Scaffold Elements

Version 3.0

User’s Guide

Release Information

The following release information applies to this version of the *Scaffold Elements*. This document is applicable for Scaffold Elements, Release 3.0 or greater, and is current until replaced.

Copyright

© 2021. Proteome Software, Inc., All rights reserved.

The information contained herein is proprietary and confidential and is the exclusive property of Proteome Software, Inc.. It may not be copied, disclosed, used, distributed, modified, or reproduced, in whole or in part, without the express written permission of Proteome Software, Inc.

Limit of Liability

Proteome Software, Inc. has made its best effort in preparing this guide. Proteome Software, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this guide and specifically disclaims any implied warranties of merchantability or fitness for a particular purpose. Information in this document is subject to change without notice and does not represent a commitment on the part of Proteome Software, Inc. or any of its affiliates.The accuracy and completeness of the information contained herein and the opinions stated herein are not guaranteed or warranted to produce any particular results, and the advice and strategies contained herein may not be suitable for every user.

The software described herein is furnished under a license agreement or a non-disclosure agreement. The software may be copied or used only in accordance with the terms of the agreement. It is against the law to copy the software on any medium except as specifically allowed in the license or the non-disclosure agreement.

Trademarks

The name *Proteome Software*, the Proteome Software logo, *Scaffold, Scaffold Q+*, *Scaffold Q+S*, *Scaffold LFQ, Scaffold PTM, Scaffold DIA* and *Scaffold Elements* logos are trademarks or registered trademarks of Proteome Software, Inc. All other products and company names mentioned herein may be trademarks or registered trademarks of their respective owners.

Customer Support

Customer support is available to organizations that purchase *Scaffold, Scaffold Q+, Scaffold Q+S, Scaffold LFQ,Scaffold PTM, Scaffold DIA or Scaffold Elements* and that have an annual support agreement. Contact Proteome Software at:

Proteome Software, Inc.
1340 SW Bertha Blvd
Suite 10
Portland, OR 97219
1-800-944-6027 (Toll Free)
1-928-244-6024 (Fax)
www.proteomesoftware.com

-

Table of Contents

Chapter 1: Getting Started with Scaffold Elements 5

Chapter 2: Identifying Compounds with Scaffold Elements 19

Chapter 3: Loading data in Elements 31

Chapter 4: Scaffold Elements Main Window 45

Chapter 5: The Organize View 71

Chapter 6: The Samples View 91

Chapter 7: The Analytes View 107

Chapter 8: The Visualize View 121

Chapter 9: The Library View 133

Chapter 10: Analysis View 139

Chapter 11: The Publish View 141

Chapter 12: Metabolomic Flux Analysis 147

Chapter 13: Quantitative Methods and Tests 151

Chapter 14: Reports 165

Appendix 169

Appendix A. Creating A Personal Spectral Library 170
Appendix B. Creating a Custom Spectral Library using a tab-delimited text file ...211
Appendix C. Elements Scoring Algorithms 213
Appendix D. Rolling up Values 220

-

Appendix E. Agglomerative Point Clustering Feature Finding Algorithm 223
Appendix F. Isotopic Clustering 231
Appendix G. Forming Consensus MS1 peak groups 234
Appendix H. Exporting a Transition List to Skyline 237
Appendix I. Structure of Scaffold Elements files (*.metdb) 240
Appendix J. Terminology 243
Appendix K. Heat map clustering 246
Appendix L. Techniques to Control the Family-wise Error Rate 249
Appendix M. Using Principal Component Analysis in Scaffold Elements 250
Appendix N. How PCA is Performed in Scaffold Elements 260
Appendix O. Description of Mouse Right Click Context Menu Commands 267

Preface

Welcome to the Scaffold Elements User's Guide. Its purpose is to answer users' questions and guide them through the procedures necessary for using Scaffold Elements efficiently and effectively.

Using the manual

A Table of Contents and an Index are provided in this manual for the user's convenience. This Preface also provides a brief discussion of each chapter to further assist users in locating needed information.

Special information about the manual

This User's Guide has a dual-purpose design. It can be distributed electronically and printed on an as-needed basis, or it can be viewed on-line in its fully interactive capacity. If users print the document, for best results it is recommended that they print it on a duplex printer; however, single-sided printing is also possible. When the document is viewed on-line, a standard set of bookmarks appears in a frame on the left side of the document window for navigation through the manual. For better viewing, users can decrease the size of the bookmark frame and use the magnification box to adjust the display according to their viewing preferences.

Conventions used in the manual

The *User's Guide uses the following conventions:*

- Information that can vary in a command—variable information—is indicated by alphanumeric characters enclosed in angle brackets; for example, <Analyte Name>.
- A new term, or term that must be emphasized for clarity of procedures, is *italicized*.
- Page numbering is "on-line friendly." Pages are numbered from 1 to x, *starting with the cover* and ending on the last page of the index.
- This manual is intended for both print and on-line viewing.
- If information appears in blue, it is a hyperlink. Table of Contents and Index entries are also hyperlinks. Click the hyperlink to advance to the referenced information.
- A sample set of Demo data, available for download from *http://www.proteomesoftware.com/products/demo-data* is used as the basis for most screen captures, examples, and data manipulations that are shown in the manual.

-

Chapter 1
Getting Started with Scaffold Elements

System Requirements

For information about the system requirements for Scaffold Elements, see:

https://support.proteomesoftware.com/hc/en-us/articles/213578086-Scaffold-Software-System-Requirements

Installing Scaffold Elements

Scaffold Elements runs on Windows, MAC or Linux systems. Follow these instructions to install the application on your system:

Request an evaluation by filling in the form found at . You will receive download instructions and a license key to activate the software via email.

2. Download and launch the installation executable.
3. Carefully follow the instructions provided in the installation wizard, accepting the user agreement when prompted and moving through the screens by clicking Next.

-

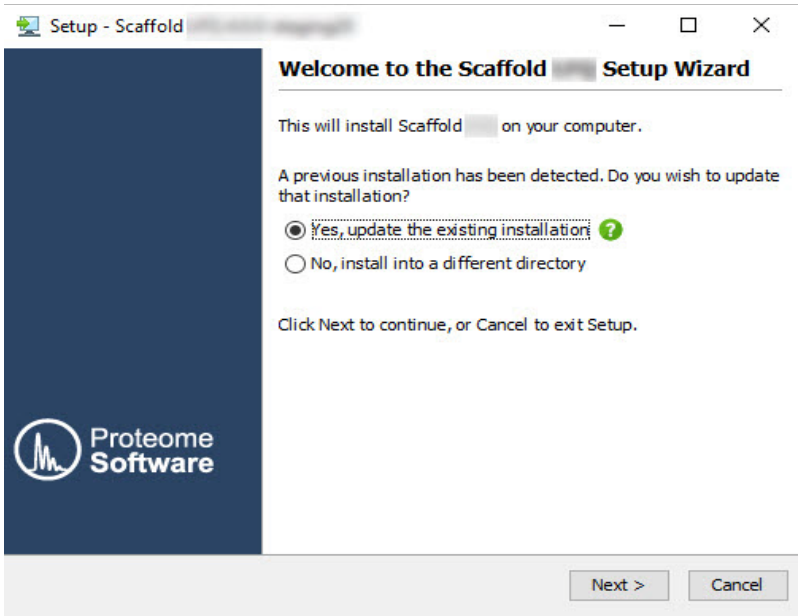


Figure 1-1: Scaffold Elements installation Setup Wizard

4. In order to process raw data, Scaffold Elements requires a working installation of MSConvert, which is an application included in the open source cross-platform tool kit **ProteoWizard**[1]. During the initial installation of Scaffold Elements, the installer will prompt the user to download and install ProteoWizard. If you already have MSConvert installed on your system and the version number is greater than or equal to the indicated version, you may skip this step and proceed to step 5.

5. If a ProteoWizard installation is not yet available on your computer or the proper version is not installed, click the button labeled “Download ProteoWizard”. A web browser will open; select the option, “I agree to the licensing terms -download ProteoWizard” and install ProteoWizard.Once you have completed the ProteoWizard installation, return to the Scaffold Elements dialog box to finish the installation of Scaffold Elements.

6. Click “Next”. If Scaffold Elements has located an acceptable version of MSConvert, its location will be displayed. If the box is empty, use the Browse button to locate and select the MSConvert.exe file on your system; generally, this will be located in a subfolder of “C:\Program Files\ProteoWizard”. Click “Next”.

7. The installer will then provide you an opportunity to allocate memory to Scaffold Elements. We recommend that you set the Maximum Memory to approximately 80% of the amount of physical RAM on your system. Click “Next”.

8. You may then select a Start Menu Folder for the application and choose whether or not to create shortcuts for all users of the system. The next screen allows you to set a file association between

1. Chambers, M. et *al*. Nature Biotechnology, 30, 918–920 (2012) doi:10.1038/nbt.2377

-

METDB files and Scaffold Elements, and the following screen allows creation of desktop icons. Clicking “Next” begins the installation.

9. Finally, Scaffold Elements allows you to select the option to have the program open at the closing of the wizard. Click “Finish”..

---

*For better performance you should allocate as much RAM as possible to Scaffold Elements. The memory setting can be adjusted after installation by selecting the menu option **Edit > Preferences - Memory tab**. You must close Scaffold Elements and restart the program in order for the new memory setting to take effect.*

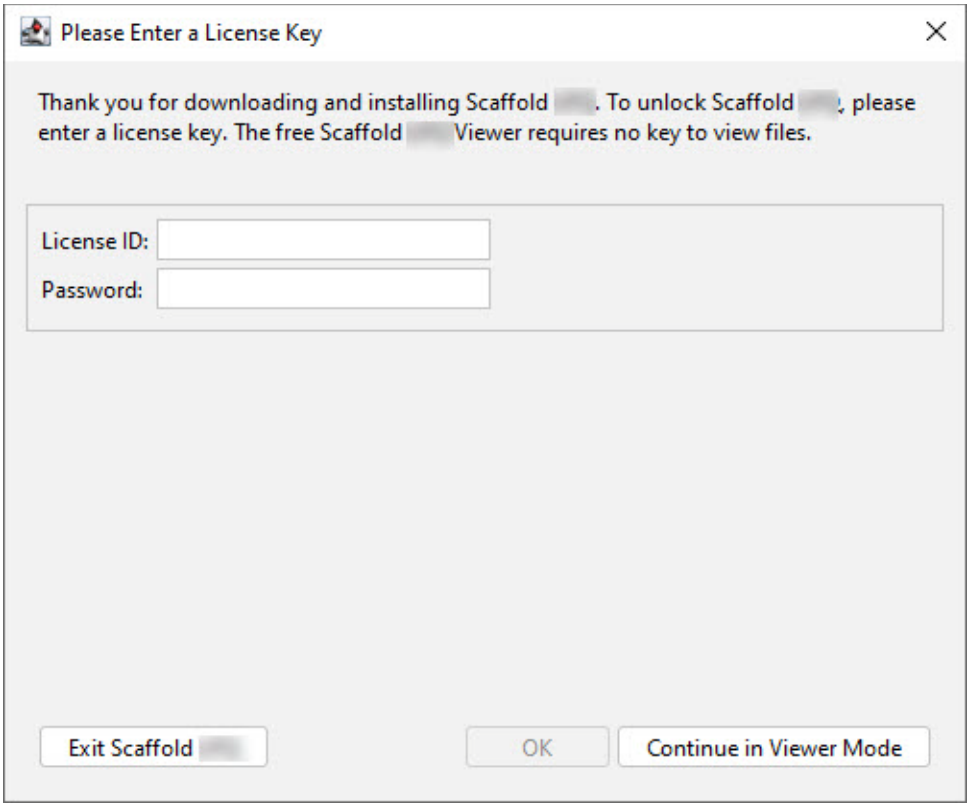
After Scaffold Elements has been installed on a computer, a shortcut icon for the application is placed on the desktop. An option is also available from the Start menu. Double-clicking the desktop icon launches Scaffold Elements, as does, for Windows computers, selecting the option from the Start menu (**Start > All Programs > Scaffold Elements > Scaffold Elements**)

Licensing

The first time Scaffold Elements opens after installation, the Enter License Key dialog box opens.

Keys and passwords may be typed, pasted or dragged into the appropriate fields. Both items may be pasted or dragged together.

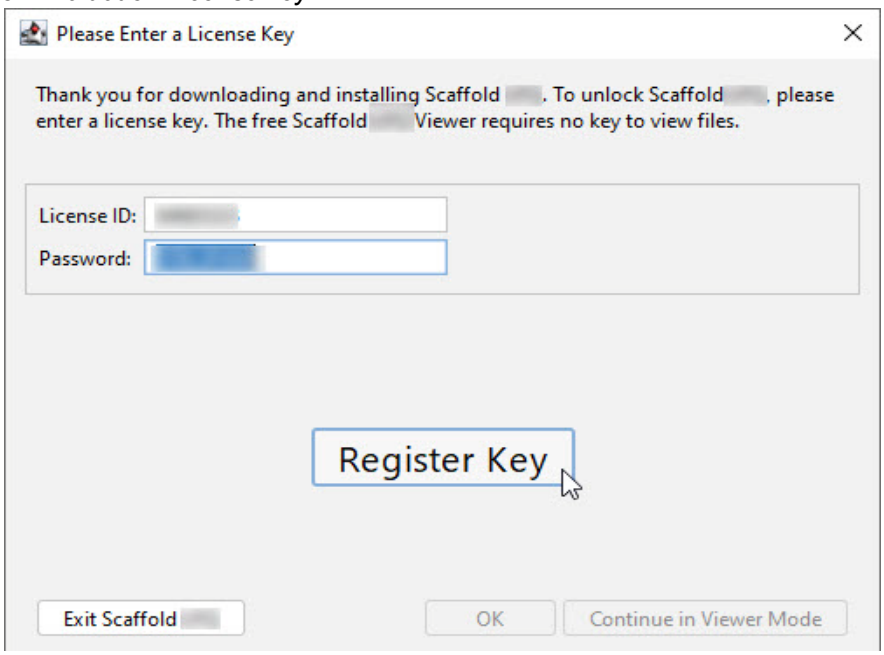
Figure 1-2: Scaffold License Key messages



Two kinds of keys are available to activate the software:

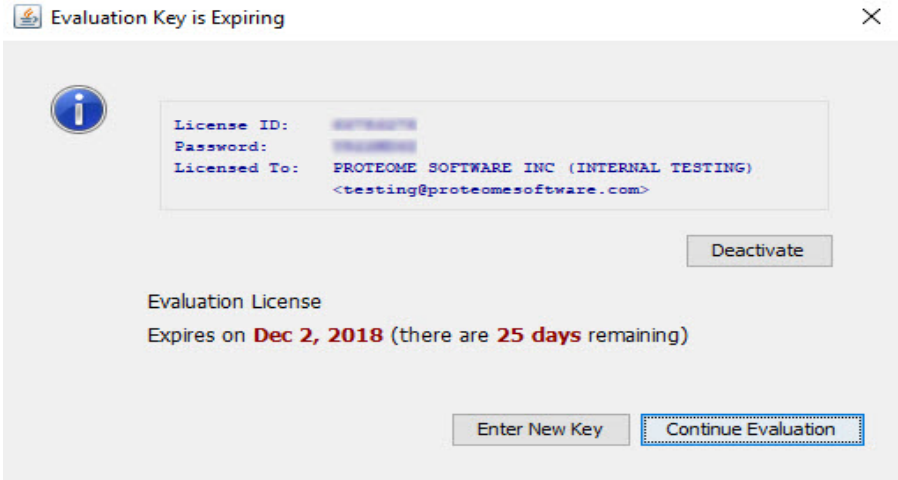
Evaluation key - An Evaluation key is valid for a limited period. A free evaluation key for Scaffold Elements may be obtained through *www.proteomesoftware.com*. An evaluation key may be used on two computers. Once the key and password have been copied and pasted into the license key dialog box, a message will appear below it, displaying confirmation of the key registration. Pressing OK starts the application.

Figure 1-3: Evaluation License key



Every time Scaffold Elements is launched in evaluation mode, a message appears showing the remaining time available for evaluation and offering the option to enter a new key.

