

Compounding insights for small molecule research

Thermo Scientific Compound Discoverer software

Authors

Tim Stratton,
Thermo Fisher Scientific, Austin, TX
Ralf Tautenhahn,
Thermo Fisher Scientific,
San Jose, CA

Keywords

[Compound Discoverer software](#),
small molecule data analysis,
[Orbitrap technology](#)

Introduction: challenges in small molecule data analysis

Thermo Scientific™ Orbitrap™ mass spectrometers produce information-rich data. The challenge for small-molecule analysis is to efficiently extract high-confidence understanding from this data, particularly determining and identifying components of interest in chemically complex samples. When performing small molecule research, scientists often have to manually process and integrate their data using multiple third party applications or rely on third party experts.

[Thermo Scientific™ Compound Discoverer™ software](#) is the solution. With a comprehensive, integrated set of libraries, databases, and statistical analysis tools able to link in customizable workflows, the software streamlines unknown identification, determination of real differences between samples, and elucidation of biological pathways. Compound Discoverer software is the only small-molecule analysis solution able to make full use of the rich high-resolution accurate-mass (HRAM) data produced by [Orbitrap mass spectrometers](#).

This white paper describes how Compound Discover software streamlines small molecule research with easily customizable workflows that include powerful unknown identification capabilities, advanced statistical tools, and links to biological databases. All data processing is specified and results reviewed using one state-of-the-art application—Compound Discoverer software.

Workflow wizard and templates accelerate startup

Compound Discoverer software is data processing workflow software. To ensure new users get started quickly, the Study and Analysis Wizard guides them through a simple six-step setup of a data processing workflow. The key to a fast startup is step two, where the user can select from an extensive list of workflow templates (Figure 1). The templates cover numerous small-molecule research applications from metabolite identification to food safety, forensic toxicology, and others. Together, the Wizard and templates reduce the number of mouse clicks needed to specify a

Applications

- Metabolomics
- Metabolite identification
- Metabolic flux analysis
- Environmental research
- Food safety including food contact materials
- Pharmaceutical safety including impurities, degradants, extractables, and leachables
- Forensic toxicology
- Sports anti-doping
- Explosives

data processing workflow from dozens to a few. For users wishing to practice or learn more, Compound Discoverer software includes example data sets for many applications.

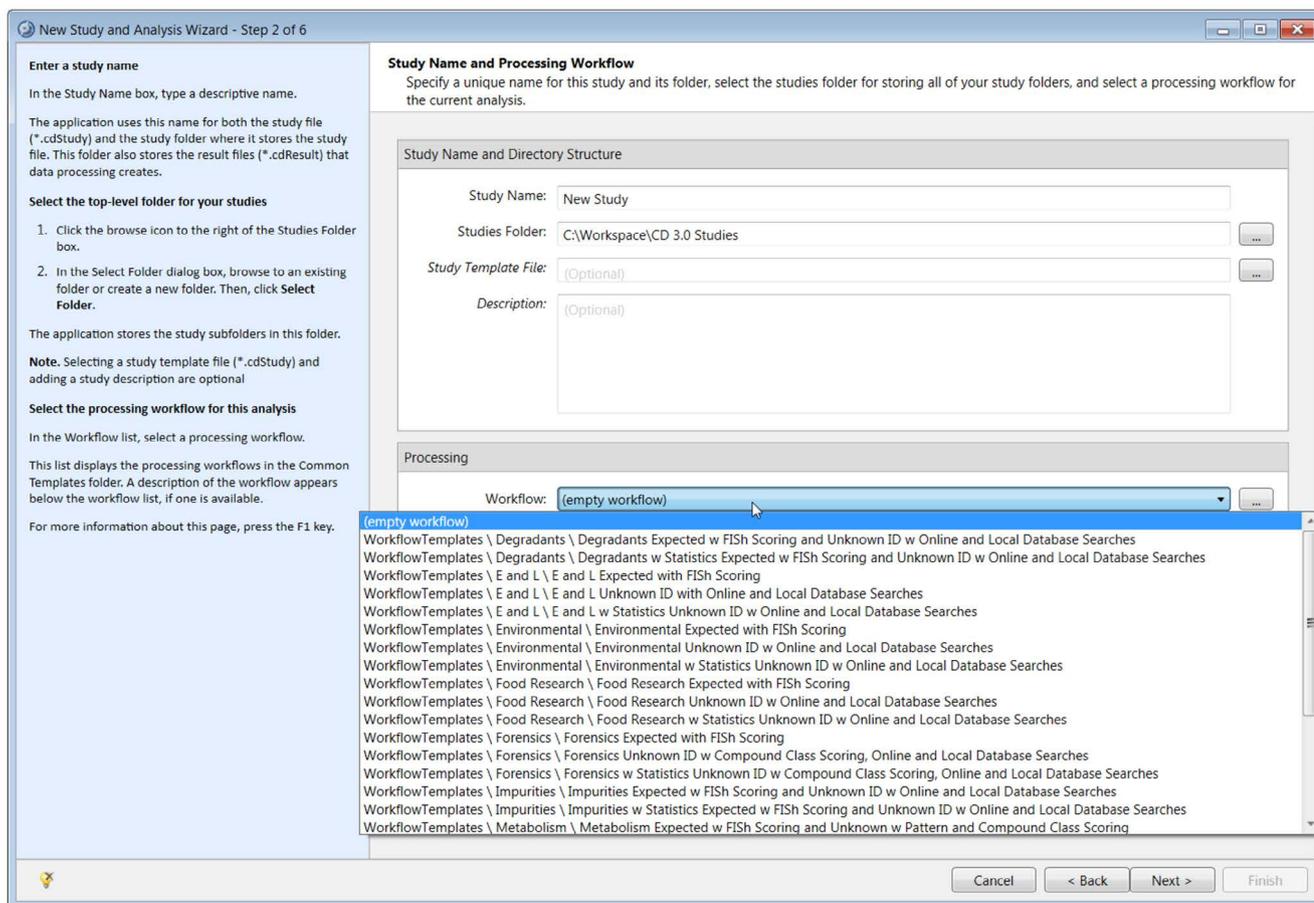


Figure 1. The Study and Analysis Wizard provides an extensive list of templates that can be used to set up a data processing workflow. Workflow templates can be used as is, or modified, renamed, and saved for later use.

Fully flexible, easily customizable, data processing workflows

Compound Discoverer software provides unlimited flexibility, making it easy to design a workflow that matches any small molecule research requirement. The flexibility of the workflows you can create using the software gives you complete control of how data is processed, and with templates, you can reuse any of your methods. Data processing workflows are easily customized by modifying existing templates or by creating new ones. Any template can be renamed, saved, and reused as desired. The template list can be reorganized, for example so you only see templates you routinely use.

