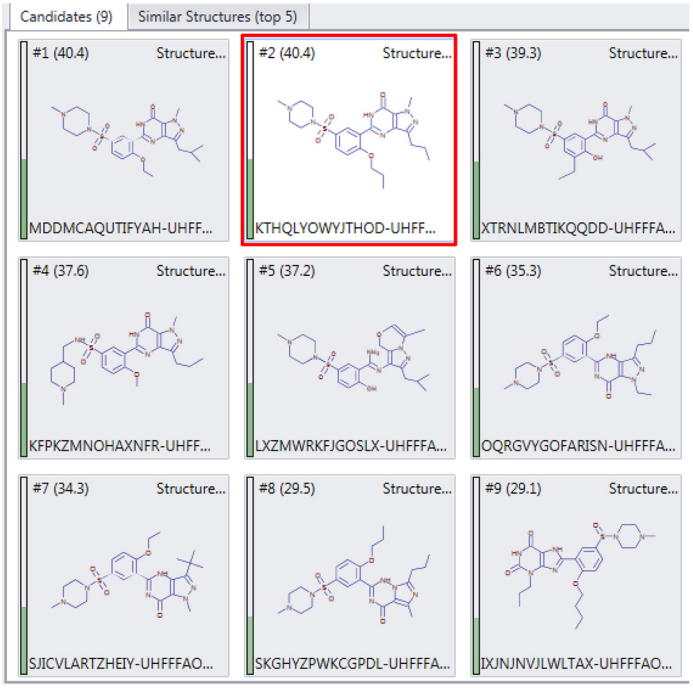


Figure 17. For the *m/z* 489 compound at 6.77 min, the correct structure was ranked #2 and had the same mzLogic score as rank #1.

thermo scientific

Conclusions

- Small molecule unknown structure elucidation is a very time- and resource-consuming task. With many in-house standard compounds, building a curated and annotated HRAM MS^n spectral library can be a valuable tool for compound identification and identifying true unknowns that are similar or share common substructures to the compounds in the library.
- mzLogic can utilize the fragmentation spectra knowledge in the spectral library to quickly narrow down the list of possible structure candidates for further validation. It is an innovative approach for unknown structure characterization utilizing the latest library searching technology.

Acknowledgements

We would like to thank Health Canada for supplying the sildenafil standard compounds.

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

ThermoFisher SCIENTIFIC