Figure 1-4: Message appearing when launching an evaluation copy of Scaffold Elements

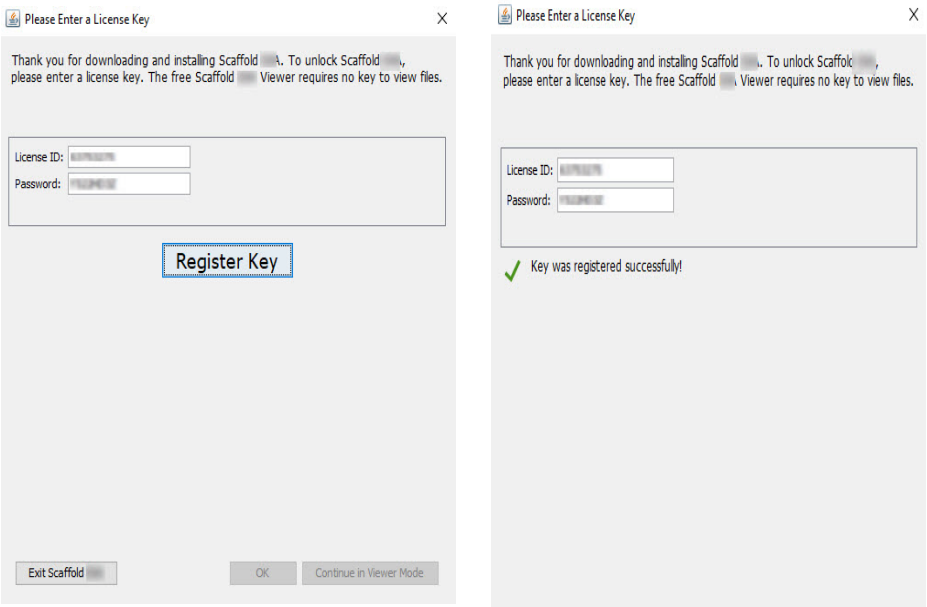


Time-Based License key—a Time-Based License key allows the user to access all features of the software permanently. It only allows upgrades within a certain time limit, however. The time tracks the length of the support contract. Once expired, Scaffold Elements will continue to work beyond the expiration date, but no upgrades are allowed unless the support contract is renewed.

Contact *sales@proteomesoftware.com* to purchase the appropriate key.

A Time-Based License key is valid only for a single computer. If it is necessary to move the Scaffold Elements installation to a different computer, see for instructions to transfer the key at no charge.

Figure 1-5: Time-Based License key



When the Time-Based License key and password are entered, pressing **Register Key** verifies their validity and a message appears describing the status of the key.

Once the key is successfully registered, pressing OK closes the dialog box and a Scaffold Elements Welcome message opens.

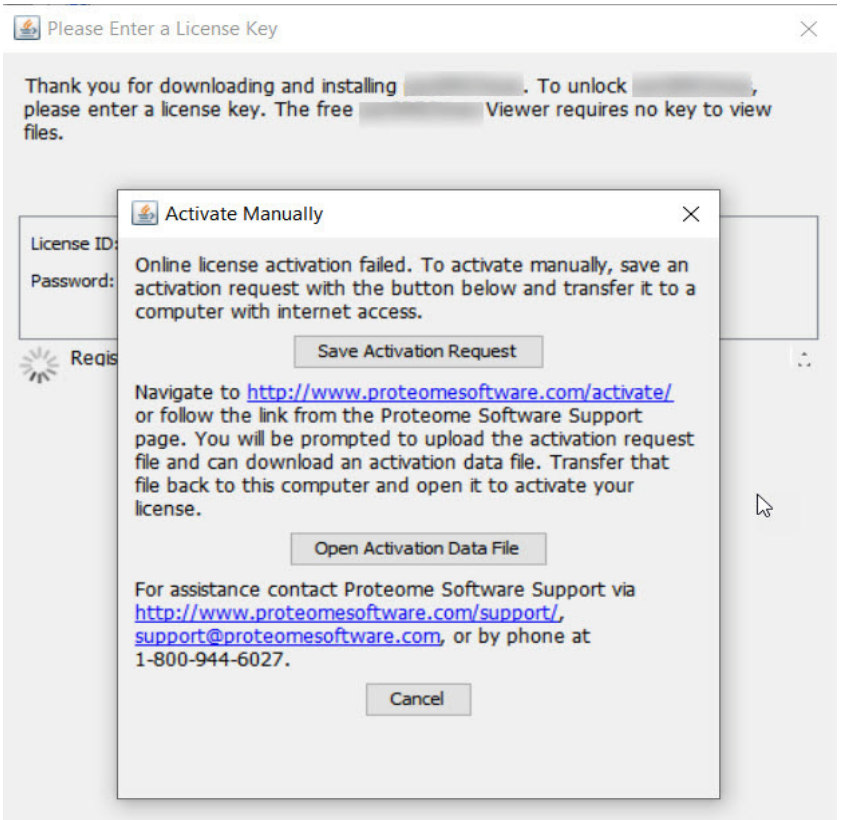
If the user is using an evaluation copy of Scaffold Elements, then an Evaluation message opens, indicating the number of days left in the evaluation period. The user must click OK to close this message and then the Scaffold Elements Welcome message opens.

From this window, the user may create a new experiment, open an existing experiment (*.METDB file), or work with the demonstration data that is provided in the Scaffold Elements installation.

Registering a Time-Based License key with no INTERNET connection

When a Time-Based License key is entered and the Register Key button is pressed, but no INTERNET connection is available, a dialog appears, providing instructions for manual activation.

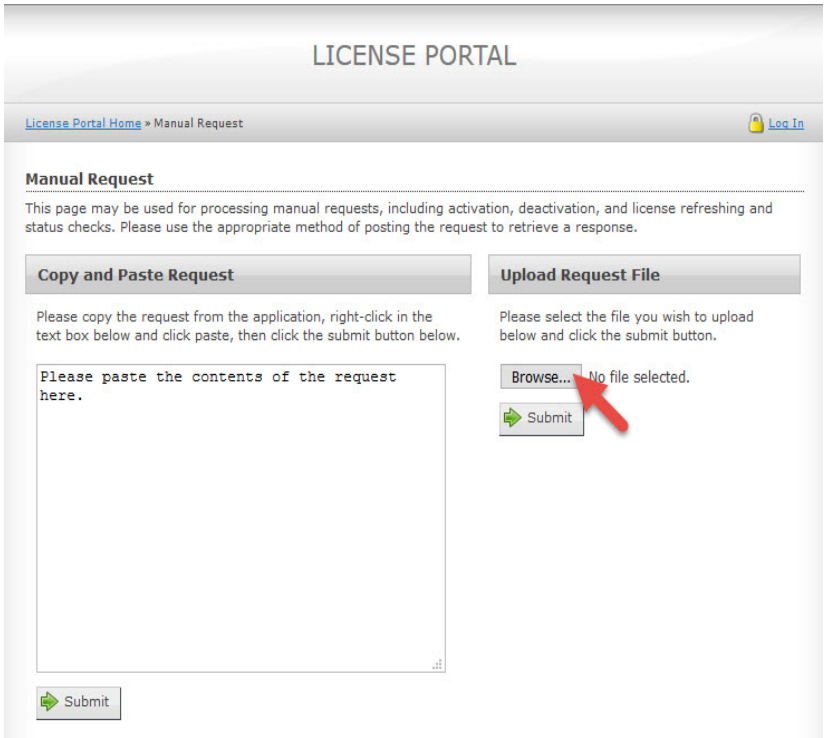
Figure 1-6: Manual or offline activation dialog



To activate Scaffold Elements without an internet connection:

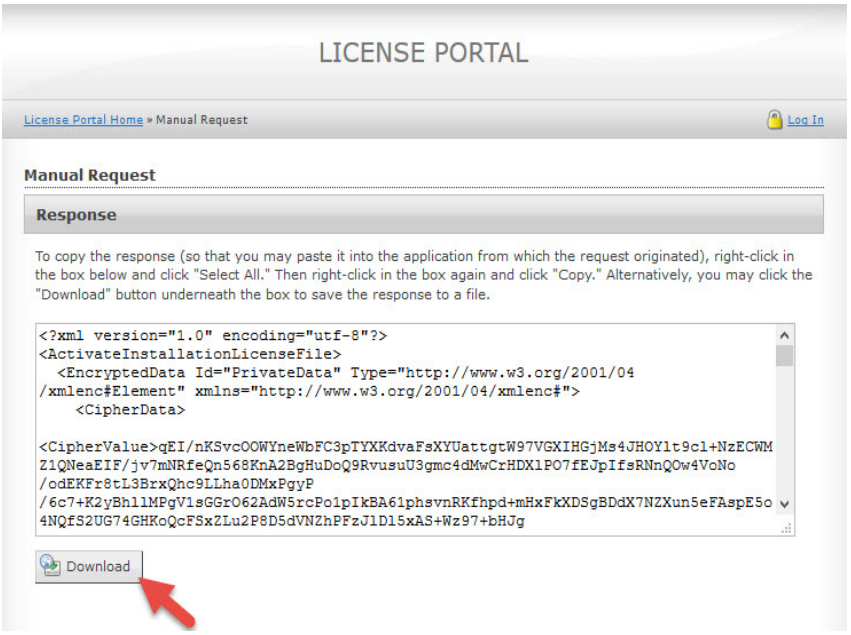
1. First, use the Save Activation Request button to create an activation request file.
2. Transfer this file to a computer with internet access (e.g. using a USB drive).
3. On the connected computer, navigate to*http://www.proteomesoftware.com/activate/*This link is also accessible from the Proteome Software Support page *(http://www.proteomesoftware.com/support/)* to make it easier to access from the internet-connected computer.
4. The License Portal will open. The Portal provides two different options for activating your software. Use the Browse button in the Upload Request File section on the right, and select the activation request file that was transferred from the offline computer(See Figure 1-7 below).

Figure 1-7: The Proteome Software License Portal



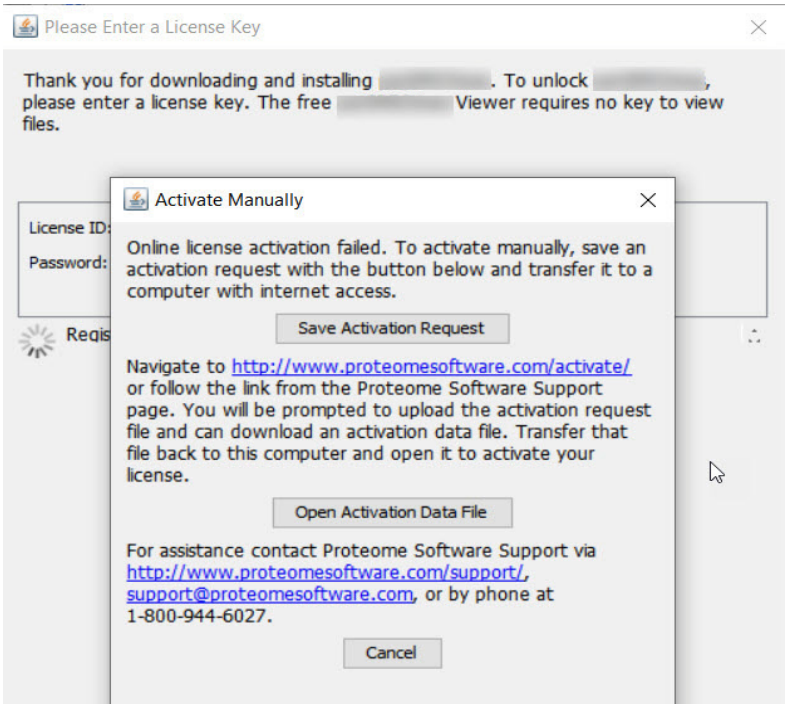
5. Click the Submit button just below the Browse button to upload the activation request file. The license portal will respond with a long text sequence(See Figure 1-8 belowFigure 1-8 below).

Figure 1-8: .License Portal Response to Activation Request



6. Click the Download button to save the response to a file named response.xml, which will be downloaded to the default download location.
7. Transfer the response.xml file to the computer on which Scaffold Elements has been installed.
8. Return to Scaffold Elements on the disconnected computer. Select Open Activation Data File.

Figure 1-9: Select Activation File returned by the License Portal



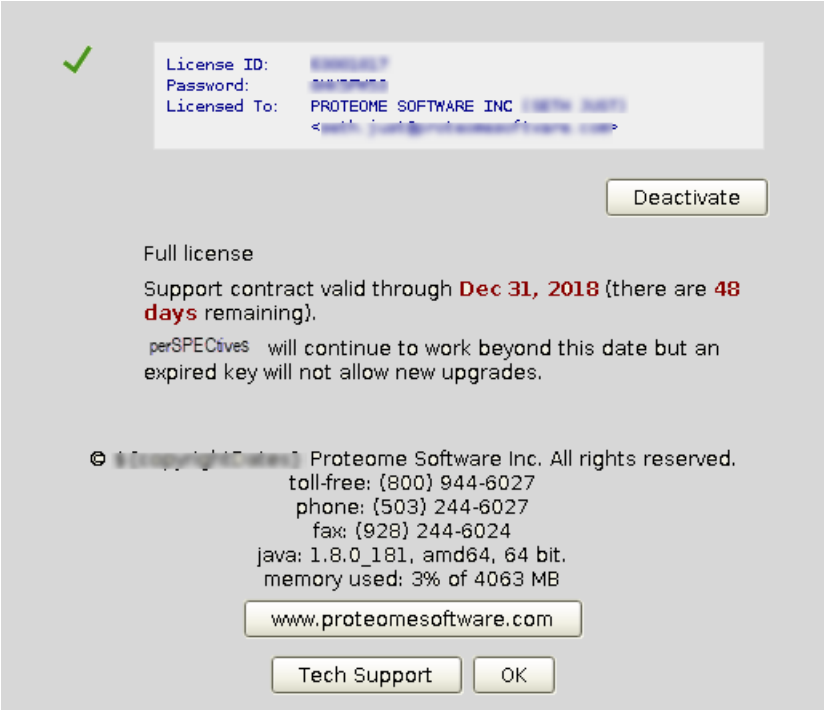
9. Browse to locate the response.xml file and click Open.
10. Scaffold Elements should report that the key was registered successfully. If not, please contact Proteome Software Support for assistance.

Time based license key renewal

Time based license keys have time limits connected to the term of the user's support contract. When the support contract expires, Scaffold Elements continues to work but upgrades are not allowed until the contract is renewed. The status of the Scaffold Elements license key may be checked by selecting **Help > About Scaffold Elements** from the main menu.

If the contract has expired and the user wishes to upgrade Scaffold Elements, clicking the **Renew** button in the dialog opens the **Key reset Request** page on the Proteome Software website. The user should complete the request. A sales representative will promptly contact him/her providing further information.

Figure 1-10: About Scaffold Elements dialog



Moving Scaffold Elements to a different computer

Each permanent Scaffold Elements key allows activation of the program on a single computer. If it becomes necessary to reinstall the program either on a different computer or on the same computer following an operating system upgrade or hardware replacement, the user may deactivate the key and then reactivate it on the new system. This may be done once per support contract period. If additional reinstallations are required within the same period, please contact Proteome Software Support.

To deactivate a key:

1. Be sure you have a record of your key and password. These were sent via email at the time of purchase, or may be copied from the Help>>About Scaffold Elements dialog.
2. Select Help>>Update License Key and click the Deactivate button.

To reinstall Scaffold Elements:

1. Download the program from the Proteome Software website to the new system and run the installation program.
2. Paste in the key and password and register as described in Installing Scaffold Elements.

Scaffold Elements Viewer

A free Scaffold Elements Viewer may be downloaded from*www.proteomesoftware.com.* The Viewer can open and display any *.METDB file created by Scaffold Elements, and allows users to distribute Scaffold Elements results to colleagues, collaborators or reviewers.

The Viewer may be installed on any number of computers, and multiple instances of the Viewer may be run on a single computer simultaneously. It performs most of the functions of the full Scaffold Elements program, but it cannot load search results files and analyze data.

Only a single fully-licensed instance of Scaffold Elements may be run on a computer at one time. Additional instances will function as Viewers.

Scaffold Elements Viewer

-

Scaffold Elements highlights

Scaffold Elements is a software tool designed to help researchers in the field of metabolomics identify analytes present in samples analyzed using liquid chromatography-mass spectrometry (LC-MS), LC-MS-MS or MS^E.

Supported analytic methods

Scaffold Elements supports untargeted search methods, which are global in scope and have the aim of simultaneously identifying and measuring as many analytes as possible from biological samples. It provides a graphical platform in which the user can load a large number of acquired MS, MS/MS or MS^E data files, which are then feature-picked, aligned and searched against spectral libraries for analyte recognition.

Graphical Views

Once the data has been searched and analyzed, the results are displayed in various graphical views. These views are designed to help the user inspect and validate the list of identified and unidentified analytes, and to perform visual inspection of the spectra.

Organize

The Organize View allows the researcher to group samples as required by the experimental design of the study. The Organize View is a very versatile graphical interface that allows the inclusion of meta-data. The user can create Categories and assign the samples to them by specifying the attributes of the samples with respect to those categories. For example, a Category might be Sex, with attributes Male and Female. Attributes may be added through a graphical interface or loaded from a file. This is the foundation for evaluating the metabolomics experiment from various viewpoints.

Summarize

Once attributes are applied, Scaffold Elements provides the ability to use them to and similarities across groups. Flexible summarization allows the user to select categories and use them to create a hierarchical categorization of the data. This makes it easy to:

- Compare similarities and differences in analytes at the sample level or at any level of summarization.
- Specify technical and biological replicates
- Compare the impact of tissue types, treatment types, demographic differences, measurement conditions and more.

Visualize

All identified analytes are listed in the Samples table, along with their measured intensities in each of the searched MS samples. Many specialized visualization tools are provided, along with the ability to:

- Cluster analytes.
- Use customizable colors to easily visualize quantification differences between samples or summarization levels.
- Apply thresholds and filters to focus on meaningful analytes.

Statistical Tests

Once the data has been properly organized, a number of statistical tests are available to help the user assess quantitative differences among the various attribute groups.

Scaffold Elements displays patterns of analyte levels across many samples with various attributes to provide new insights into an experiment.