Unique to Compound Discover software, customized workflows are easy to build by dragging and dropping workflow nodes—data processing algorithms—into workflow trees that specify desired data flows (Figure 2). Workflow trees make it easy to visualize and modify workflows.

Compound Discoverer software includes an extensive set of data processing nodes. The nodes are color coded to make it easier to see which nodes are related, either because they do similar tasks or because they support one another. Some nodes are applied to the output of others (post processing nodes). Other nodes can take output from many similar nodes and compare it. You can view the parameters for any node by clicking on it.

If the selected nodes know where they should logically connect, the connection arrows appear automatically. If the software knows that nodes don't logically connect, it will not allow the connection or will issue a warning. If the software knows that all of the information needed to perform a step correctly has not been specified, the software will prompt the user to do so.

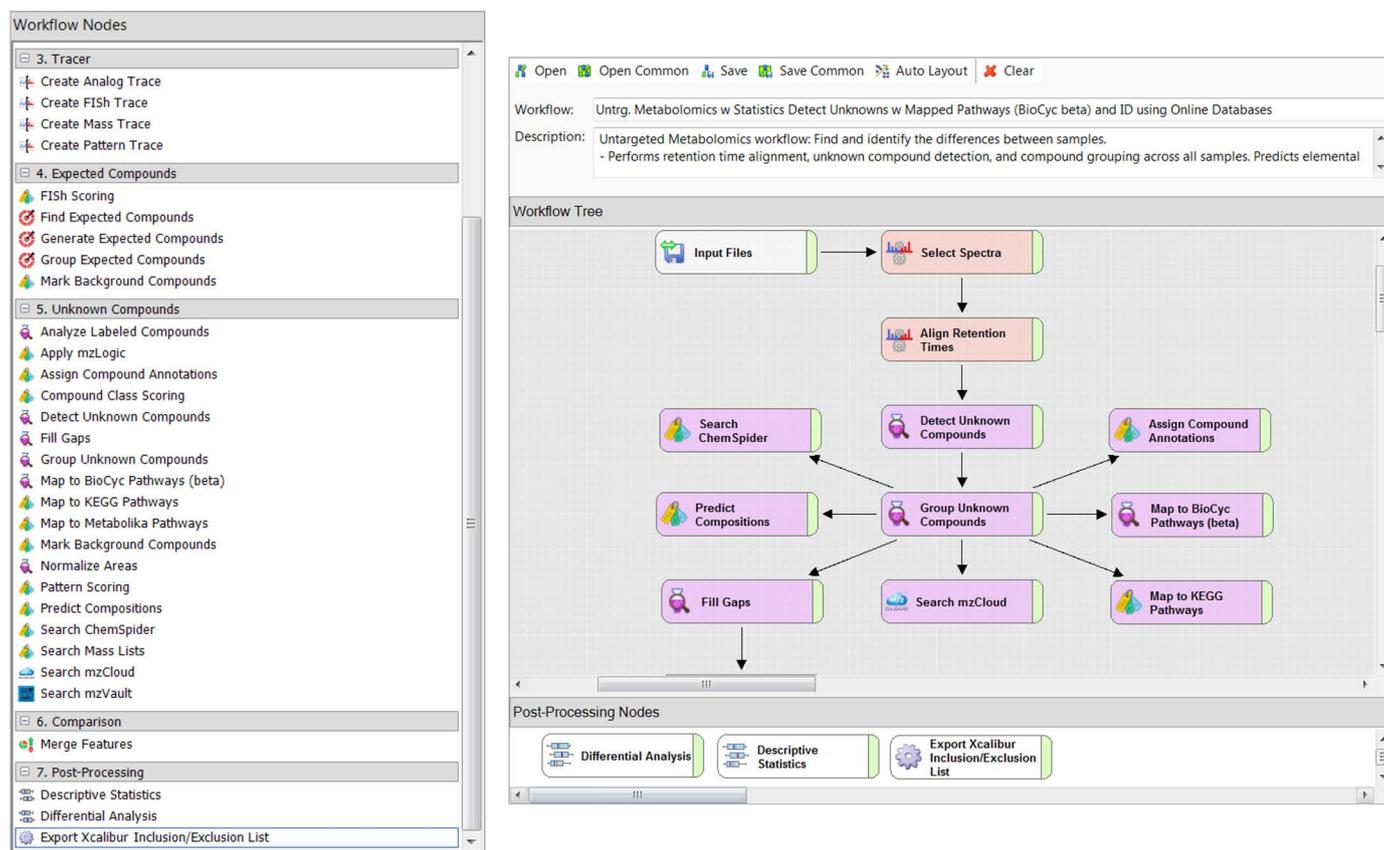


Figure 2. Customized workflows are easy to build by choosing from a comprehensive set of processing nodes and then dragging and dropping them into workflow trees.

Powerful unknown identification capabilities

Unknown identification is one of the toughest challenges in many small molecule applications. By integrating multiple tools to support unknown identification, Compound Discoverer software enables you to identify more of your unknowns—without leaving the software application. Its extensive capabilities include:

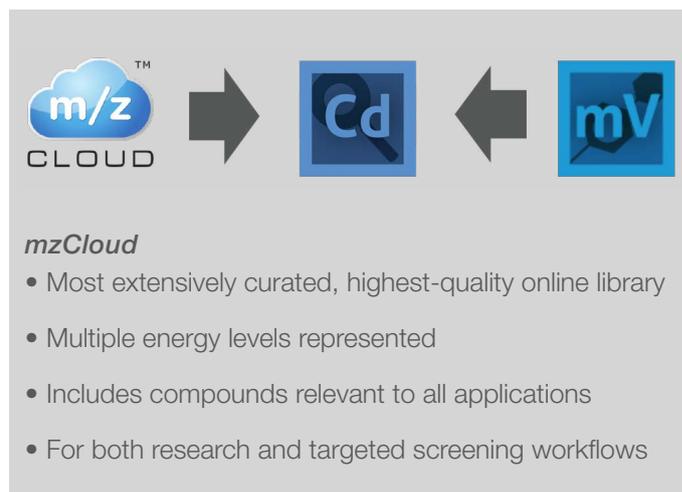
- Spectral library searches of the online mzCloud™ Advanced Mass Spectral library and local Thermo Scientific™ mzVault™ spectral libraries
- Spectral similarity searches of mzCloud
- Online chemical database searches
- Elemental composition determination using HRAM MS data, including fine isotope and MS/MS fragmentation data
- Automated annotation of spectra with predicted fragments

Comprehensive spectral library searches: mzCloud and mzVault

Compound Discover software allows you to search multiple libraries, including the mzCloud and mzVault spectral libraries, in one data processing workflow.

mzCloud is the world's largest (in terms of total spectra and data per compound) and most extensively curated mass spectral fragmentation library. With the largest number of HRAM spectra publicly available, mzCloud includes high-resolution MS/MS and multi-stage MSⁿ spectra arranged into spectral trees. These spectra have been acquired at various collision energies, and for multiple adducts, using collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD). Each raw mass spectrum has been filtered, recalibrated, averaged, and annotated. Unlike other mass spectral libraries, a professional curator evaluates each spectrum before it is incorporated into the library.