Referencing Scaffold Elements Results

Please cite:

Incorporating In-Source Fragment Information Improves Metabolite Identification Accuracy in Untargeted LC−MS Data Sets. Seitzer, P.M., Searle, B.C. Journal of Proteome Research (2018) [*https://pubs.acs.org/doi/10.1021/acs.jproteome.8b00601*]

-

Chapter 2

Identifying Compounds with Scaffold Elements

Scaffold Elements performs analyte identification by searching spectral libraries. Multiple libraries may be searched in a single run, and libraries may consist of only MS data or may also include MS2 spectra. Elements can analyze raw data output files from most of the modern mass spectrometers (MS) available on the market, see Files supported by Scaffold Elements. Once imported, the MS data is analyzed to perform feature (or peak) detection. The features are matched to library spectra for identification. When multiple samples are present, they may be aligned in retention time among the loaded samples, and similar features are combined to form composite features.

The NIST library is included with the purchase of Scaffold Elements, and the METLIN library (as distributed by Wiley) is available at additional cost. Other libraries in standard formats, including HMDB, LIPID MAPS and MoNA may be downloaded and searched, and one of the most powerful features of Scaffold Elements is the ability to create personal spectral libraries, which provide better matching of MS2 spectra and the ability to use retention time information in identifying compounds (see Appendix "Creating A Personal Spectral Library," on page 170).

Scaffold Elements provides tools to validate, organize, and interpret the search results so that the user can easily manage large amounts of information and quantitatively compare samples. The loaded data can also be annotated with meta-data and the user can perform statistical tests using many possible summarization hierarchies.

This chapter covers the following topics:

- "How Scaffold Elements can help streamline metabolomics studies" on page 20, which provides a brief description of how to organize data in Elements to facilitate a variety of metabolomics experiments.
- "Scaffold Elements Views" on page 27, which briefly describes the different views that help analyze and validate the loaded experiment.

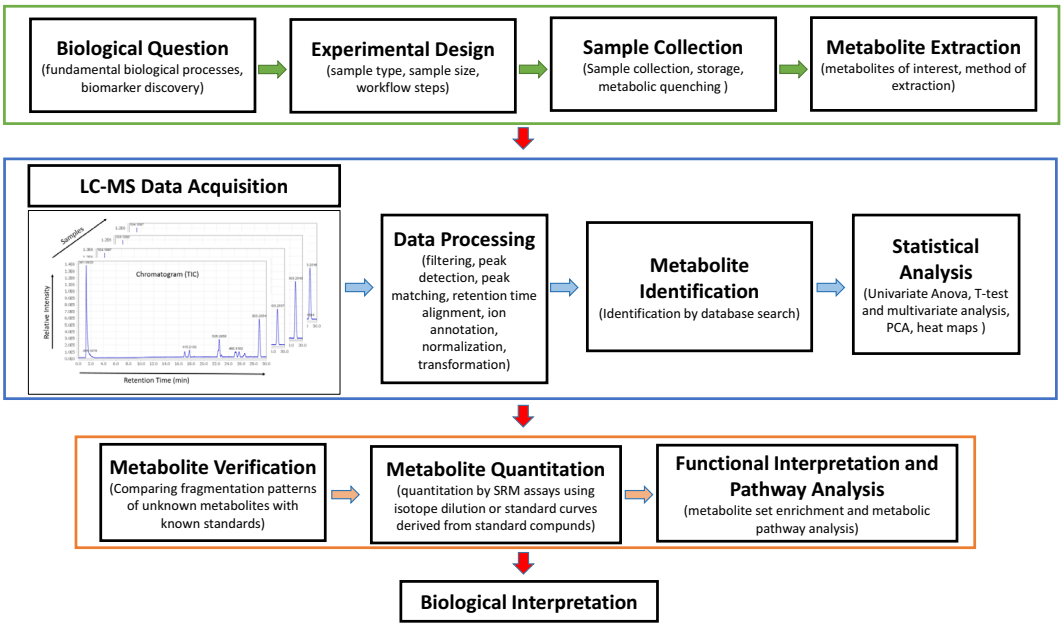
-

How Scaffold Elements can help streamline metabolomics studies

Scaffold Elements is an application designed to accommodate all aspects of a typical untargeted metabolomic or lipidomics study that uses LC-MS1 and (optionally) MS2 data to identify and quantitatively compare the analytes present in biological samples. The purpose of untargeted analysis is to generate new hypotheses for further tests by measuring all the analytes present in a certain biological system. For example, as reported by Zhou *et al.* [1], a typical metabolomic study aims at comparing multiple biological groups to identify analytes that are significantly altered. After establishing the main alterations present in a particular experiment, more specific quantitative studies can be performed through targeted approaches like SRM or MRM experiments, see Figure 2-1.

- "Scaffold Elements - supported experimental designs" on page 20
- "Scaffold Elements - data processing workflow" on page 21

Figure 2-1: A typical workflow for a metabolomic study, from Zhou et al.



Scaffold Elements - supported experimental designs

Generally, experimental design is tightly connected to the biological questions researchers ask, see Figure 2-1 above. The organization of a particular design can vary according to the type of explanatory variables or factors that are being manipulated in trying to formulate an answer to the question at hand. Typical factors can be time, which defines a time series experiment, or treatment which may define a case-control study, etc.

1. Zhou B., Xiao J.F., Tuli L. and Ressom H.W., Mol Biosyst. 2012 Feb;8(2):470-81. doi: 10.1039/c1mb05350g. Epub 2011 Nov 1

To be able to derive correct statistical inferences it is important to have a sufficient number of biological replicates; three are typically suggested but five are even better. Pilot studies are also recommended as well as are quality control samples. A proper experimental design for metabolomic investigations also includes analytical replicates, blanks, and negative and positive controls to infer the contributions of analytical and biological variation and to asses data quality.

Scaffold Elements has been designed to help the user easily replicate the structure of most experimental designs and to provide a reliable search engine for analyte identification. It also includes statistical tools to help answer biological questions.

Once the raw data is loaded in the program and searched for identification, Scaffold Elements opens a graphical interface displaying results of an LC-MS, LC-MS2 or MS^E metabolomic experiment. Results may be easily shared with other researchers through the freely available Scaffold Elements viewer.

A variety of possible experimental designs can be handled by the Organize View "The Organize View" on page 71, which provides tools to accommodate biological and technical replicates, define levels of blocking and assign custom meta-data attributes, allowing the researcher to analyze the data from different view points and thus to gain a deeper understanding of the factors responsible for patterns of variation seen in the results.

Scaffold Elements - data processing workflow

In a typical study,see Figure 2-1, the analytes, once extracted, are further separated and analyzed for identification using techniques such as gas chromatography (GC) or LC-MS1 and MS2. Scaffold Elements processes LC-MS, MS2 and MS^E raw data files acquired with a variety of mass spectrometers, see Loading data in Scaffold Elements.

Figure 2-1shows the Scaffold Elements' raw data processing workflow. This workflow is in line with a well established approach generally accepted by the scientific community.

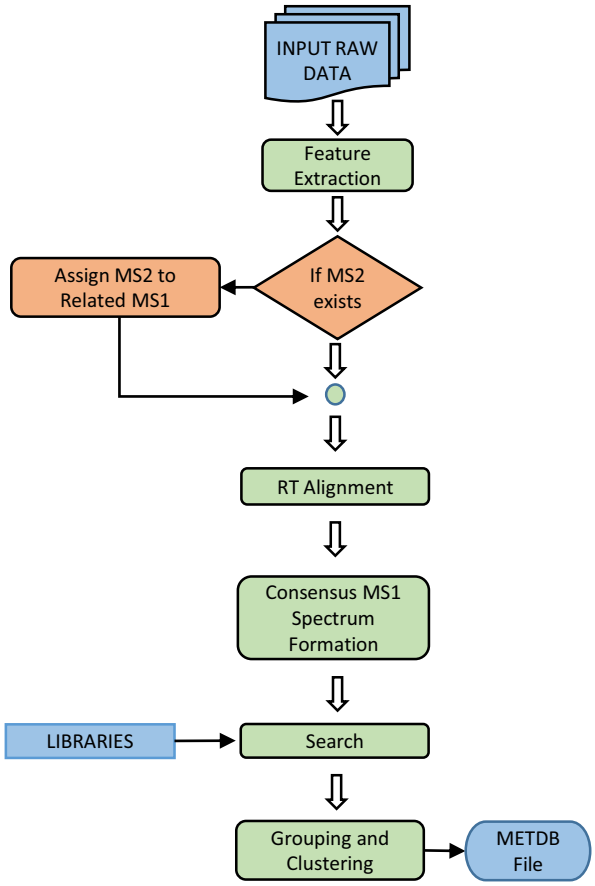
Before performing the identification search against the spectral library of known analytes selected by the user, the loaded raw data files are run through a series of preprocessing steps to properly extract the features recorded by the mass spectrometer. Coeluting features are grouped into MS1 peak groups. When multiple samples are loaded into the program, the identified features are aligned in retention time and similar features are grouped into composite or consensus features.

- "Input raw data" on page 22
- "Feature Extraction" on page 23
- "Assigning MS2 to related MS1" on page 24
- "Retention Time Alignment" on page 24
- "Consensus MS1 Peak Group Formation" on page 25
- "Identification by searching against spectral libraries" on page 25
- "Scaffold Elements search engine" on page 25
- "Grouping and Clustering" on page 26

-

- "Scaffold Elements Views" on page 27

Figure 2-2: Scaffold Elements file processing workflow

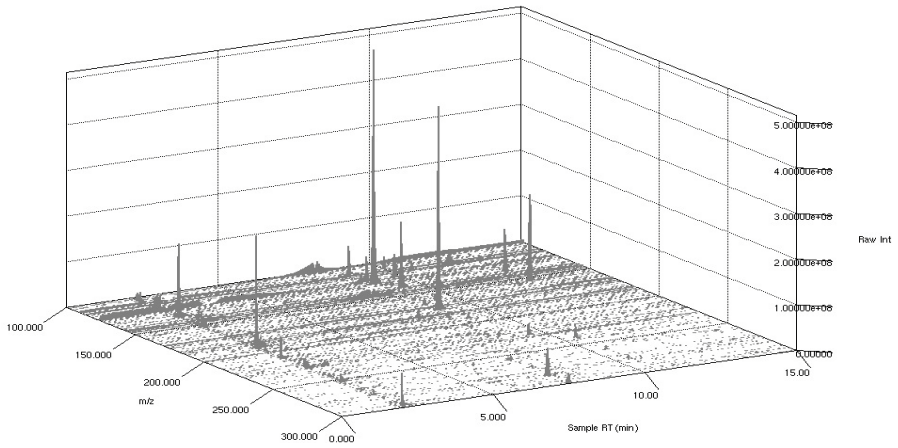


Input raw data

Figure 2-3shows an example of a raw dataset from a Liquid Chromatography Mass Spectrometry (LC-MS1) run in its 3D representation. In LC-MS1, the output of the LC column is inducted to a mass spectrometer periodically throughout the elution process. Each time point is typically referred to as an elution sampling point. At each elution sampling point the mass spectrometer produces an MS1scan which registers the m/z values and the corresponding abundance (intensity) of all the ionized molecules. One way of organizing the intensities is to show the collection of elution time profiles or Extracted Ion Chromatograms (XICs) at different m/z values.

In addition to resolving ions by their m/z values and obtaining estimates of their molecular masses, mass analyzers can further aid analyte identification by acquiring highly resolved and accurate MS2 spectral.

Scaffold Elements imports raw data files through the Workflow dialog. Raw files are initially converted into an internal format using the application msconvert, which is included in the open source cross-platform tool kit ProteoWizard[2].Once the data is imported the program goes through a number of processes, see Feature Extraction, that lead, for each raw data file, to the compilation of a list of features.



Feature Extraction

Feature extraction is the process of identifying local maxima in a three-dimensional (m/z, RT, intensity) landscape, where each maximum corresponds to the presence of a population of ionized chemical species. Generally, feature extraction algorithms aim to tease apart features generated by real ions from those generated by variations due to random electrical and chemical noise.

Feature Extraction in Elements

Scaffold Elements detects features in LC-MS using an algorithm developed in-house which is based on the fact that LC-MS data is inherently discrete (instruments have physical detection and resolution limits), and so individual features appear as clusters of raw (m/z, RT, I) data points. Scaffold Elements' feature finding algorithm seeks to first organize the raw (m/z, RT, I) data points into clusters of points and derive a single (m/z, RT, I) value for each cluster. These single (m/z, RT, I) values are used in all spectral library matching and analyte association steps. Details of the algorithm are exhaustively described in the Appendix "Agglomerative Point Clustering Feature Finding Algorithm," on page 223.

2. Chambers,M et al. Nature Biotechnology, 30, 918–920 (2012) doi:10.1038/nbt.2377

-

Assigning MS2 to related MS1

MS2 spectra, or fragment spectra, are extremely helpful in producing confident identifications. In the Elements scoring algorithm, the most heavily weighted aspect of analyte identification compares agreement between an experimentally derived MS2 spectrum and a previously-generated library MS2 spectrum.

It is possible to generate data that does not have MS2 spectra, and/or to match experimental data against a spectral library that does not have MS2 spectra. In that case, an MS2 score can not be produced, so analyte identifications necessarily rely on other metrics. However, we highly recommend using an experimental protocol that would facilitate the generation of MS2 spectra, and comparing results against a spectral library that contains MS2 spectra, such as the NIST spectral library or the METLIN Experimental Library (which are included in Elements).

There are two major classes of mass spectrometry data from which Scaffold Elements generates MS2 spectra: DDA (data-dependent acquisition) and DIA (data-independent acquisition).

DDA Data

For DDA data, the mass spectrometer automatically generates MS2 spectra, with an associated precursor m/z. From the list of features we discover in feature finding, we associate MS2 spectra to features if

- The RT of the MS2 scan is within the RT bounds of the feature RT range
- For profile mode data, the precursor m/z of the MS2 scan is not less than the minimum m/z - 1/2 the m/z feature width, and not more than the maximum m/z + 1/2 the m/z feature width.
- For centroid mode data, the precursor m/z of the MS2 scan is not less than the minimum m/z - 0.001, and not more than the maximum m/z + 0.001

MSE Data

For DIA data (note that currently, Elements only supports DIA data from Waters MS^E files), Scaffold Elements must generate MS2 spectra. It performs feature finding on the MS2 scans (or, in the case of Waters Ms^E, the high energy scans) to identify all of the individual fragment features. It then removes all non-monoisotopic features from both the MS1 and MS2 feature lists (or the low energy and high energy feature lists).

Next, it constructs “pseudo-DDA MS2 spectra” by associating features from the MS2 feature lists to a single precursor feature from the MS1 feature list. The algorithm used to perform the MS1-MS2 feature association step is a re-implementation of the “DIA Umpire” algorithm[3]

Retention Time Alignment

Coeluting features in each sample are grouped to form MS1 peak groups. Rough consensus

3. Tsou, Chih-Chiang, et al. “DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics.” Nature methods12.3 (2015): 258-264.

-

MS1 peak groups are formed by matching features from MS1 peak groups across samples and retention time alignment is performed across samples using these single-sample MS1 peak groups.

A subset of this initial set of consensus MS1 peak groups is designated as “anchors.” From the anchors, a monotonic mapping is formed between each sample and the reference sample. Linear interpolation is used to align the remaining features.

Retention time alignment is only performed between samples with the same polarity. Positive samples may be aligned with other positive or with mixed mode samples. Similarly, negative samples may be aligned with other negative or with mixed mode samples. If mixed mode samples are present, they are preferred for use as reference samples in the RT alignment. For more details.

Consensus MS1 Peak Group Formation

If all samples have undergone similar chromatographic procedures, an analyte would be expected to elute at essentially the same time in each sample in a study. Further, in most cases, true co-elution of different analytes is rare. As a result, it can be helpful to group ions that elute together into MS1 peak groups before attempting to identify them. MS1 peak groups in which many ion forms can be explained by a specific analyte identification are more likely to be correct.

Another advantage of grouping ions into MS1 peak groups is that unidentified ions in an MS1 group may be compared to MS2 spectra for other ion forms in the group, and in many cases may be identified as in-source fragments. This can be very helpful in improving the accuracy of analyte identification when no experimental MS2 spectra are available.

As a result, after RT alignment, coeluting ions in the various samples are grouped together to form consensus, or cross-sample, MS1 peak groups. Peaks from different samples are compared for similarity of MS2’s, m/z, isotopic distribution and peak shape, and if they are judged to be similar enough they are grouped.

Identification by searching against spectral libraries

Analytes are identified in Scaffold Elements by comparing the mass value identified for each feature in the consensus MS1 peak groups to the calculated exact mass of library analytes. The calculated exact mass is determined entirely from the chemical formula of the analyte. This mass is adjusted by the list of expected ion forms for the analyte (in this program, these ion forms, or adducts, are specified by the user when they load data). These analyte identifications are ranked according to a scoring algorithm (described in the next section).

Scaffold Elements search engine

An **ID Score** is computed for each MS1 peak group by comparing features to Analyte Records in the search library, using several kinds of additional evidence. This evidence is represented as individual match scores, as well as an assessment of the explanatory value of the MS1 peak group as a whole.

An Analyte Record is an entry in a spectral library corresponding to a specific experiment. This typically means a purified standard is run on a specific instrument, and fragmented at a specific collision energy. Variations will exist in the fragment spectra generated on different

-

instruments run with different collision energies.

When a feature matches multiple different Analyte Records, the highest ID score indicates the identification most likely to be real.

Elements computes four match scores for every Feature – Analyte Record match, three of which are used to determine the ID score:

- MS2 Sub-Score"MS2 Score" on page 216
- XIC Sub-Score, not used in ID score"XIC Score" on page 217
- Mass Accuracy Sub-Score"Mass Accuracy Score" on page 214
- Isotopic Distribution Sub-Score"Isotopic Distribution Score" on page 214

Then it integrates them through a linear combination of their values into an overall ID Score, and adds the MS1 Annotation Score. see"Analyte ID Score" on page 213.

Grouping and Clustering

All ions assigned to the same consensus MS1 peak group are organized together into a single analyte cluster. Analyte grouping, however, depends on which option is selected for the "Treat each MS1 peak group as a single analyte" parameter in the Search tab of the Workflow dialog.

If "Treat each MS1 peak group as a single analyte" is checked, the ions are also organized into a single analyte group. This is the default behavior, because unless there is a reason to expect coeluting ions, it is generally the case that all of the ions that elute at the same retention time are different forms derived from the same analyte.

If "Treat each MS1 peak group as a single analyte" is not checked, then only analyte identifications made with identical ions are organized into the same analyte group. Unidentified coeluting ions are treated as separate analytes (or discarded if "Retain unknown analytes" is not selected).

Scaffold Elements Views

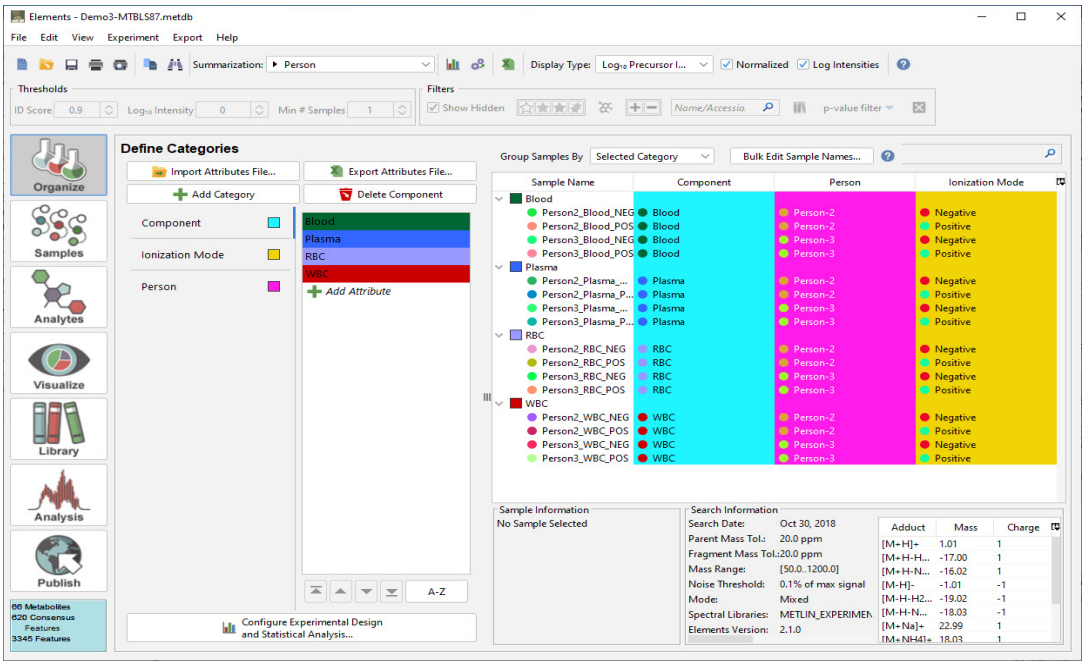
Scaffold Elements offers both a high-level overview of the list of analytes and a detailed look at supporting data. Scaffold Elements presents the more detailed levels in a coherent structure, helping the user in verifying critical findings. The information is organized into different views that can be easily accessed through the main Scaffold Elements window.

Organize View

This view shows the list of MS samples loaded in Scaffold Elements. Tools in the view help the user assign meta-data information to the loaded files and organize them according to an experimental design. At he bottom of the view, file and search information is provided along with the list of adduct/loss ions used in the search.

For more information see “The Organize View” on page 71.

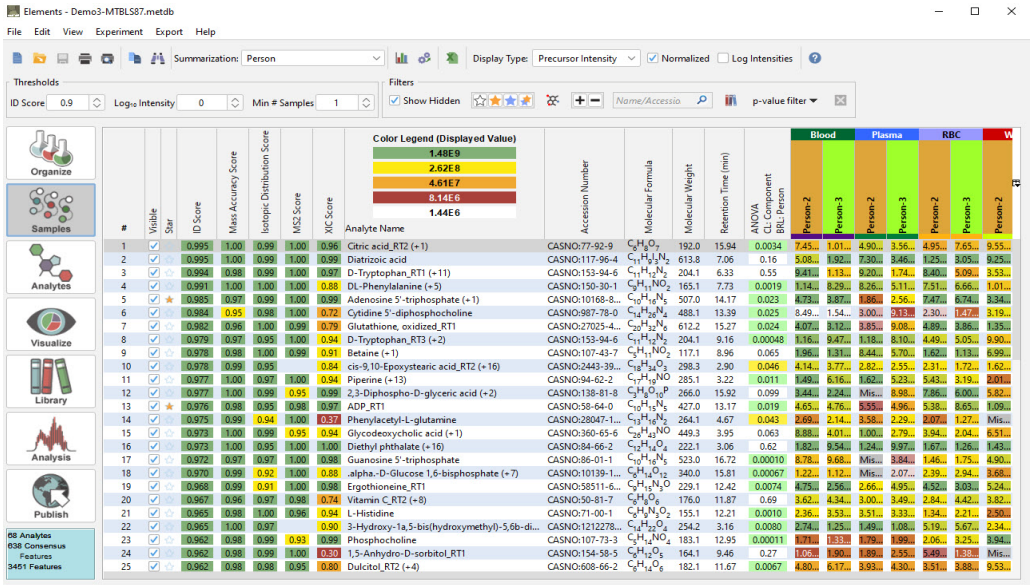
Figure 2-3: Scaffold Elements: Organize View



Samples View

The Scaffold Elements Samples View provides overviews of the loaded data that help the user make direct comparisons among categories of samples summarized in different ways. For more information see “The Samples View” on page 91

-

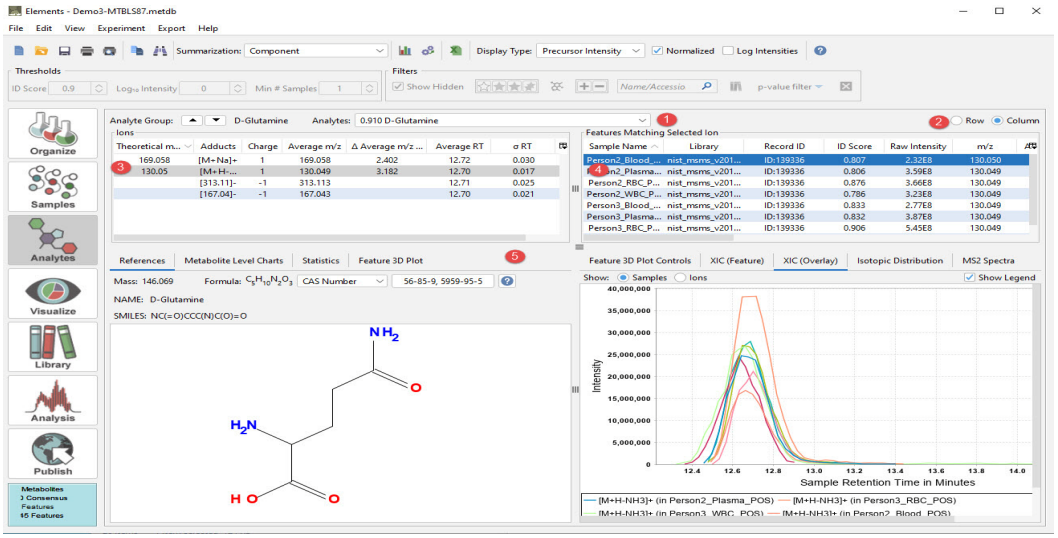


Analytes View

The Scaffold Elements Analytes view structures a great deal of detailed information about an analyte, showing the ions that identify the compound within each sample. It also shows the MS2 spectra, when present, that confirm the identification. For more information see "The Analytes View" on page 107.

Scaffold Elements: Analytes View

Figure 2-4: Scaffold Elements: Analytes View

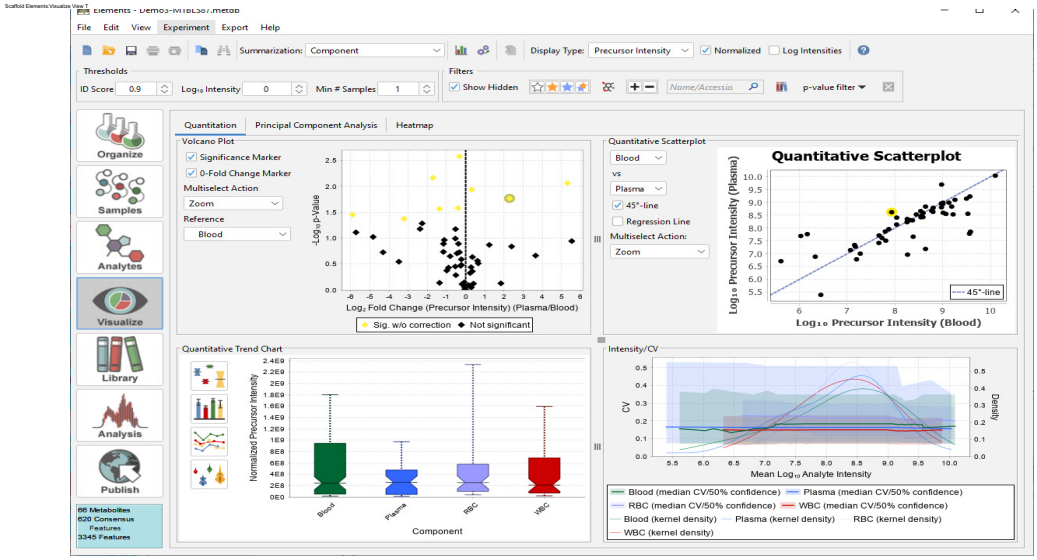


Visualize View

The Visualize View has three tabs: Principle Component Analysis, Quantitation, and a Heat Map of the filtered analytes list shown in the Samples table. For more information see "The

Visualize View” on page 121.

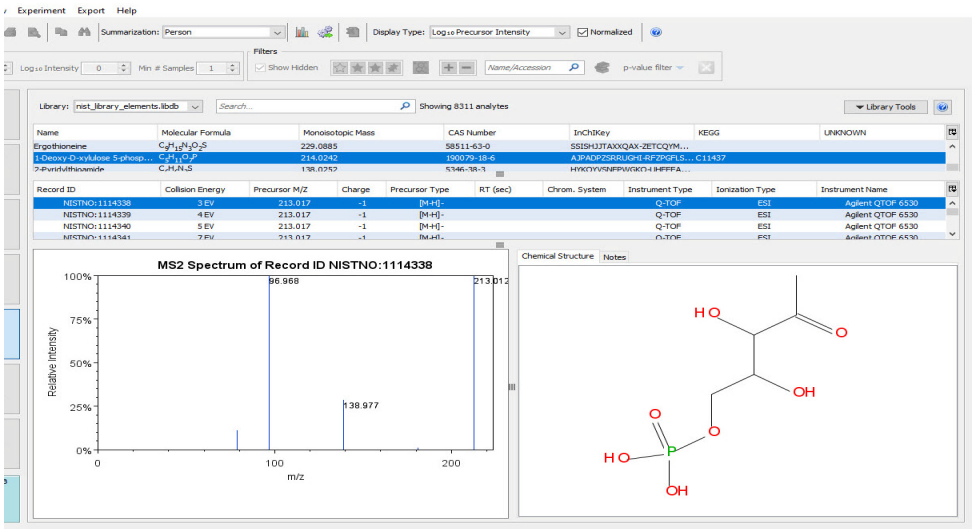
Figure 2-5: The Visualize View



Library View

The Library View provides tools to view detailed information about any member of a selected spectral library loaded in the Scaffold Elements. For more information see “The Library View” on page 133.

Figure 2-6: Scaffold Elements: Library View



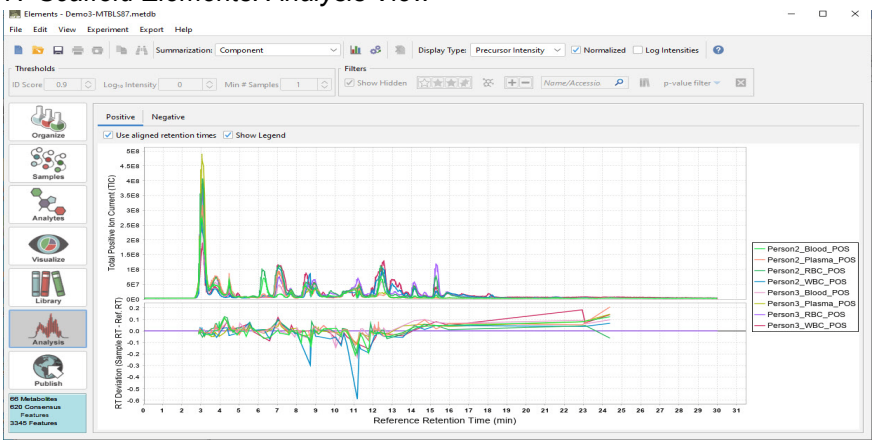
Analysis View

The Analysis view displays the Total Ion Current (TIC) as a function of retention time for each of the raw data files loaded in Scaffold Elements. It shows the chromatograms aligned

-

in retention time and unaligned. It also provides the retention time deviation from the reference sample. For more information see “Analysis View” on page 139.

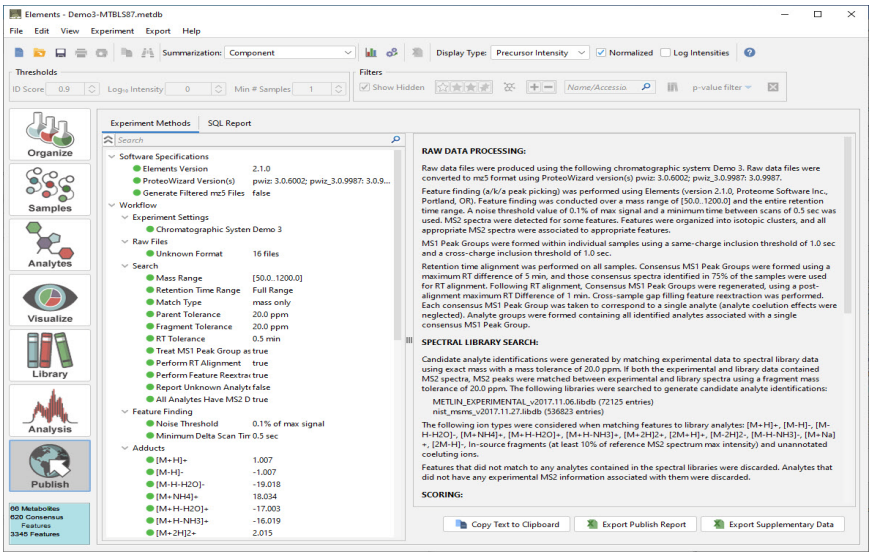
Figure 2-7: Scaffold Elements: Analysis View



Publish View

The Scaffold Elements Publish view contains two different tabs: the Experiment Method tab which records the experimental information and the SQL Report tab through which the user can query the current METDB file using SQL. Queries can be saved and reused.

For more information see “The Publish View” on page 141.



Chapter 3
Loading data in Elements

Scaffold Elements imports and analyzes data from various types of Mass Spectrometers. Data loading is accomplished through a dialog that guides the user in setting all of the parameters needed for data analysis and for identification using one or more spectral libraries.

Files supported by Scaffold Elements

Scaffold Elements reads and processes both regular and high mass resolution MS data with or without MS2 scans and Waters MS^E data.

Scaffold Elements supports a number of vendors' data file formats and open file formats. The Figure shows the list of supported formats.

Vendor's File Formats			
Vendor	Software Application	MS Instrument	File Format
SCIEX	Data Explorer (4.2 and higher)	4X00 and higher TOF series	*t2d
	Data Explorer (prior to 4.2)	AB SCIEX Voyager (MALDI-TOF)	*.dat
	Analyst	Qstar, Qtrap	*.wiff
Agilent	Mass Hunter	Q-TOF	*.d directory
Bruker	XMASS/XTOF	Flex Series	*.d directory
	flexAnalysis	APEX, microQTOF, microQTOF-Q	
	flexAnalysis	Esquire Series	
		FTICR	
Thermo	XCalibur	LCQ, LTQ, Orbitrap	*.raw
Waters	MassLynx	All Waters Mass Spectrometers DDA, MSE supported HDMSE summed over drift times	*.raw directory
Open Formats			
HUPO Proteomics Standards Initiative mzML			*.mzML

Vendor's proprietary format raw data files can be processed only on Windows operating systems.