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 other identifiers.



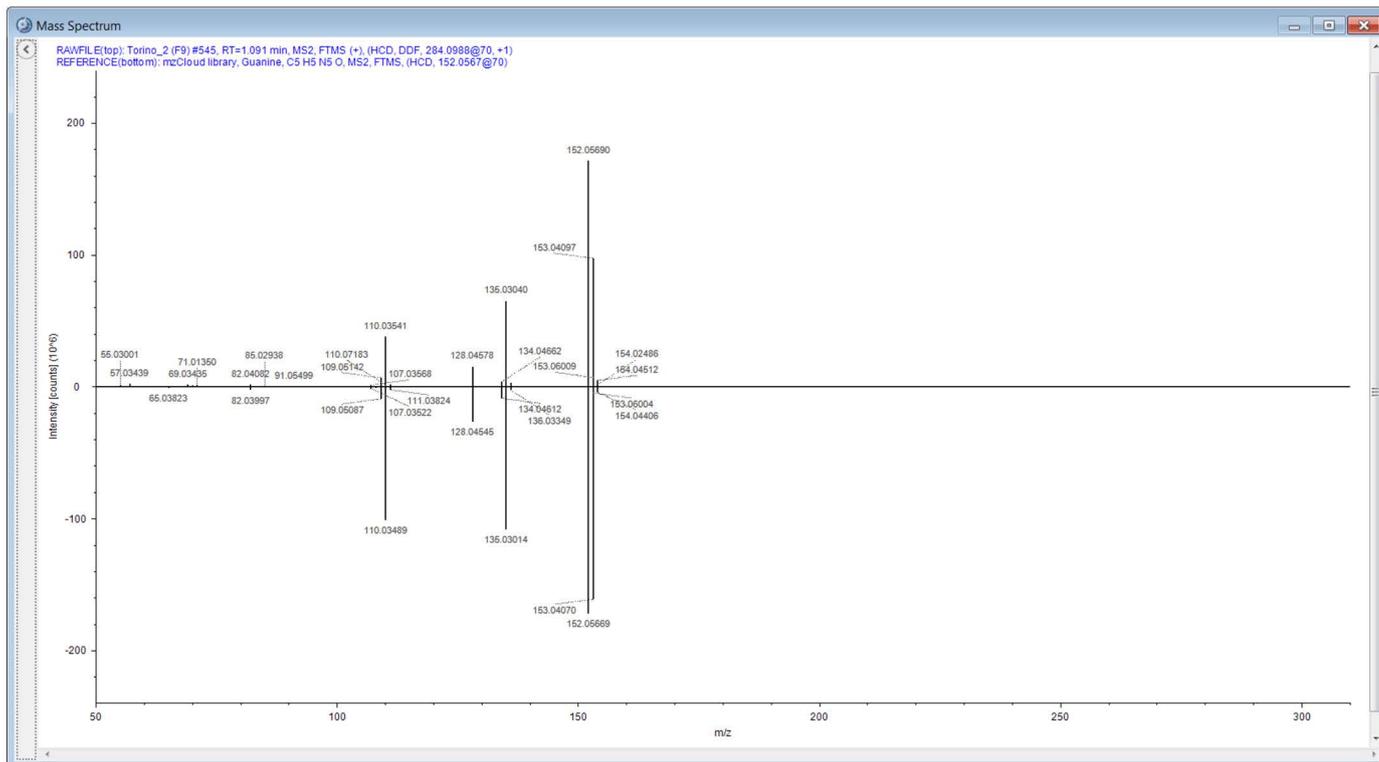
mzVault is a software application for curating and storing high-quality custom in-house MS/MS spectral libraries. Spectra of interest can be directly exported from Compound Discover software to mzVault. Because mzVault ships with the MS²-level spectral data provided in mzCloud, it can be useful when a data system used for data processing is not connected to the Internet.

Database searches alone only produce putative candidates for identification. Using the mzLogic algorithm, putative results can be ranked using the extensive fragmentation spectra in mzCloud.

Similarity and orthogonal database searches yield more identification information

When no good matches result from a spectral library search, structural similarity matches can provide useful structural information. Similarity searches 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.

In the example shown in Figure 3, an unknown compound in a river water sample did not yield an identity match with any spectra in mzCloud. The similarity search produced a very good match with guanine (Figure 3, top), with a mass difference of approximately 132, which corresponded to the molecular formula C₅H₈O₄.



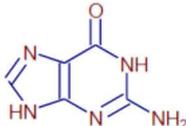
Type	Structure	Name	Formula	Molecular Weight	Δ Mass [Da]	mzCloud ID	Scan #	Best Sim. Match
Similarity		Guanine	C5 H5 N5 O	151.04941	-132.04211	436	545	97.3

Figure 3. Top: Peak and associated spectra from a river water sample; bottom: spectral similarity result for the unknown peak

The molecular formula of the peak was searched against ChemSpider™, resulting in a list of possible structures ranked by number of references (Figure 4). Top ranked is guanosine, which is guanine with a ribose attached.

Stepped collision energy enhances search result quality

Orbitrap mass spectrometers can be set to acquire data with stepped collision energies. This is important because one single fragmentation energy will not be optimal for all small molecules. mzCloud includes more than ten energy levels for each compound in the library. This ensures that there will be relevant spectra regardless of the energy levels chosen to operate the MS instrument.

Compound Discoverer software searches make use of stepped collision energy scans to model matches using the multiple energy levels of fragmentation data stored

in mzCloud. In this way, Compound Discoverer software makes best use of the stepped collision energy capability of Orbitrap mass spectrometers to enhance the quality of search results.

Figure 5 shows mzCloud search results where the query spectrum (top) is a spectrum acquired at 25, 40, and 55 energies. The reference library spectra (bottom), is not one library spectrum, but an *in silico* averaged representative spectra based on the collision energies of the query data. In other words, when the stepped collision energy query spectrum was sent to the library, the library performed a search for possible candidates, and then pulled spectra from the database at close matching energies to create an average spectra to produce a final match score. In this case, the search used spectra at 30, 40, and 60 energies to create the final reference spectra.