-

Loading data in Scaffold Elements

To create a new experiment the user can either select ***File > New*** or click the **New** icon located in the tool bar below the main menu in the Scaffold Elements window, or click on the **New** button appearing in the "Welcome to Scaffold Elements" dialog when the program is launched. The Workflow dialog appears and the user can conveniently set up the parameters for the analyte search and select the raw data files to be analyzed. Once the parameters have been defined and the raw data files selected, the start button will become available to begin the analysis.

Another option for starting a new analysis is provided by the command ***File > Reanalyze***. It allows the user to start a new analysis, but with all of the parameters and files from an existing analysis. If the mz5 and index files are still available, Scaffold Elements will not need to repeat file conversion or feature extraction. Some typical situations in which this might be useful include:

- **Adding data files to the experiment** - It may be desirable to load a small number of files in order to check the parameter settings, and then to add the remaining files. In this case, the user can select ***File > Reanalyze***, add the file(s), and continue.
- **Adjusting parameters in an existing experiment** - The user might open a METDB file and decide to adjust one or more parameters. Selecting ***File > Reanalyze*** opens the **Workflow** dialog opens showing all the parameters used and the list of raw data included in the experiment. At this point the user can easily adjust the parameters and rerun the search, to produce new results.

Workflow dialog

The Workflow dialog contains two panes: the "Search Setup" pane, on the left side of the dialog and the "Load Data" pane, on its right side. It also includes a number of functional buttons below or within each pane.

- **Workflow Buttons** - The buttons located below the "Search Setup" pane allow the user to either save the current selection of options and parameters listed in the pane to a named workflow or to retrieve them from an existing one.
  - *Load Workflow* - This selection opens a file browser to locate the WORKFLOW file to be loaded.
  - *Save Workflow* - After all parameters and options have been properly defined and when each tab shows a green check, the user has the option to save the information to a WORKFLOW file. Clicking the button opens a file browser so the user can assign a name and save the WORKFLOW file to a convenient directory.

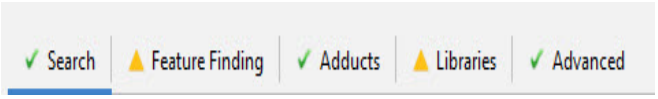
  A message located at the top left corner of the dialog reports the name of the workflow currently loaded.

-

- **Search Setup pane** - Through this pane the user defines the search parameters, chooses the adducts and selects the spectral libraries to be used for the search. The information is organized into the following tabs:
  - Search tab
  - Feature Finding tab
  - Adducts tab
  - Libraries tab
  - Advanced Tab

Note: A small warning icon may appear near each tab name. It alerts the user when not all of the required options have been specified. When information is missing, a yellow triangle is shown. When all of the options have been defined a green check replaces the triangle.

Figure 3-1: Tab warning icons



Once the user has defined the search parameters, selected the correct adducts and imported the spectral libraries to be used for the search, the next step is to select the files to be analyzed. This is done by populating the list in the Load Data pane.

- **Load Data pane** - This pane contains a list of raw data files to be imported, analyzed and searched for identifications in Scaffold Elements. Three buttons are used to populate and adjust the list. To select files to be included, the user clicks the "Add" button and navigates to the location containing the raw data files. The "Remove" button can be used to remove undesired files from the list.Some data files require additional information for processing and will display a yellow triangle warning indicating that additional information is needed. A tooltip explains what is missing. For example, when loading Waters MS^E data, the user may be prompted to enter or confirm a Lock Mass Correction.

  The "Edit" button brings up a dialog to allow the user to supply the missing parameters for a selected file. Multiple files may be selected for editing together, provided that they share the same values for the data to be confirmed or modified. If incompatible files are selected together, Scaffold Elements will not allow simultaneous editing and the user will need to change the set of selected files.

See Files supported by Scaffold Elements for a comprehensive list of supported file formats.

- **Chromatographic System** - The Chromatographic System is an identifier to allow the user to track all of the information needed to describe the chromatography associated with the files loaded. This may include the mobile phase, stationary phase, filter size, column length, and pore size, just to name a few important factors. It is the responsibility of the user to keep track of all of the pertinent details associated with the chromatography setup and assign a single string representation of the chromatographic system. Scaffold Elements will accept any character string.

-

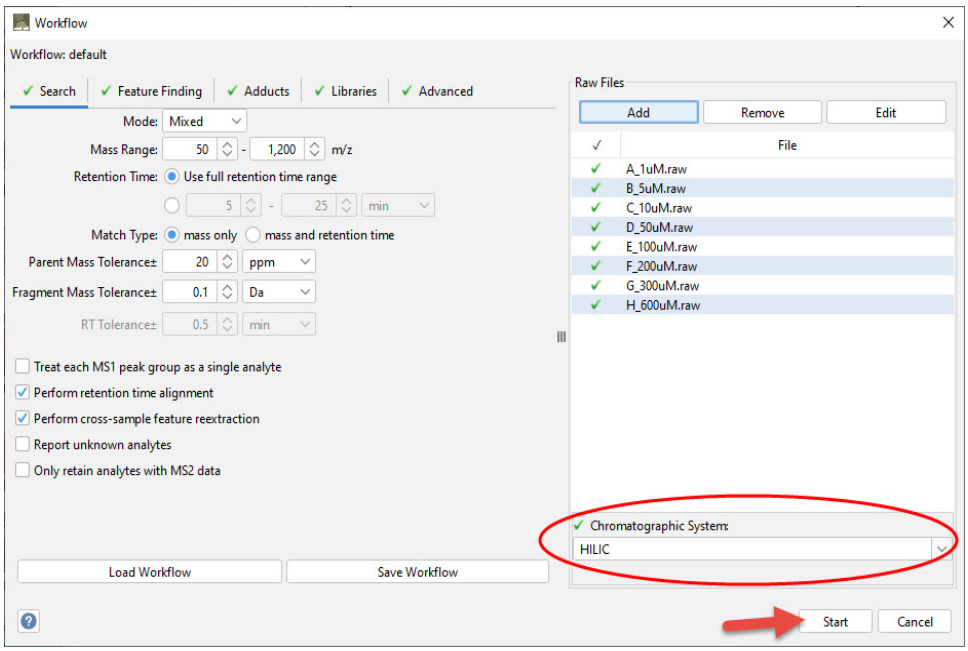
Specifying the chromatographic system is required when the “Match Type” selected in the search options is “mass and retention time” as it is important that only retention times derived from similar chromatographic systems be used in matching. This type of search is usually used with a personal spectral library (see) when the spectra comprising the library were acquired under the same conditions and the retention times are fairly reliable for use in identification.

When matching only on “mass”, specification of the chromatographic system is optional, but it is important if the experimental results are to be saved in a personal spectral library.

The dropdown list is populated with all chromatographic systems found in entries in the currently selected library or libraries.

Once all of the panes are properly filled, the “Start” button located on the lower right corner of the dialog becomes available for use. When it has been clicked, loading and analysis of the data begins.

Figure 3-2: Starting the search.



Depending on the amount of raw data submitted for analysis, the loading and analysis phase might take a considerable time. Once processing is complete the Samples View opens with its table populated with the list of identified analyte groups meeting the default score threshold. If **Report Unknown Analytes** was selected during the search, the user will have the option to show unidentified features.

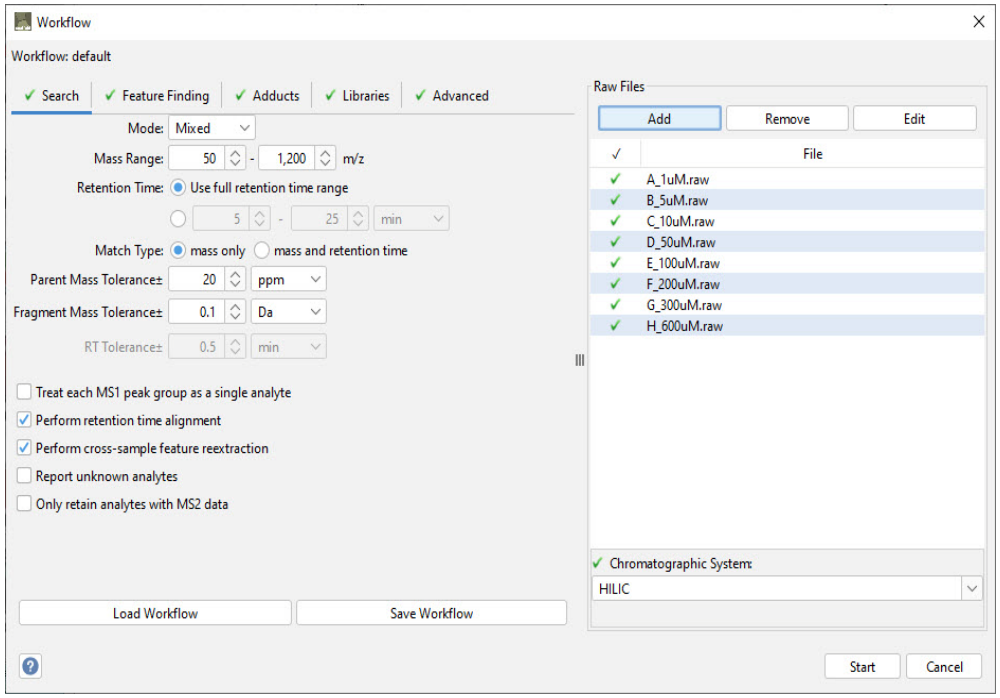
-

Search tab

This tab allows the user to adjust the parameters used in the search:

- **Mode** - Specifies the ionization mode used for data collection in the mass spectrometer. Elements allows analysis of positive, negative and mixed-mode data in a single experiment. The mode setting determines the adduct list from which the user may choose. The default mode is mixed. If only one mode is present in a raw file when mixed mode is selected, Scaffold Elements picks the correct mode automatically.
- **Mass Range** - Specifies the m/z range included in the analysis. Default values are proposed by the program. Limiting the mass range speeds the analysis.
- **Retention Time** - A radio button allows the user to choose to search over the full retention time range contained in the input data or to specify a specific retention time range in either minutes or seconds. Selecting a specific range will speed processing and may eliminate extraneous identifications of compounds that exit the column very early or very late that are unlikely to be of interest.

Figure 3-3: Workflow Dialog: Search tab



- **Match Type**- Determines whether the scoring of potential matches with the spectral library will be based on mass alone or will also consider retention time. If “mass and retention time” is selected, the Chromatographic System must be selected.

-

- **Parent Mass Tolerance** - Size of the mass tolerance window of the parent ion which typically depends on the mass accuracy and mass resolution of the instrument that collected the data (units can be ppm or Dalton).
- **Fragment Mass Tolerance** - Size of the mass tolerance window for the MS/MS fragment ions (units can be ppm or Dalton). We recommend using the proposed default value or a larger value than the mass accuracy of the instrument used to collect the data.

- **Search Option Checkboxes**:
  - Treat each MS1 peak group as a single analyte - When this option is checked, every ion that has met the criteria for inclusion in an MS1 peak group is retained in the analyte cluster that results from it and labeled an "Unnatotated Ion".

    If "Treat each MS1 peak group as a single analyte" is unchecked, any ions within an MS1 speak group which are not identified are presumed to be other coeluting analytes and are removed from the cluster and treated as "Unknown Analytes."
  - Perform retention time alignment - If this option is selected, retention times alignment across samples is performed and retention times are adjusted accordingly.
  - Perform cross-sample feature reextraction - When this option is selected, if a feature is found in one sample but not in another sample, the program will reexamine the corresponding portion of the data landscape in that other sample in an attempt to find a lower-intensity or lower quality peak that may indicate presence of the feature.
  - Report unknown analytes - When checked, non identified features are recorded in the METDB file and shown in the Samples view when Thresholds are minimized to zero. Its default status is checked. This option is useful when in need to reduce the size of the METDB file.
  - Only retain analytes with MS2 data - When this option is selected, only composite features including at least one MS2 spectrum will be included in the search. MS2 information results in more accurate identification. This option is also useful when data is being searched for the purpose of building personal spectral libraries.

Feature Finding tab

- **Noise Threshold** - Specifies when a signal is considered noise and provides the user with the flexibility of trading sensitivity for reduced processing time and memory usage. The threshold is applied during feature extraction and reextraction across samples, but the lower intensity data is searched during reextraction for isotopic peaks. There are two options available for defining the noise level: (1) Percentage of the maximum signal and (2) Input a specific minimum intensity value.

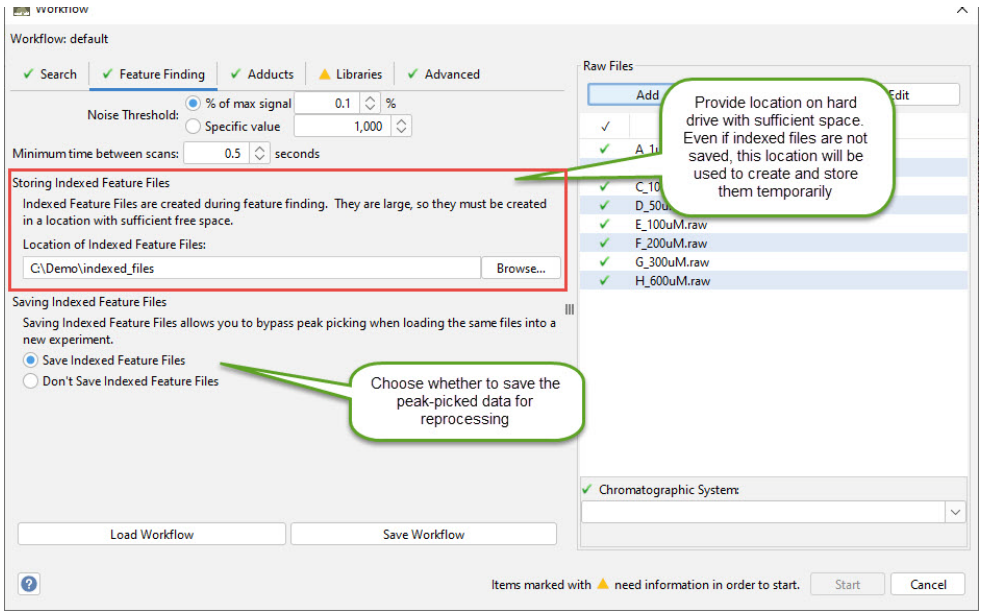
  A preference setting, reached via Edit>Preferences>ProteoWizard, allows the user to specify that the noise threshold should be applied during conversion of the raw data to MZ5 format by the ProteoWizard program MSConvert. If this option is selected, the resulting MZ5 files contain no peaks below the specified noise thresholds. This results in

smaller MZ5 files and reduces memory requirements for processing very large files. It is useful, for instance, in processing FT-ICR data. Because the lower intensity data has been completely discarded, however, Elements is unable to access it during isotopic distribution reextraction.

- **Minimum Time Between Scans** - this option is provided for use when data is collected with an extremely fast sampling rate, resulting in very large files with many scans. In such cases, Elements will respect the minimum time between scans setting and skip scans, which supply no additional information for peak-picking. In most cases, this value should be left at the default setting.
- **Storing Indexed feature Files** - When Scaffold Elements reads raw data files, it creates MZ5 files using MSConvert, an application included in the ProteoWizard tool kit[1]. During the feature finding phase, Scaffold Elements also creates index files of the MZ5 files, or indexed feature files, that need to be stored in a location with sufficient free space. The default location is the local temp folder, but because of the size of these files we advise the user to provide a more suitable storage location by typing a folder address in the text box or by browsing to the desired storage location using the button “Browse...”.

Note: The new location will revert to the system temp directory when the user closes and reopens Scaffold Elements, but it is retained in the Workflow if the “Save Workflow” option is selected.

Figure 3-4: Storing Indexed feature Files option.



1. Chambers, M. et *al*. Nature Biotechnology, 30, 918–920 (2012) doi:10.1038/nbt.2377

-

We highly recommend selecting a location other than the system temp folder with enough memory so that Scaffold Elements can store the index files. This expedites processing and it prevents memory-related errors when analyzing large datasets.

- **Saving Indexed feature Files** - This option allows the user to choose whether to keep the MZ5 and Indexed feature files. Saving the files allows the user to reanalyze the data without repeating the time-consuming steps of file transformation and feature extraction, and also allows full three-dimensional viewing of the data. Without the feature index files, only the selected features may be visualized in the Feature 3D Plot. If the user chooses not to save the files, they are deleted when loading is complete

Adducts tab

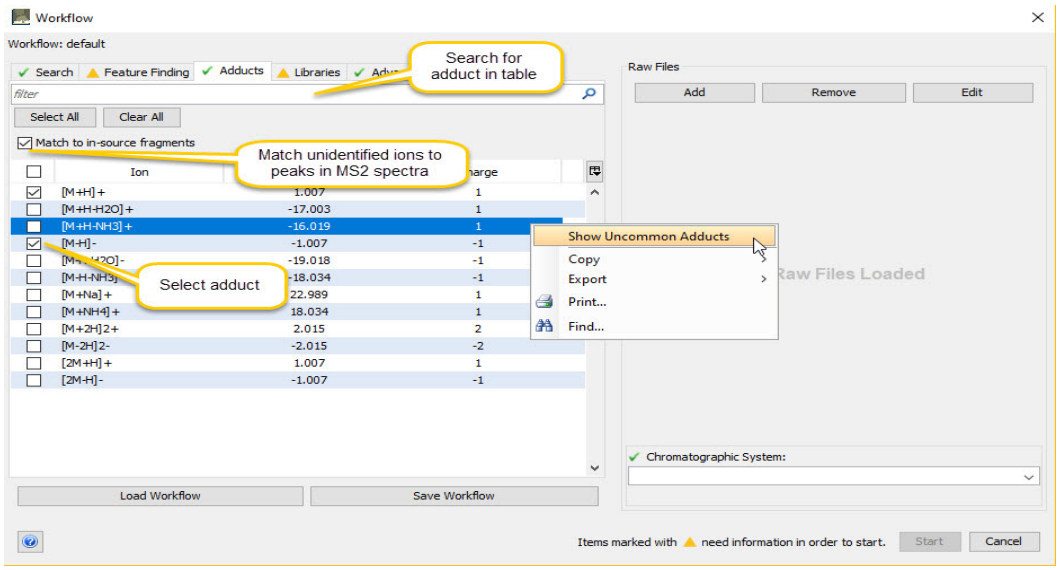
Through this tab the user selects the adducts Scaffold Elements will use for the search. The tab includes a table where each row lists the type of adduct ion, indicated as, for example, [M+H]+, a check box for selection and some of the adduct properties like the Mass Shift and charge. The adducts appearing in the list vary depending on the ion mode selected in the Search tab. The user can either select all the adducts appearing in the list by checking the top check box or choose specific adducts by check marking the ones to be included in the search.

By default the list shows only the most common adducts. The option **Show Uncommon Adducts**, available from the context menu which appears with a right click of the mouse, provides a more extensive list of adducts.

A search filter is also available above the table to make it easier to find adducts of interest.

Note that choosing too many adducts may increase the number of false identifications, so we recommend choosing only the ion forms that are expected to be common in the current experiment.

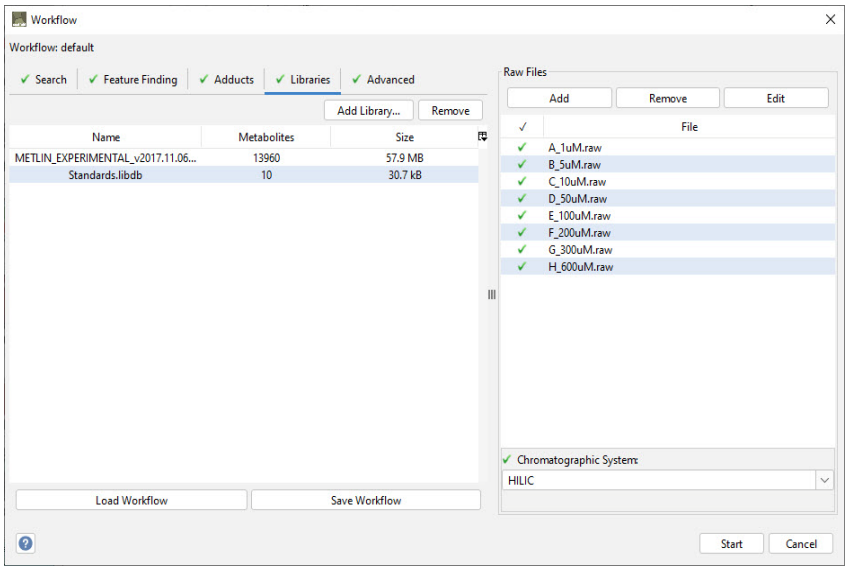
Figure 3-5: Workflow dialog: Adducts tab



Libraries tab

Through this tab the user loads and selects spectral libraries for the search. The Library tab includes a table and two action buttons used to populate or pare the content of the table. Once populated, the table shows the list of spectral libraries Scaffold Elements will use in the analyte identification search. The user can inspect the content of a spectral library through the Library View, see "The Library View" on page 133

Figure 3-6: Workflow dialog: Library tab



When first selected, the Library tab pane is empty and as a warning a yellow triangle appears in the title tab. Clicking the "Add Library" button opens the from which the user can select

-

one or more libraries to use for the metabolomic identification search, load a new library or download additional libraries from the Proteome Software website using the tools provided by the dialog.

After one or more spectral libraries have been selected from the “Library Manager”, a table appears in the Library tab pane which lists in each of its rows the name of the library, the number of entries included in it, and its file size and the library’s vendor.

The “Remove” button can be used to pare the list of spectral libraries available in the table.

Searching Multiple Libraries

Analyte vs. Analyte Record

When thinking about spectral libraries, it is useful to distinguish between an **Analyte** and an **Analyte Record**. **Analyte** refers to an annotation of a theoretical chemical entity, and will typically have a name, molecular formula, identifiers, and metadata, such as a KEGG Pathway. **Analyte Record** refers to an example of a specific analyte. In the vast majority of cases, an analyte record is associated with an experimentally observed MS2 spectrum of the analyte, along with the associated precursor type (or ion form), collected on a particular instrument with a specific collision energy.

When Elements matches experimental data to spectral libraries, it actually matches the data to analyte records, not to analytes themselves.

When multiple libraries are searched, the same analyte may appear in more than one library. In such cases, prior to comparing experimental data to the set of libraries, Elements combines analyte records for the same analyte.

Merging Analyte Records

In order to be merged, the exact mass of the analytes must be within 1×10^{-6} Da, the molecular formulae must be equal, and they must share at least one analyte identifier.

When two analytes are merged, all analyte metadata, records, and identifiers are combined into the merged analyte. For example, if Analyte X from library A has records X1, X2, and X3. Analyte X from library B has records X4, X5, and X6. In the search, Elements would combine these to form one instance of Analyte X, with records X1, X2, X3, X4, X5, and X6.

-

Library Licensing Information

*Depending on the type of licensing different spectral libraries will be available for download. A purchased copy of Elements gives full access to a copy of the NIST Mass Spectral Library (*NIST/EPA/NIH Mass Spectral Library (EI)*) and allows download of the METLIN Mass Spectral Library, developed by the Scripps Metabolomics Center and distributed by Wiley* (WILEY/METLIN Mass Spectral DatabaseWILEY/METLIN Mass Spectral Database*) for use in Elements searches. For evaluation purposes, temporary and limited access to the NIST and METLIN libraries is provided for the time defined by the evaluation key. Various other libraries may be downloaded from the Proteome Software website.*

- The Elements for Metabolomics End User may not redistribute the NIST/EPA/NIH Mass Spectral Library in any manner without explicit written approval by NIST. Contact Proteome Software atsupport@proteomesoftware.com for more information.
- Copyright protection on the compilation of data in this Library is secured by the US Department of Commerce in the United States and in other countries that are parties to the Universal Copyright Convention, pursuant to Section 290(e) of Title 15 of the United States Code.
- It is expressly understood and agreed that unauthorized copying of this Library is not permitted.

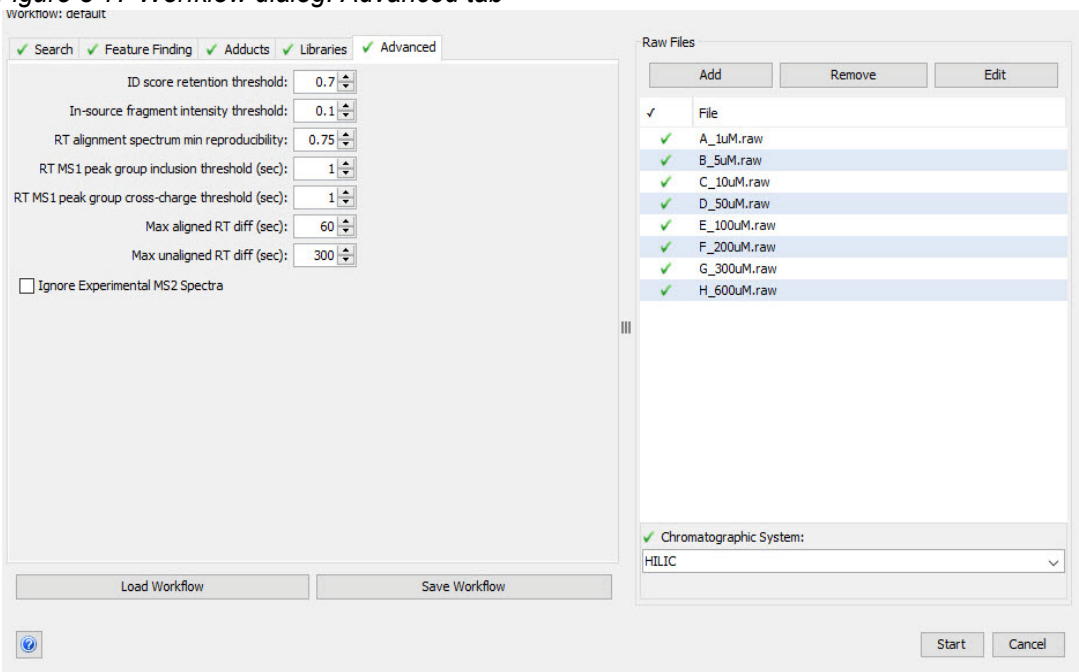
- The Elements for Metabolomics End User must comply with the Wiley METLIN End User Agreement. see http://www.proteomesoftware.com/company/metlin-eula/ for more information.
- It is expressly understood and agreed that unauthorized copying of these Libraries is not permitted.

Advanced Tab

The advanced tab contains internal thresholds and parameters which are used in the analysis, but which, in most cases, will not need to be changed.

-

Figure 3-7: Worfklow dialog: Advanced tab



- **ID score retention threshold:** the minimum ID score which must be attained by an analyte-library match in order to be included in the MS1 peak group.
- **In-source fragment intensity threshold:** the minimum intensity of the peak in a library MS2 to which a feature can be matched in order to label it as an in-source fragment of the analyte identified by the MS2. Expressed as a proportion of the maximum intensity peak in the MS2.
- **RT alignment spectrum min reproducibility:** the proportion of samples which must contain an identification in order for it to be used as an anchor point during RT alignment.
- **RT MS1 peak group inclusion threshold (sec):** the maximum allowable difference in retention time between features for inclusion in the same MS1 peak group.
- **RT MS1 peak group cross-charge threshold (sec):** the maximum allowable retention time difference when aligning samples of different polarities.
- **Max aligned RT diff (sec):** the maximum difference in retention time allowed when forming consensus MS1 spectra across retention time aligned samples.
- **Max unaligned RT diff (sec):** the maximum difference in retention time allowed when aligning features to perform retention time alignment. This should be a higher tolerance than the Max aligned RT diff(sec)/
- **Ignore Experimental MS2 Spectra:** causes the program to consider only MS1 data in its analysis, even if the input files contain MS2 spectra. This may be useful if the MS2 spectra were incorrectly collected or collected under very different conditions than the library MS2 spectra.

-

File type created by Scaffold Elements

Scaffold Elements creates its own file type called METDB. It is a lightweight, high performance SQL database file. Scaffold Elements provides the user the opportunity to query the experiment file using Structured Query Language (SQL) and to save these queries for future use. This direct access to the data structure gives the user of Scaffold Elements a unique capability to manipulate and analyze their data.

METDB files may be opened with the free **Scaffold Elements Viewer**, allowing users to easily share their results with clients or collaborators. The Viewer retains all functionality of the fully licensed program except the ability to load and process new data and the Library Manager.

-

-

Chapter 4
Scaffold Elements Main Window

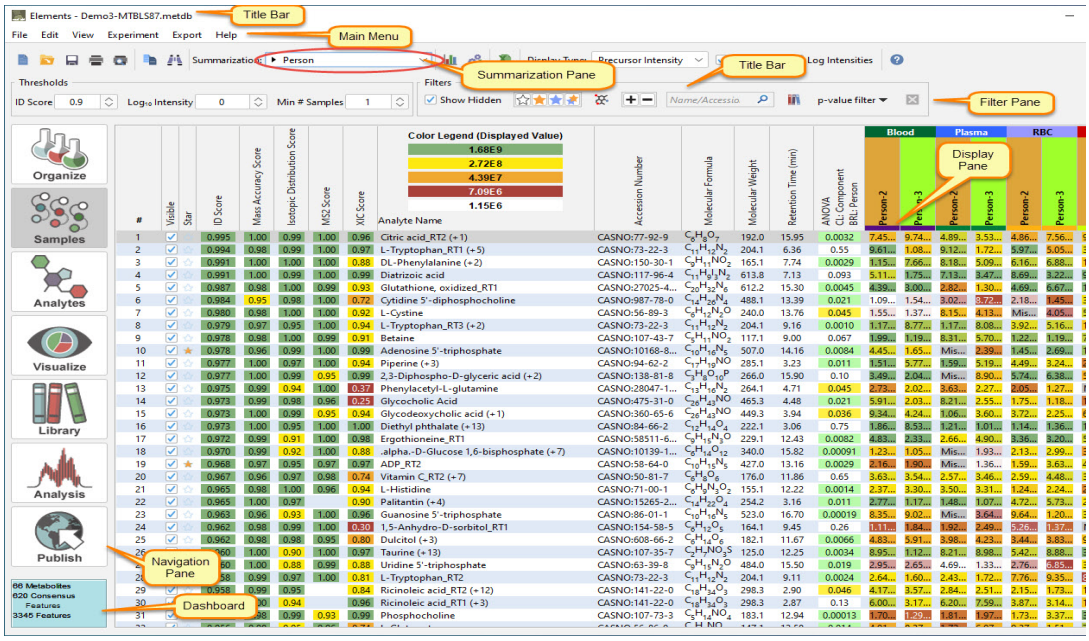
Like most Scaffold applications, Scaffold Elements consists of a main window which provides access to a number of specific views. In each view, the data loaded into a Scaffold Elements experiment are organized so that a user can easily view the results from various points of view.

The Scaffold Elements Main Window provides quick access to all of the Scaffold Elements features and functions through the following features:

- The “Title bar” on page 46
- The “Main menu commands” on page 47
- The “Tool-bar” on page 56
- The “Display Type Bar” on page 57
- The “Navigation pane” on page 64
- The “Summarization Bar” on page 66
- The “Display pane” on page 67
- The “Dashboard” on page 65

-

Figure 4-1: Scaffold Elements window



Title bar

Figure 4-2: Title bar

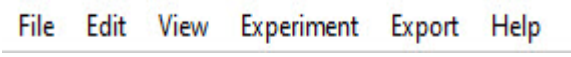
The title bar at the top of the Scaffold Elements window always displays “Elements”. Additional text is displayed in the title bar depending on the actions that the user has performed. For example, when a new experiment is created, the default experiment name “-Elements_Experiment” is appended. When a file is saved with a different name, the default name is replaced by the new file name. When a .METDB file is opened, the title bar displays the file name.

*The version of Scaffold Elements in use is not displayed in the Title bar. The version may be accessed through the **Help > About Elements** option in the main menu. See “Main menu commands” below.*

-

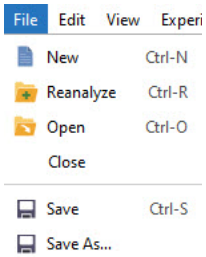
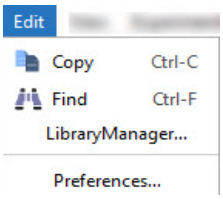
Main menu commands

Figure 4-3: u

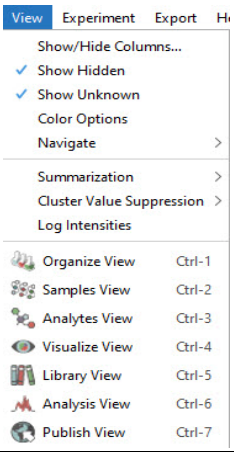
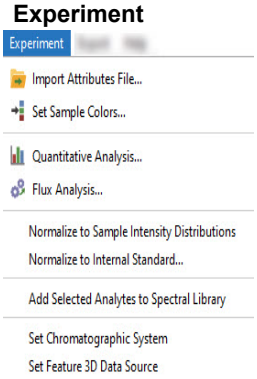


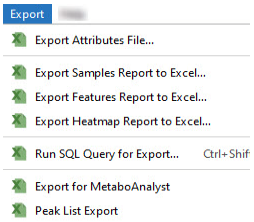
The Scaffold Elements main menu is organized in a standard Windows menu format with commands grouped into menus (File, Edit, View, Experiment, Export and Help) across the menu bar.

Some of these menu commands are also available in other areas of the application.