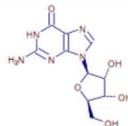
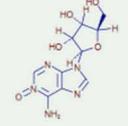
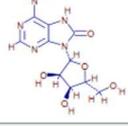
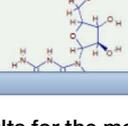
Checked	Compound	Structure	Name	Formula	Molecular Weight	Δ Mass [Da]	Δ Mass [ppm]	CSI
<input type="checkbox"/>	<input checked="" type="checkbox"/>		Guanosine	C ₁₀ H ₁₃ N ₅ O ₅	283.09168	0.00016	0.55	
<input type="checkbox"/>	<input type="checkbox"/>		Adenosine1-oxide	C ₁₀ H ₁₃ N ₅ O ₅	283.09168	0.00016	0.55	10
<input type="checkbox"/>	<input type="checkbox"/>		8-Oxoadenosine	C ₁₀ H ₁₃ N ₅ O ₅	283.09168	0.00016	0.55	
<input type="checkbox"/>	<input type="checkbox"/>		9-(.beta.-D-Xylofuranos	C ₁₀ H ₁₃ N ₅ O ₅	283.09168	0.00016	0.55	

Figure 4. ChemSpider search results for the molecular formula C₁₀H₁₃N₅ displayed in Compound Discoverer software

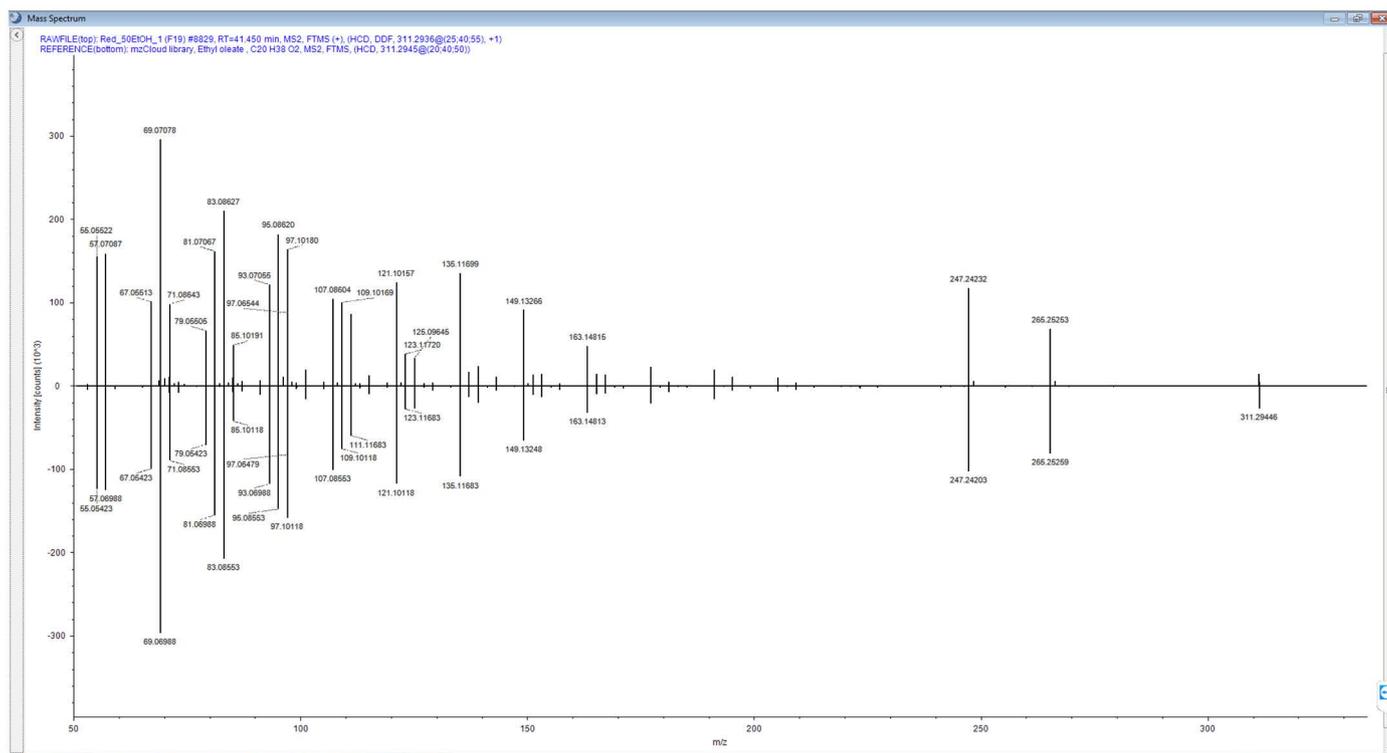
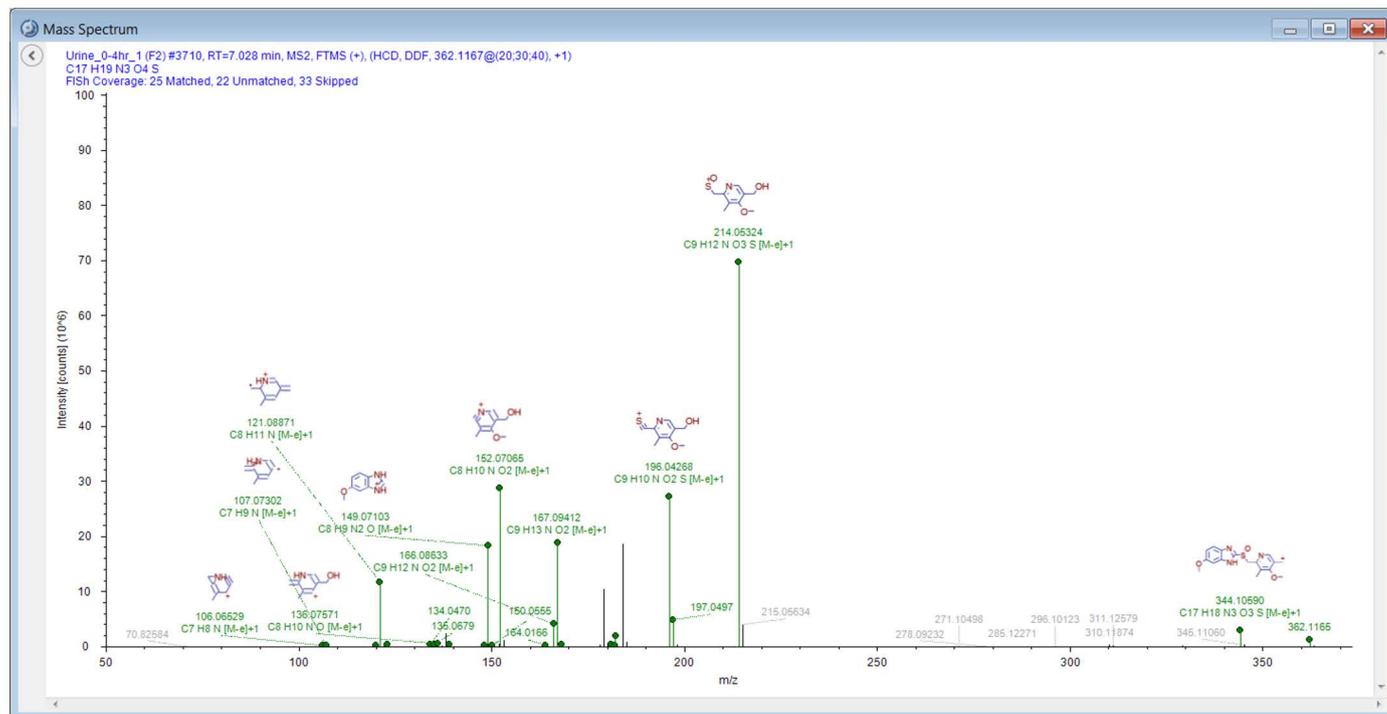


Figure 5. mzCloud search results. The query spectrum (top) is an averaged spectrum of spectra acquired at 25, 40, and 55 energies. The reference library spectra (bottom), is not one library spectrum, but *in silico* averaged spectra based on user specified collision energies.

Structure searches and peak annotations make unknown identification easier

In addition to proposing structures, Compound Discoverer software can annotate structures onto spectra of interest to aid unknown identification (Figure 6). Each peak can be annotated with one or more proposed structures that have been obtained from various sources. The proposed structures are merged directly into the result tables, making them easier to view.

Structure proposals can be used to rank and evaluate the search results generated by an online database. Shown in Figure 7, a ChemSpider search produced over 200 hits. The proposed ChemSpider structures can be ranked by applying Fragment Ion Search (FISH) scoring of *in silico* generated fragments. Compound Discoverer software will generate fragments for each proposed structure with a score based on how well the structure explains the fragments observed in the unknown peak.



Structure	Name	Formula	Molecular Weight	FISH Coverage	Comments
		C17 H19 N3 O4 S	361.10963	57.82	

Figure 6. Structure proposals automatically annotated on an experimental spectrum

Class-based fragment searching rapidly reveals structural analogs

Class-based fragment searches are a revolutionary yet simple approach to detecting and identifying structural analogs in complex samples. These searches enable you to use knowledge of known compounds to find unknown related compounds.

Using local or mzCloud spectral libraries, a compound class fragment list can be created. The Compound Discover software node, “Compound Class Scoring,” can then be applied to score the components for one or more distinct chemical structure classes. Fragmentation is automatically annotated on suspect peaks, enabling putative structure identification. By building multiple lists, it's possible to search for a wide range of specific chemical classes simultaneously.

Figure 8 shows class-based search results for methylenedioxyphenethylamine analogues. Here, it was suspected that the sample contained Ecstasy-like compounds, so the data were processed to find peaks having fragments common to this compound class. The fragmentation spectra are annotated with the fragments from the compound class fragment list that point to parts of the molecule that are common to known Ecstasy analogs.

Integrated tools turn unknown identifications into knowledge

Compound Discover software supplies a fully integrated suite of tools—KEGG™, BioCyc™, and Thermo Scientific™ Metabolika™ biological pathway databases, advanced statistics, hypothesis testing, flexible visualization

of complex data, and automated normalization and retention time alignment—that take you far beyond compound identification, enabling you to advance your understanding of complex systems. These tools build on the certainty provided by the identification and structural elucidation capabilities of Compound Discoverer software.

Advanced statistics illuminate real differences

The powerful Compound Discoverer software toolset enables you to quickly find real statistical differences between sample sets. You can view trends in components across a study or identify the key components of interest between multiple sample groups, all using interactively linked displays. Interactivity allows you to move between from high-level views and individual component views using one mouse click.

In addition to unsupervised statistical analysis tools such as principal component analysis (PCA) and differential analysis with volcano plots, Compound Discoverer software provides supervised statistical tools such as partial least squares-discriminant analysis (PLS-DA). PLS-DA is useful when you know there is a difference between two or more sets of samples, and you are looking for the components that account for most of the difference.

Figure 9 shows a PLS-DA plot pointing to the five compounds most important in describing the difference between the sample sets. Like all Compound Discoverer software displays, the PLS-DA plot is fully interactive. You can click on the components and view the peaks in the data set.

Checked	Name	Formula	Annotation Sc	Molecular Weight	RT [min]	Area (Max.)	Class Coverage	
<input type="checkbox"/>		C10 H10 O2	■■■■■	162.06786	2.629	505319	54.55	33.33
<input type="checkbox"/>		C11 H15 N O2	■ ■ ■ ■ ■	193.11016	2.784	3052442	45.45	33.33
<input type="checkbox"/>	3-(4-Methylbenzoyl)acrylic acid	C11 H10 O3	■ ■ ■ ■ ■	190.06272	4.196	1960786	45.45	33.33
<input type="checkbox"/>	3,4-Methylenedioxyamphetamine (MDA)	C10 H13 N O2	■ ■ ■ ■ ■	179.09438	2.423	258756	36.36	0.00
<input type="checkbox"/>	UNII:734CNR85EV	C13 H14 O4	■ ■ ■ ■ ■	234.08872	3.353	268815	36.36	33.33
<input type="checkbox"/>	Safrole	C10 H10 O2	■ ■ ■ ■ ■	162.06786	2.426	113567	36.36	0.00
<input type="checkbox"/>	3,4-Methylenedioxyamphetamine (MDA)	C10 H13 N O2	■ ■ ■ ■ ■	179.09438	2.629	1067070	36.36	0.00