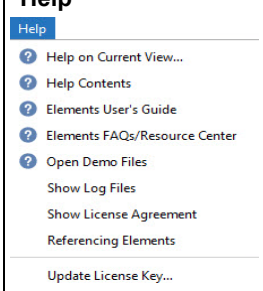
Menu	Menu Commands
File  	• **New**—Starts a new experiment and opens a file browser to allow the selection of files to load into Scaffold Elements. See "Loading data in Scaffold Elements" on page 32. • **Reanalyze**—Add files or modify parameters and reanalyze the experiment • **Open**—Opens a saved Scaffold Elements experiment file, *.METDB, through a file browser. • **Close**—Closes the current experiment, standard Windows behavior. • **Save**—Saves the current experiment, standard Windows behavior. • **Save As**—Saves the current experiment offering the option to use a different name, standard Windows behavior. • **Print**—Prints the current view. • **Print Preview**—Previews the current view with the option of printing the document. • **Exit**—Closes the Scaffold Elements window.
Edit  	• **Copy**—For each View, copies the first table appearing at the top of the View to the clipboard so it can be pasted into a third-party program such as Excel or Microsoft Word. • **Find**—Opens a find dialog box that searches the first table present in the Current View • **Library Manager**—See Library Manager • **Preferences**—see View

-

Menu	Menu Commands
View  	• **Show/Hide Columns...**—Opens the Tables Column Control menu • **Show Hidden**—Show table rows for which the visibility chceckbox has been unchecked • **Show Unknown**—Displays the unidentified compounds, despite their not meeting thresholds. • **Color Options**— Opens the Adjust Display Options dialog, see • **Navigate** — Allows selection of tabs in a View. • **Summarization**—Equivalent to the Summarization Bar pull down menu • **Cluster Value Suppression**—Helps the user select the clustering values that need to be visible, see Cluster Value Suppression • **Log Intensities** — Toggles the Log Intensities checkbox. Determines whether intensity values are shown as base 10 logarithms.
Experiment  	• **Import Attributes File**— See Import Attributes File... • **Set Sample Colors** - Switches sample coloring mode • **Quantitative Analysis**— See Configure Quantitative Analysis dialog • **Flux Analysis**—Opens a dialog to launch a metabolomic flux analysis. See Flux Analysis in Scaffold Elements • **Normalize to Sample Intensity Distributions**- See • **Normalize to Internal Standard**... • **Add Selected Analytes to Spectral Library**—Opens the dialog to “Add Experimental Spectra to Library...” • **Set Chromatographic System**— Allows setting the Chromatographic System outside of the loading process • **Set Feature 3D Data Source**— Connects the file to the mz5 and index files needed for 3D visualization of the full data landscape

Export
 
 	
	• **Export** :
	  • **Attributes file...**—Generates a tab-delimited text file of the meta-data attributes assigned to each ms sample in the current experiment, see Sample Organization tree table.
	  • **Samples Report to Excel...**—Generates a tab-delimited text file of the Samples table appearing in the Samples View, that can be opened and viewed in Excel.
	  • **Analyte Features Report to Excel...**— Generates a csv file containing information from the Analytes View for all analytes.
	  • **Heatmap Report to Excel...**—Generates a tab-delimited text file of the information depicted in the Heatmap appearing in the Visualize View, that can be opened and viewed in Excel.
	  • **Run SQL query for Export...**—Opens the SQL dialog box see SQL Export tab
	  • **Export for MetaboAnalyst**— Generates a file formatted for import into the MetaboAnalyst online analysis tool.
	  • **Peak List Export**—Exports peaklist of all features in the experiment

-

Menu	Menu Commands
Help  	• **Help on Current View**—Opens the Online Help that is specific for the currently displayed topic. • **Help Contents**—Opens the Contents page for the Online Help. • **Scaffold Elements User's Guide** —Opens the current Scaffold Elements User's Guide. • **Elements FAQs/Resource Center**—Opens the user's default web browser to the Home page of the Proteome Software's resource center. • **Open Demo Files**—Opens the folder where Scaffold Elements demo files are stored. The user can choose any of the pre-loaded files to test Scaffold Elements capabilities. • **Show Log Files**—Opens the folder containing Scaffold Elements error_log and output_log files • **Show License Agreement** • **Referencing Scaffold Elements**—Opens the Online Help that contains a sample of how to reference Scaffold Elements when publishing data analyzed with this application. • **How to Purchase** — (in Viewer mode) Opens the user's default web browser to the Purchase page of the Proteome Software web page: www.proteomesoftware.com/products/purchase.**Update License Key...—**(in licensed mode) Opens a dialog where the user can paste a purchased license key that will allow full use of the application. For more info see The first time Scaffold Elements opens after installation, the Enter License Key dialog box opens.. When a full licensed application is in use this option is not visible. • **About Elements**—Provides the release information for the current version of Scaffold Elements, license information, contact information for Proteome Software, Inc.. It also reports information about the system where Scaffold Elements is installed, the amount of memory available to the software and the percentage of memory used by the application.

Edit

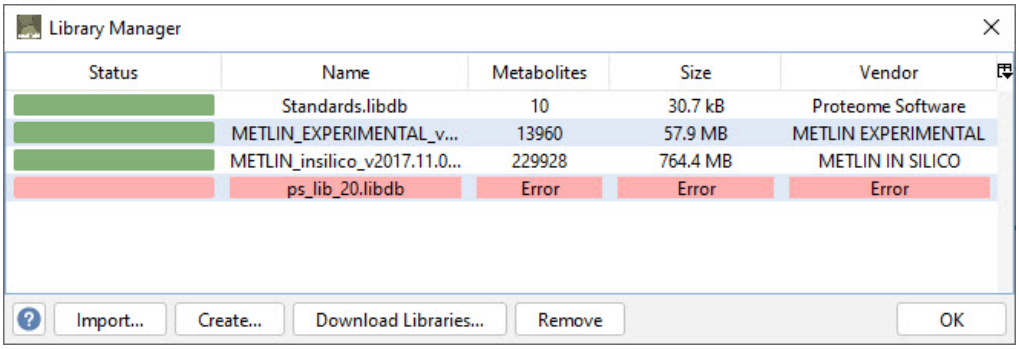
Some commands appearing under the Edit Menu:

Library Manager

The **Library Manager** dialog contains the **Imported Spectral Libraries** table, four action buttons and a **Help Online** button. The **Imported Spectral Libraries** table shows the listing of spectral libraries already loaded into Scaffold Elements and available for selection. The action buttons can be used to populate or pare the content of the **Imported Spectral Libraries** table and to confirm currently selected libraries to be imported into the **Search Library** tab.

-

Figure 4-4: Library Manager Dialog

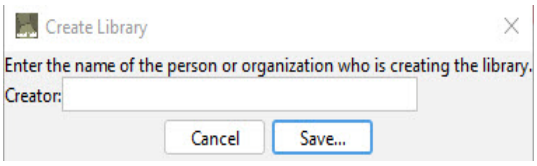


Selecting the **Import...** button opens a file manager so the user can navigate to the location of a spectral library of interest. The accepted file formats are: SDF (Structured Data Format), LIBDB (the Elements spectral libraries format), MSP and custom spectral libraries stored as tab delimited text files and saved with extensions TXT or TSV. For more information on how to create a custom spectral library using a text file, see Creating A Personal Spectral Library in the appendix.

When one or more spectral libraries have been selected, the status column in the table displays the progress of the library loading process.

Each imported library occupies a row in the table. Columns in the table display each library's: Status, Name, Entries, Size, Vendor and Location.

The **Create...** button is used to create an empty library to which experimental spectra may be added to create a personal library. The button brings up a dialog box asking for the **Creator** of the library. This information will be displayed in the Vendor column of the Library Manager.



After a string has been entered in the Creator field, clicking **Save...** opens a file browser which allows the user to select a location for the new library file. An empty library is created. Spectra must be added to the library before it can be used in searching.

The **Download Libraries...** button, opens a local browser to the **Download Elements Library** page where spectral libraries compatible with Scaffold Elements are available for download.

Depending on the type of license under which Elements is operating, different libraries may be downloaded. A purchased copy of Elements gives access to the NIST Mass Spectral Library (NIST/EPA/NIH Mass Spectral Library (EI)) and for an extra fee, also the METLIN library *(WILEY/METLIN Mass Spectral Database)*. For evaluation purposes, temporary and

limited access to these libraries is provided for the time defined by the evaluation key.

Download the library files through the Download Elements Library page. If the downloaded file is a zip file (the file name contains the ".zip" extension), unzip the file. Place the file(s) in a permanent location from which Elements may access them. To add the downloaded libraries to Elements, use the Import... button in the Library Manager.

The NIST library (NIST20 Tandem Library) as distriburted with Scaffold Elements consists of two different files in Elements LIBDB format:

NIST HRAM MS/MS Library (hr_msms_nist) - high resolution, accurate mass spectra only. This libarary contains 1,026,712 spectra of 94,472 precursor ions from 27,840 chemical compounds.

Combined NIST HRAM and NIST Low Resolution MS/MS library (hr_lr_msms_nist) - In addition to the spectra contained in NIST HRAM, this library also contains 215,649 lower resolution spectra of 163,512 precursor ions from 28,559 chemical compounds.

For more information about the NIST20 spectral libraries, see https://www.nist.gov/srd/nist-standard-reference-database-1a.

The METLIN Mass Spectral Library[1], which was created by The Scripps Center for Metabolomics and is distributed by Wiley, consists of three separate library files:

METLIN Experimental[2] - contains over 72,000 experimental spectra for over 14,000 chemical compounds.

METLIN In Silico[2] - with over 699,500 in silico spectra for over 233,000 compounds. Each of these unique chemical structures features an in-silico spectrum at collision energies of 10, 20, and 40 eV.

METLIN Spectraless [2]- provides information such as name, structure, elemental formula, mass, CAS number, systematic name, KEGG ID, HMDB ID, PubChem ID, and commercial availability for over 700,000 compounds. No MS2 spectra are included.

- The Elements for Metabolomics End User may not redistribute the NIST/EPA/NIH Mass Spectral Library in any manner without explicit written approval by NIST. Contact Proteome Software at for more information.
- Copyright protection on the compilation of data in this Library is secured by the US Department of Commerce in the United States and in other countries that are parties to the Universal Copyright Convention, pursuant to Section 290(e) of Title 15 of the United States Code.
- It is expressly understood and agreed that unauthorized copying of this Library is not permitted.

1. METLIN Mass Spectral Database: Scripps Center for Metabolomics, Gary Siuzdak (Editor), H. P. Benton (Editor), ISBN: 978-1-119-37705-4, October 2017
2. Scripps Center for Metabolomics, Wiley 2017

-

- The Elements for Metabolomics End User must comply with the Wiley METLIN End User Agreement. see http://www.proteomesoftware.com/company/metlin-eula/ for more information.
- It is expressly understood and agreed that unauthorized copying of these Libraries is not permitted.

The **Remove** button allows elimination of imported spectral libraries from the list. When selected, a library row is highlighted in blue; clicking **Remove** deletes the highlighted libraries from the table. Before the operation is finalized, a confirmation dialog appears offering, as an option, to completely remove the libraries from the computer. If this option is selected, the .libdb file is deleted from the computer, otherwise the file remains available and may be imported again and added to the list in the future.

Preferences

The main menu option: **Edit > Preferences**, opens the **Preferences** dialog which contains the following tabs:

- "Accession Number" on page 52
- "Web Links" on page 52
- "Units" on page 53
- "System" on page 53
- "User Interface" on page 53
- "Proteowizard" on page 54

Accession Number

This tab includes a list of accession number or identifier formats that Elements can use to label analytes, ordered by priority of selection. Under the list there are five action buttons and a check box.

The highest priority accession number identifier is located at the top of the list. The order may be changed by selecting a row and dragging it to the desired level.

When reporting an identified analyte, Elements will choose the first identifier type in the list which can be found in the library entry.

Functional buttons and check boxes in the dialog tab:

- Save this order as User Default--The user can check this box to save the order he prefers.
- Reset to User Default--When a custom default order has been previously saved, this button will readjust the list according to that preferred order.
- Reset to Elements Default --Set the order to the original sequence.

Web Links

The tab allows the user to choose the online database to be used for accession number searches when web links are clicked in the Accession Number field of the Samples View.

-

Units

The tab contains a drop-down list for selection of the preferred retention time units.

System

This tab provides a number of options related to system settings:

- Memory Usage - allows the user to specify the amount of RAM to be allocated to Elements. The more RAM allocated the better the program will perform. It is recommended that this be set to approximately 75% - 80% of the physical RAM available on the system.

.

- *The new memory setting will take effect only after the application has been closed and restarted.*
- *MAC OS, once Elements Viewer is installed, does not allow the memory allocation to be reset unless the user is an administrator. A non-administrator user can update the memory only when reinstalling the software.*

- Number of Processors - allows selection of the maximum number of processors available to Scaffold Elements for threading computations. The default value is the maximum number of processors available on the system where the application is installed.

  Internet Settings - determines whether the program is permitted to connect to the internet. If this is allowed, an option is provided to specify an HTTP proxy server. Proxy servers may be used by an organization's IT departments to filter communications to and from the Internet. If a proxy is in use, the user needs to set the Proxy Server Name and Port Number.

  If the user is unsure whether there is a need to use proxy server settings, he/she may check how his/her web browser is connected to the internet.

User Interface

This tab offers options related to the function of the Elements graphical user interface.

- Search Fields - a checkbox allows the user to specify whether or not search fields should be interpreted as regular expressions.
- Messages - a button allows users to re-enable any messages that have been disabled by checking a box labeled “Do not show this message again” in any dialog.
- Views - Chooses the View that will open by default when files are loaded or when an Elements file is opened.

-

Proteowizard

- ProteoWizard location - Allows the user to select the msconvert.exe file to be used for conversion of raw files to MZ5 format. MSConvert is a program in the ProteoWizard suite, and the executable file may be found in the folder into which ProteoWizard was installed. Generally this will be in a folder named C:\ProteoWizard followed by the version number.
- Download ProteoWizard - Allows the user to download the recommended version of ProteoWizard. After downloading and installing, use the ProteoWizard location to link Scaffold Elements to the new msconvert.exe.
- Create filtered mz5 files - If this box is checked, Elements instructs MSConvert to filter the data as raw files are converted to MZ5 format. The thresholds used will be the noise threshold and retention time range selected in the Search Parameters Tab.

  If this option is checked and the user has set a retention time range and/or a noise threshold using specific value, the mz5 files created by msconvert will not contain any raw data outside of the specified retention time range and/or noise threshold. This will affect isotopic feature cluster formation, since no raw signal below the noise threshold will be included in the isotopic peaks, and browsing the raw data landscape, where only filtered data will be shown.

View

Some items appearing in the View menu:

- **Color Options** - Selecting the command **View > Color Options** opens the dialog **Edit Coloring for Display Type "<selected display option>"**.

Through this dialog, the user can adjust the coloring used to highlight the range of values of the selected Display Type appearing in the Samples Table. The new colors are then reflected in the color legend appearing at the head of the Samples Table.

Controls are available to set the specific color for each interval as well as to define the range of values over which the color scheme is to be applied. Sliding one of the colored squares located above the legend changes the range of the selected color. A color range can also be set by typing a value in the **Selected Value** box. The color gradient check box determines whether a color gradient or discrete colors will be applied over the designated ranges.

Double clicking a color in the color bar or a colored square opens a color choice dialog. This dialog allows selection of different colors using either swatches, HSV, RGB or other methods. By double clicking a specific colored square the user can change the specific color using the color choice dialogsee Figure 4-7.

The button **Restore Scaffold Elements Defaults** provides a way to go back to the default settings of the program. The **OK** command finalizes the current changes.

- **Show/Hide Columns** - see Tables Column Control

-

- Show Unknown - allows unidentified analytes to be displayed even though they do not have scores and therefore do not meet any scoring threshold in effect.
- **Navigate** - selects the tab to be displayed. The options are to Select previous tab or Select next tab. These options are disabled when the currently displayed View contains only a single tab.
- **Summarization** - see Summarization Bar
- **Cluster Value Suppression** - see Cluster Value Suppression
- **Log Intensities** - when this item is checked, intensity values in the Samples View are displayed as base 10 logarithms. When it is not checked the intensity values are not shown in log form. This selection coordinates with the Log Intensity checkbox.

Cluster Value Suppression

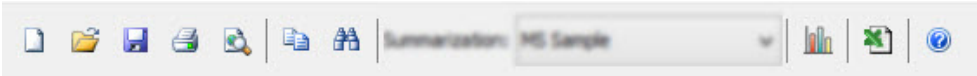
This option allows the user to show or hide the rolled up values at the cluster level. The option is meant to help the user navigate through the samples list when clusters are applied.

- **Show Values for all clusters** - The rolled up values are shown in any of the rows that represent the top level of a cluster
- **Only Show values for collapsed clusters** - The rolled up values are shown only for the rows that represent the top level of a collapsed cluster
- **Do not show values for clusters** - The rolled up values are not shown for any of the rows that represent the top level of a cluster

-

Tool-bar

Figure 4-5: Scaffold Elements tool bar



The Scaffold Elements tool-bar contains icons that represent equivalent commands for frequently used main menu options.