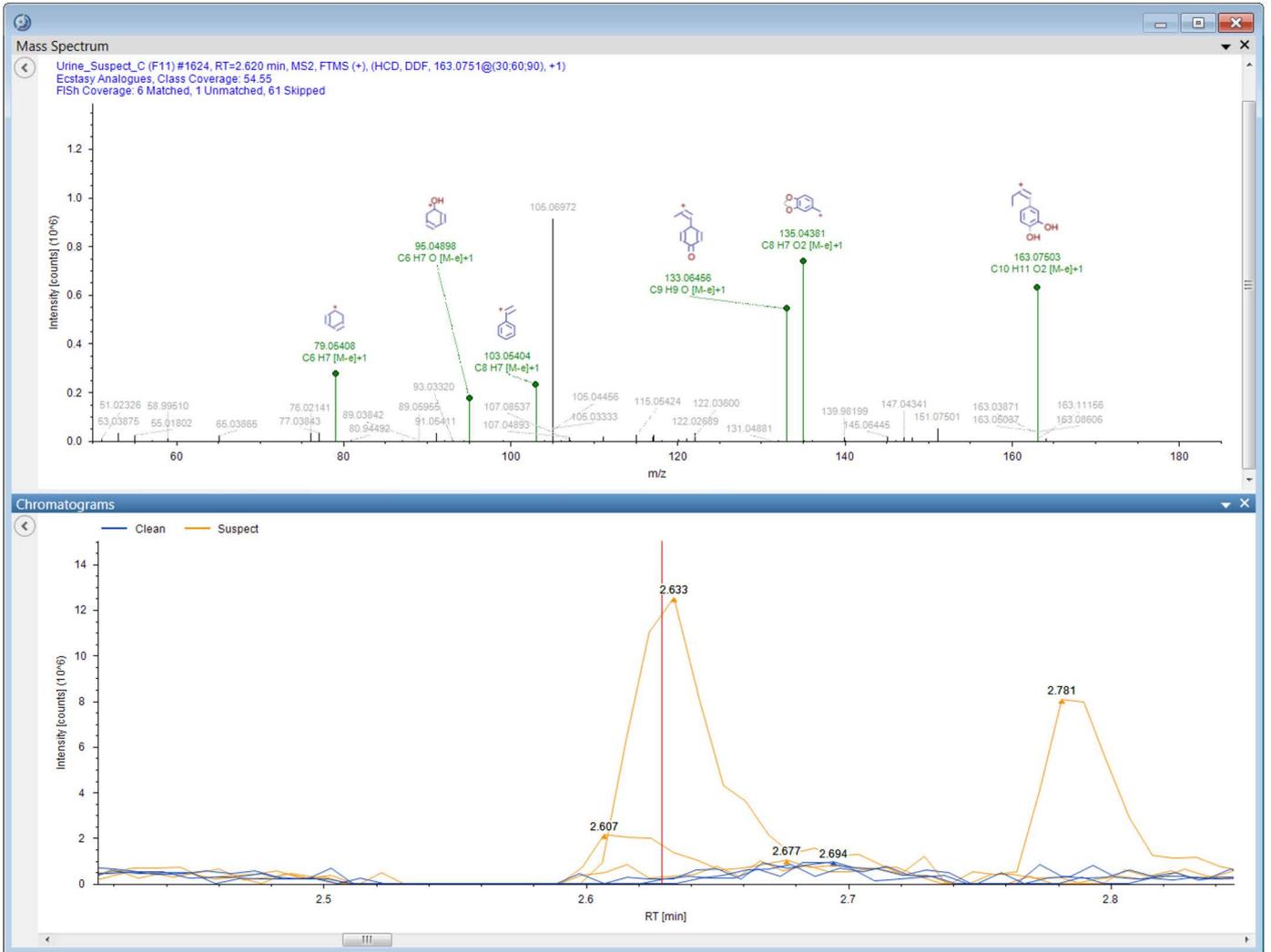


Figure 8. Class-based fragment search results for methylenedioxyamphetamine analogs. Top: results table and class list. Middle: fragmentation spectra for the peak at RT 2.62 annotated with the fragments from the Ecstasy analog compound class fragment list. Bottom: chromatogram.

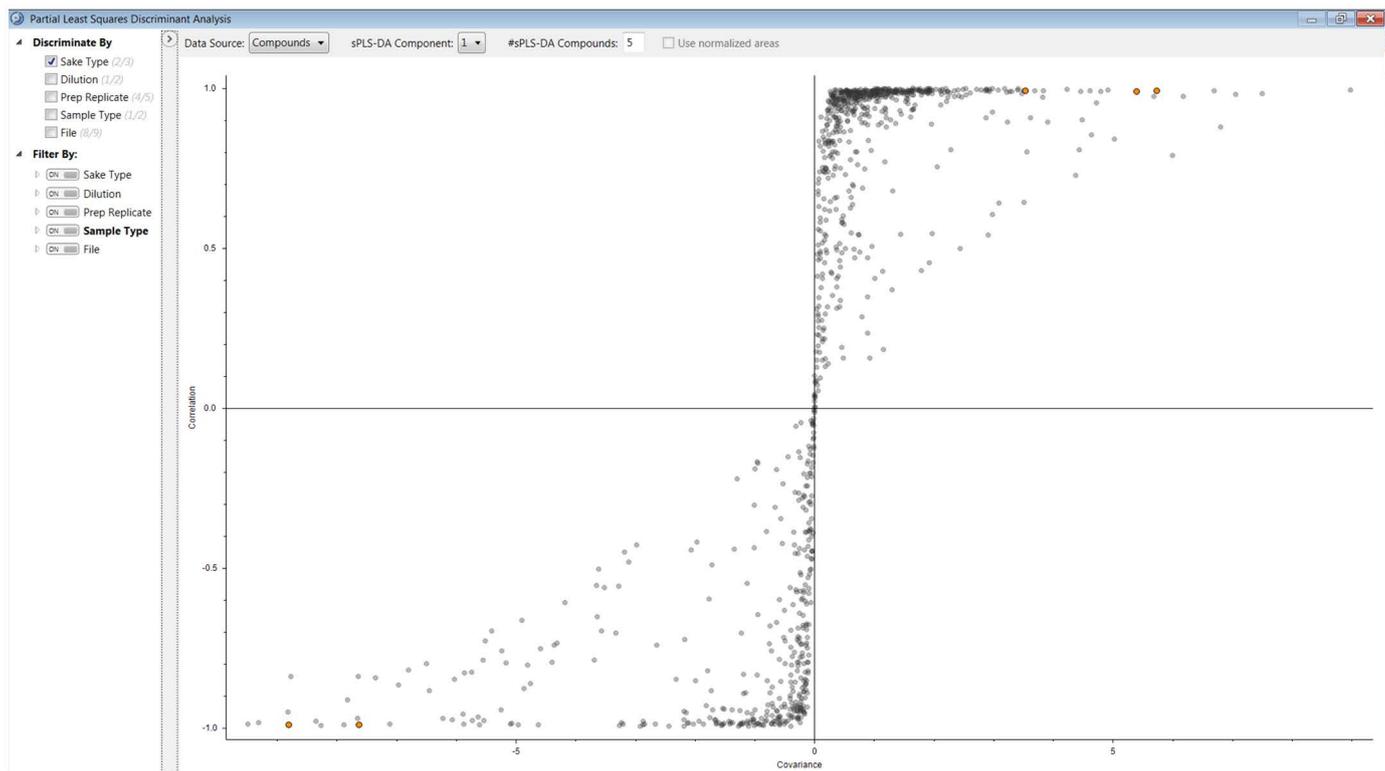


Figure 9. PLS-DA plot. The five compounds most important in driving the difference between sample sets are shown in orange.

Statistical tools

- Interactive differential analysis; PCA; ANOVA; volcano, box-whisker, and trend line plots; PLS-DA; and *t*-tests
- Heat-map-like tables for peak areas, ratios, two-fold log change, CV values, *p* and adjusted *p*-values.
- Interactive heat-maps with hierarchical clustering
- Isotopologue distribution charts

Nested study design improves study factors

Compound Discover software supports nested study designs in which one subject is used as a control for itself, for example when a study aims to profile the same subject with and without a treatment. Figure 10 shows an example of a nested study design. Compound Discoverer software automatically detects a nested study when the study factors are entered and will treat it as such during data processing.

Metabolic flux workflows and Metabolika suite streamline stable isotope labeling flux analysis

Stable isotope labeling (qualitative flux analysis) experiments are extremely useful in understanding metabolic pathways. Compound Discoverer software makes full use of HRAM Orbitrap mass spectrometer data to offer advanced, streamlined untargeted stable isotope label analysis. Using an unlabeled reference sample, Compound Discoverer software automatically detects all compounds, determines their elemental composition, and identifies their labeled counterparts in labeled samples. Fractional label incorporation (exchange rate) after natural abundance correction is reported for each compound. Any isotopic label can be used, and the label is the only user-entered parameter required for automatic isotopologue detection. Isotopologue distribution and trend line charts provide an easy way to visualize results.

The exchange rate and other statistical data can be overlaid onto pathways using the Metabolika suite of pathway tools that are integrated with Compound Discoverer software. The Metabolika suite not only includes high-quality curated pathways; it also allows you to build your own and to visualize your data on them.

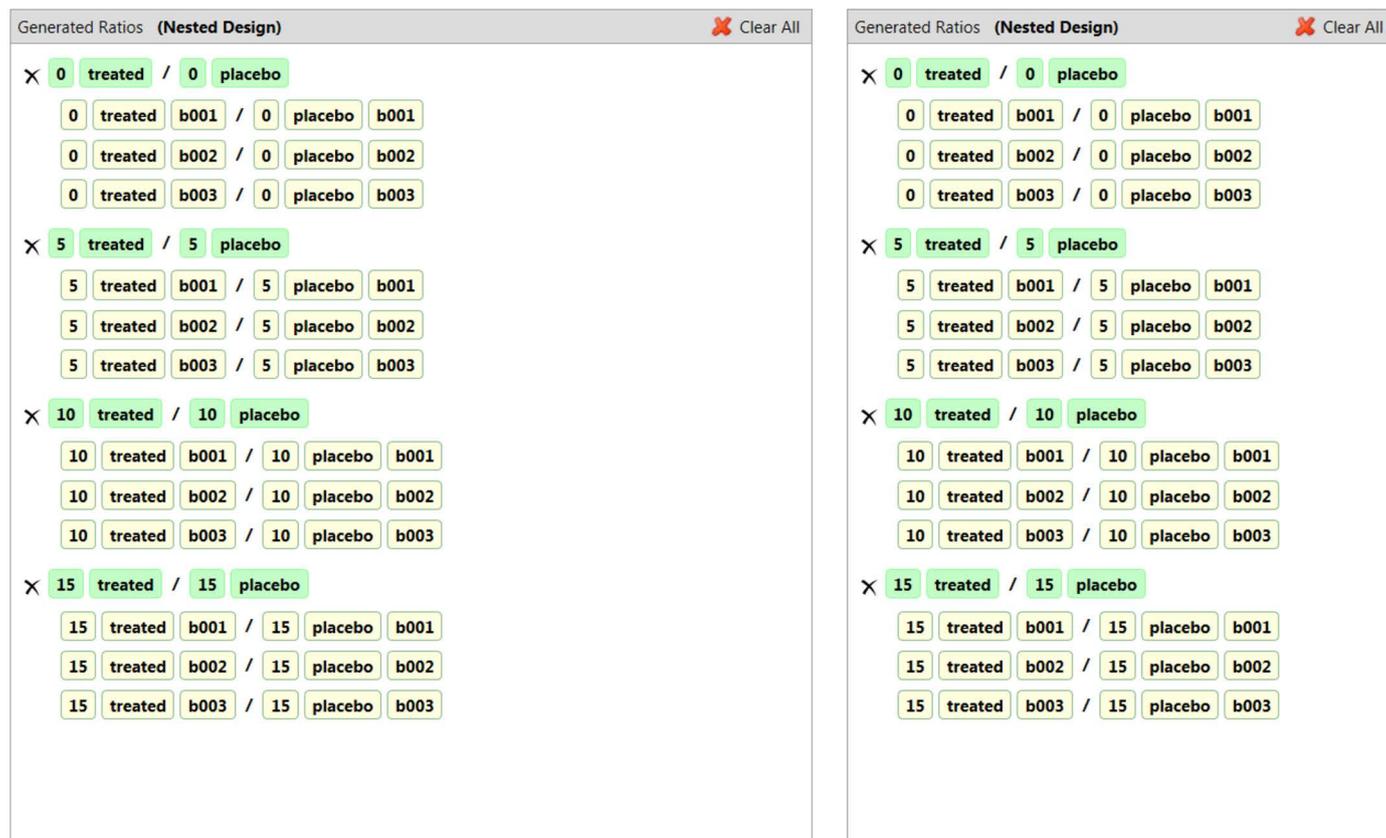


Figure 10. Nested study design

Superior normalization approach

Across large data sets, injection-to-injection variation in LC performance and MS response is common. Normalization is typically performed to account for this variation. Because every chemical exhibits a unique pattern of fluctuation in response over time, normalizing data based on total signal is not very useful.

Compound Discover software applies a superior normalization approach developed by leaders in the metabolomics community.¹ Similar to the approach used in quantification, Compound Discover software normalizes every chemical entity against itself. If all samples are pooled together to create a quality control (QC) sample, that QC sample will contain every chemical entity. The QC samples are injected regularly, allowing every individual peak—every compound in the sample—to be normalized against itself. When Compound Discoverer software is instructed which injections are the QC injections, it performs quality metrics to determine if

a component is present in enough of the QC samples to allow normalization. If it is, then Compound Discoverer software goes ahead and does the normalization.

Advanced retention time alignment and visualization saves time

An important problem for many large-scale studies is aligning retention times over hundreds or even thousands of injections over many days. Compound Discoverer software solves this problem by sampling as many features (chromatographic peaks) as possible across all raw files. These features serve as landmarks that allow related files to be aligned using a multi-step consensus approach. "Alignment-by-group," chooses a representative from each sample group. Next, "alignment-across-groups," seeks consensus across groups. In this way, all of a raw file's components across the entire retention time studied are aligned against their group first, and then against the entire study. Raw file alignment and outlier files are easy to visualize in plots (Figure 11).

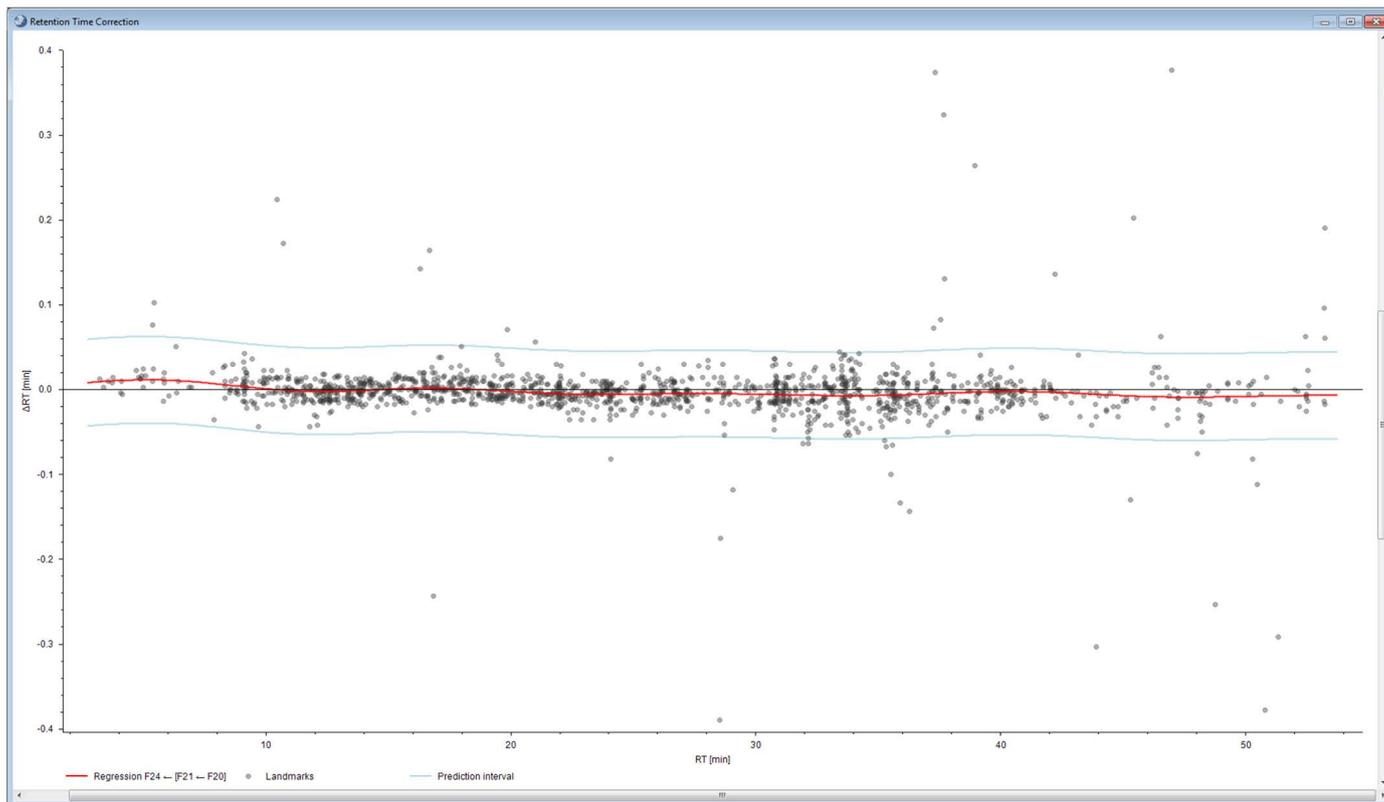


Figure 11. Plot of retention time alignment results. File 20, is aligned against File 21, and File 21 then becomes the representative of the group used to align the entire study.

Data export options eliminate manual steps

Transforming the results of small molecule research into subsequent analyses, for example for targeted work, can involve tedious manual steps. Setting up an instrument to trigger all of the necessary transitions needed to obtain high-quality fragment spectra can be tricky. Compound Discoverer software eliminates manual data entry with data export tools that allow automated export of detected peaks to targeted inclusion or exclusion lists. For example, the expected transformations can be used to create and export exhaustive target inclusion lists, or untargeted peak detection results can be used to create better MS/MS acquisition methods.

Compared to manual approaches, it is easier to use Compound Discover software to determine all the components of interest in initial sample injections and to click on the results for export out in the desired instrument format (Figure 12).

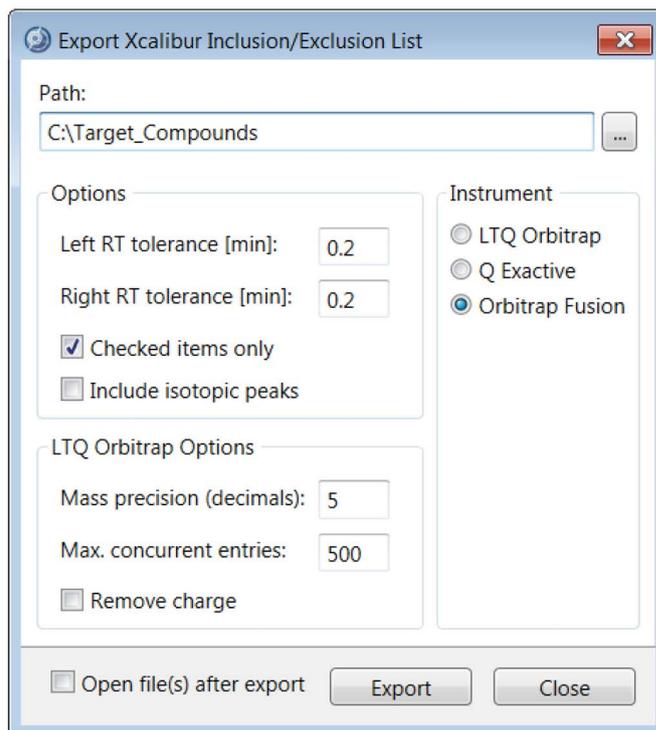


Figure 12. Compound-specific lists for each instrument, ready to use, are easily exported from Compound Discoverer software to Thermo Scientific™ Xcalibur™ software.

mycompounddiscoverer.com speeds learning

Because Compound Discoverer software is a one-stop solution for small molecule research, it is the only software you need to learn to perform your MS data processing. Though the workflow templates provide quick startup for nearly every application, mycompounddiscoverer.com is the place to go to view training videos covering specific portions of the software. The videos walk you through numerous examples using the example data sets provided with Compound Discoverer software. If you need more, use the mycompounddiscoverer.com HELP BUTTON to get help from a Thermo Fisher Scientific expert.

Conclusion

The only small-molecule analysis solution able to make full use of the rich high-resolution HRAM data produced by Orbitrap mass spectrometers, Compound Discoverer software allows you to create flexible workflows with an enormous degree of functionality and customizability. Within one software application, a comprehensive, integrated set of libraries, databases, and statistical analysis tools link in workflows that streamline small molecule research involving unknown identification, determination of differences between samples, and interpretation of biological pathways.

Reference

1. Dunn, W.B., et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **2011**, Jun 30; *6*(7):1060–83.

Learn more at thermofisher.com/compounddiscoverer