Icon	Function
	New—Starts a new experiment and opens a file browser to allow the selection of files to load in Scaffold Elements. See "Loading data in Scaffold Elements" on page 32.
	Open—Opens a saved Scaffold Elements experiment file, *.METDB through a file browser.
	Save—Standard Windows behavior.
	Print—Prints the current view.
	Print Preview—Previews current view with the option to print the document.
	Copy—For each view copies to the clipboard the first table appearing at the top of the view. From there, the user can paste it into a third-party program such as Excel or Microsoft Word.
	Find—Opens a find dialog box that searches the first table present in the current view
	Quantitative Analysis—See Configure Quantitative Analysis dialog.
	Flux Analysis - Launch a Metabolic Flux Analysis
	Excel—Exports the information that is contained in the current view to a tab-delimited text file that can be opened and viewed in Excel.
	Help—Opens the Scaffold Elements Online Help.

-

Display Type Bar

The Display Type bar contains:

- the Display Type: pull down list.
- the Normalized Check Box.

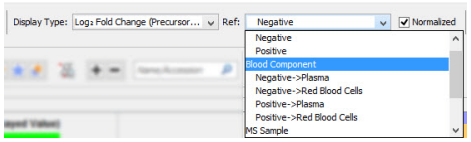
Their functionality affects the displayed values in the Samples View and in the Visualize View.

Display Type:

The Display Type drop down list offers a range of sample statistics. The values shown in the Samples table and in the Heat map depend on the display type selection. When a particular quantitative value is selected from the list, the corresponding values are reported in the table cells for each analyte group and MS sample or chosen level of summarization.

The **Display Type** drop down list offers the following options:

Figure 4-6: List of Display Types



- **Log10 Precursor Intensity**---Which shows, rolled up to the selected summarization level, the Log_{10} of the precursor intensity (or Area under the Curve) of a particular analyte in each sample or summarized group of samples.
- **Log_2 Fold Change (Precursor Intensity) ---** When selected, the **Ref:** pull down list appears, allowing the user to choose a reference against which the Log_2 Fold Change is to be computed. The menu lists the columns or groups of columns in the Samples table representing the available reference options compatible with the selected level of summarization.
- **Max Score--** Shows the maximum ID score value of the identified features for a specific analyte at the selected summarization level, see "ID Score" on page 182".

---

*The default coloring of each cell in the Samples table reflects the selected display type. Coloring can be changed in the Edit Coloring for Display Type dialog that is reached through the command **View > Color Options**. See "Color Options" on page 60.*

Normalized Check Box

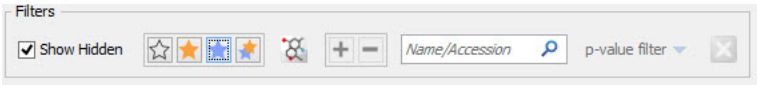
When this option is selected the quantitative data shown in the Samples table, chosen from the Display type list, is normalized using a normalization method described in "Normalization methods" on page 154.

-

Filters Pane

The Scaffold Elements Filters pane is located under the Display Type bar in the upper section of the main Scaffold Elements window. It contains an ample selection of filtering options that help the user search and choose a subset of the list of analytes included in the Samples table. Whenever one or more filters are in use, the grayed out cross icon present at the right end of the pane becomes active appearing red in color. Clicking on the cross cancels the action of any active filter.

Figure 4-7: Scaffold Elements Filters Pane



The Filters pane includes the following functionalities:

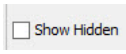
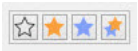
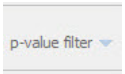
Icon	Function
	**Show Hidden**—Toggles the view of hidden analytes listed in the Samples table according to the status of the check-box appearing under the column visible. By default the status of the box is checked.
	**Star Filters** — A series of toggle buttons that can be used to filter analytes tagged with a specific star in the Samples Table, see "Star Filter" on page 58.
	Substructure Filter—Filters the list of analytes according to a chemical substructure custom drawn through the Chemical Substructure Filter dialog. See "Substructure Filter" on page 59.
	Mode Filter — When the data was measured in a mode, this filter helps selecting the features resulting from a selected mode. See "Mode Filter" on page 60.
	Text Search Box — Filters the analyte list according to the input text, see"Text Search Box" on page 61.
	**p-value filter** — When a statistical test is applied and the p-value column has been added to the table, this filter becomes available and offers the option to filter based on statistical significance, see "P-Value Filter" on page 62.
	Cross Icon—Red when a filter is active.Clicking cancels filtering.

Figure 4-8:

Star Filter

The Star Filer box contains four toggle buttons located next to the "Show Hidden" check box. Each button is characterized by one of the four possible star states the user can trigger

for a specific analyte group or cluster, by clicking the icon shown under the “Star” column in the Samples table. This action is called “starring an analyte”.

Each star button has two possible filtering states:

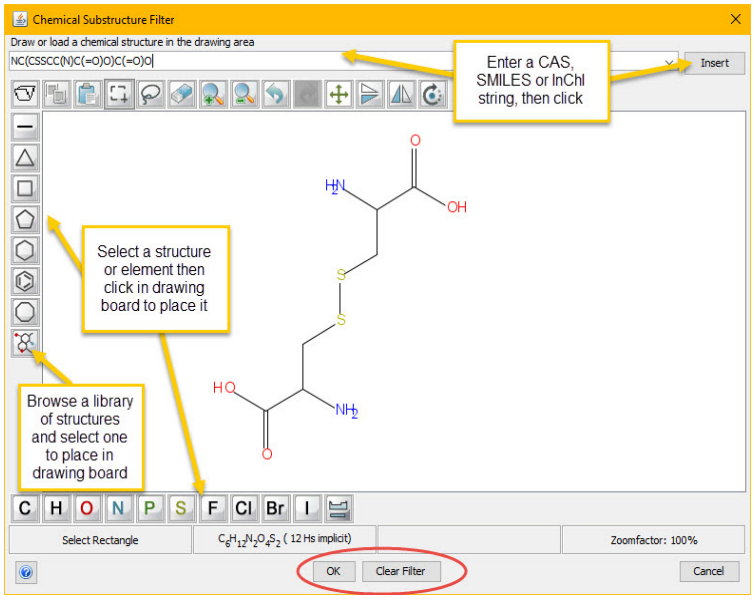
- **Unselected** - The star appears at the center of a squared box and analytes tagged with that type of star are included in the Samples table. When the button is clicked, a red diagonal bar appears across the box and the analytes that are tagged by this specific star disappear from view. The Cross icon becomes active, indicating that a filter has been applied.
- **Selected** -The star appears at the center of a squared box barred with a red diagonal and the analytes tagged with that type of star are filtered from the list. Clicking the button clears the red bar and returns the analytes to the list. If no other filters are in effect, the Cross icon is grayed out.

It is possible to select one or more star filter buttons at the same time. The analytes tagged with the selected stars will be hidden from the analyte list. Selecting the uncolored star leaves only starred analytes in the list. For more information, see “Tagging Analyte of Interest, the star function” on page 99.

Substructure Filter

Selecting the Substructure filter button opens the Chemical Substructure Filter dialog.

Figure 4-9: Chemical Substructure Filter dialog



The substructure filter allows the user to specify a specific chemical substructure and to filter the analyte list to only those analytes containing this substructure. The dialog contains a drawing board with three tool bars along its edges, a textbox for entering a search string, and buttons for applying or clearing the substructure filter. There are three methods of adding

-

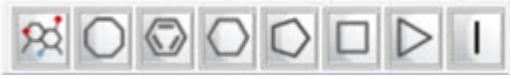
structures to the drawing board:

- Enter a CAS, SMILES or InChI string into the box (it will populate the dropdown and may be selected from there) and click Insert.
- Select specific structural elements from the left tool bar and/or elements from the lower tool bar and place them into the drawing board by clicking in the desired location.
- Browse a library of analytes and select one, which will be placed into the drawing board.

Once a structure has been placed in the drawing board, it may be edited using the Editing and Copying toolbar functions. When the substructure has been specified in the drawing area, clicking the OK button filters the analyte list and closes the Filter. Dialog. Clicking Clear Filter clears the drawing board and closes the Filter Dialog.

Editing and Copying functions tool bar

5. **Draw Bonds and Atoms** tool bar



6. **Draw Chemical Symbols** tool bar

Mode Filter

MS data can be acquired in three possible modes:

- positive, negative or in a mixed mode.

This filter is only active when the experiment contains both ions obtained in positive and in negative mode (e.g. data acquired in mixed mode). Clicking on the + button filters out analytes with spectra acquired in the positive mode, displaying only the analytes identified only by negative ions. Clicking on the - button filters out analytes with negative ions. Once selected, the button shows a red diagonal bar. and a red button at the end of the Filters pane is activated. Clicking the red button or toggling the + or - button sign one more time clears the filter.

-

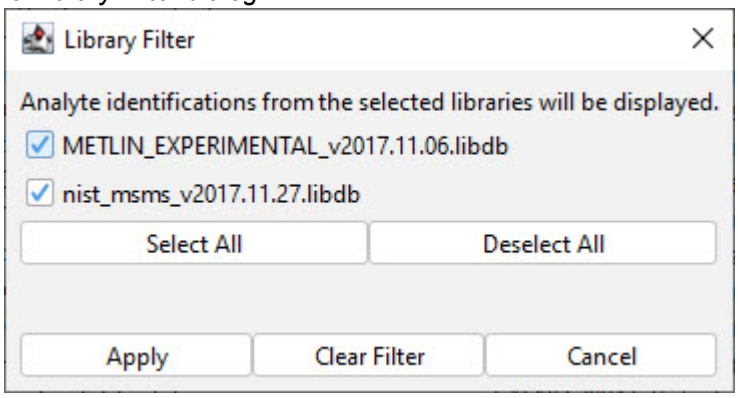
Text Search Box

The **Text Search box** filters the list of analytes in the Samples Table according to what has been typed in the box. The filter searches for the typed characters in the Analyte Name and Accession Number columns in the Samples Table.

It is possible to allow the use of regular expressions in the search field by going to the menu option **Edit > Preferences** and selecting the appropriate option in the **Searching** tab.

Library Filter

Figure 4-10: The Library Filter dialog



When more than one spectral library has been searched, the Library Filter allows the user to choose to display only identifications made from specific libraries. All libraries with matches in the current experiment are shown, and the user may check the libraries for which matches are to be displayed.

The option to Show Unknown in the View menu actually works through the Library Filter, selecting the compounds which were not identified through any library.

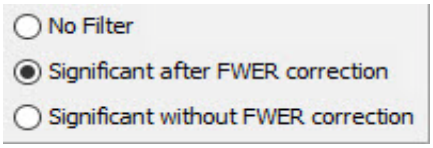
-

P-Value Filter

The filter is active only when a statistical test has been applied. When applying a test, the user may select the p-value threshold below which a result is considered statistically significant, and may also apply a FWER correction to this significance level to adjust for multiple comparisons, see "Significance Level" on page 160.

When a statistical test has been applied, a column containing the resulting p-values appears in the Samples Table, with p-values above the specified significance level colored green. P-values that meet the uncorrected significance level, but do not meet the adjusted significance level after applying a selected FWER correction are colored yellow. The p-value filter allows the user to filter out all analytes whose p-values do not meet significance criteria, either with or without correction.

Figure 4-11: p-value Filter

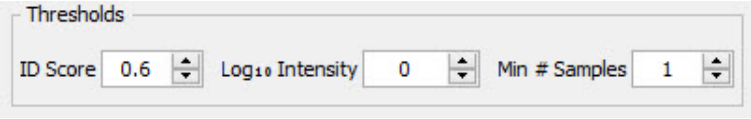


-

Thresholds Pane

Elements provides three different thresholds that can be used to increase or decrease the length of the analyte list appearing in the Samples Table. Their function is to set minimum standards for identification confidence. Thresholds can be adjusted through the Thresholds pane located in the main Elements window under the Tool bar. The level of a threshold can be changed either by typing a value in the text box or by adjusting the value using the up and down arrows.

Figure 4-12: Thresholds Pane



The pane includes the following thresholds:

- "ID Score"
- "Log10 Intensity"
- "Min # Samples"

ID Score

The ID score is a weighted average of the "MS2 Score" on page 216, the "Isotopic Distribution Score" on page 214, and the "Mass Accuracy Score" on page 214, for more information see section "Analyte ID Score" on page 213.

Log_{10} Intensity

Where intensity refers to the area under the curve or precursor intensity of the XIC associated to an analyte feature.

Min # Samples

The Min # Samples or reproducibility threshold filters analytes groups that are present in more MS samples than the number appearing in the filter.

-

Navigation pane

The Scaffold Elements Navigation pane is a vertical bar displayed on the left side of the Scaffold Elements window.

Figure 4-13: Scaffold Elements Navigation pane View selection

The bar contains large buttons that toggle the seven different views available in the Scaffold Elements main window:

- The Organize View, see The Organize View
- The Samples View, see The Samples View
- The Analytes View - see The Analytes View
- The Visualize View, see The Visualize View
- The Library View - does not appear in the Elements Viewer, see The Library View
- The Analysis View, see Analysis View
- The Publish View, see The Publish View

-

Dashboard

The **Dashboard** or **Information Box** is located under the navigation pane on the left lower corner of the Scaffold Elements main window. The box contains information about the numbers of Identified Analytes, Consensus Features, and total Features at the current level of thresholding. The information reported in the Info Box when highlighted can be copied by simply using the standard CTRL C key strokes.

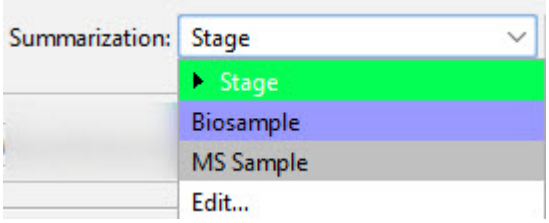
Figure 4-14: Dashboard

-

Summarization Bar

The Summarization bar allows the user to view the data collapsed or expanded at different hierarchical levels. The Summarization bar operates through a drop down menu containing a list of Categories hierarchically ordered.

Figure 4-15: Scaffold Elements Summarization bar



The last item in the list is the command **Edit...**, which, when selected, opens the The Configure Sample Organization and Statistical Analysis Dialog. through which it is possible to add Categories to the drop down list and define a different hierarchical order.

-

Display pane

The information included in the different views appears in the Scaffold Elements Display pane. Depending on the view, the type of information reported might appear framed in one or more tables or graphs included in one or more sub-panes. All panes and tables included in Elements share the following characteristics:

- Tool-tips
- Resizing of columns and panes
- Tables Column Control
- Moving columns
- Column sorting feature
- Multi selection of rows

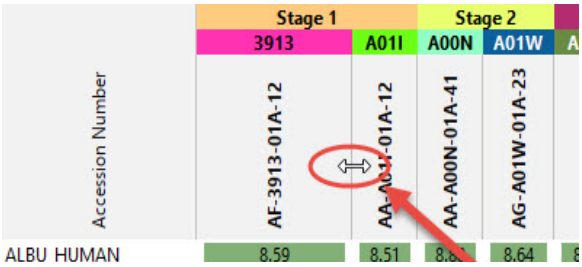
Tool-tips

The user can view information about fields or columns in a View by just hovering the mouse pointer over the location of interest.This operation opens a collapsed tool-tip. Pressing F2 opens an expanded tool-tip. Pressing the Escape (ESC) key on the keyboard closes the expanded tool-tip.

Resizing of columns and panes

The user can resize columns and different panes in each of the views to better suit his/her working needs. For example, in the Samples Table, the user can change the width of a column by resting the mouse pointer on the right side of a column heading until the pointer changes to a double-headed arrow, and then dragging the boundary until the column is the width that he or she wants.

Figure 4-16: Changing the width of a column in the Samples View

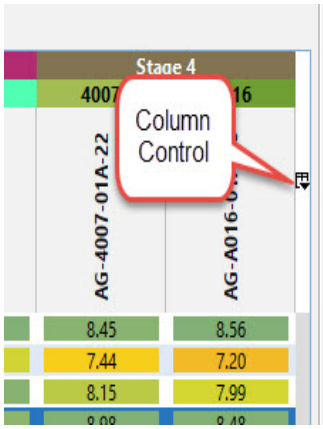


Tables Column Control

All tables throughout Scaffold Elements have a feature called Column Control. It is a vertical button placed on the right side of every table lined up with the column headers. When the user clicks the button or selects the option **View > Show/Hide Columns** from the main menu, a drop down list of all the columns opens. Each column is associated with a check box and at the bottom of the list there are included three group commands.

-

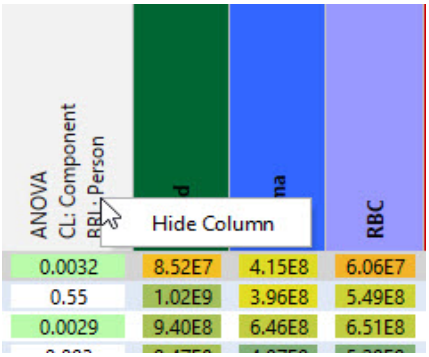
Figure 4-17: Scaffold Elements Column Control button



- Unchecking columns from the list will hide them from view in the Samples table.
- The Horizontal Scroll command if checked will add a scroll bar at the bottom of the Samples table.
- Pack all columns when selected resizes all samples columns to a common width.
- Pack selected column resizes the column that contains the current selected cell. If no cell has been selected the command is grayed out.

Columns can be hidden also by using the right-click function which brings up the context menu Hide Column when hovering over the heading of a column. To have the column reappear again use Tables Column Control

Figure 4-18: Hiding a column with Mouse Right-click



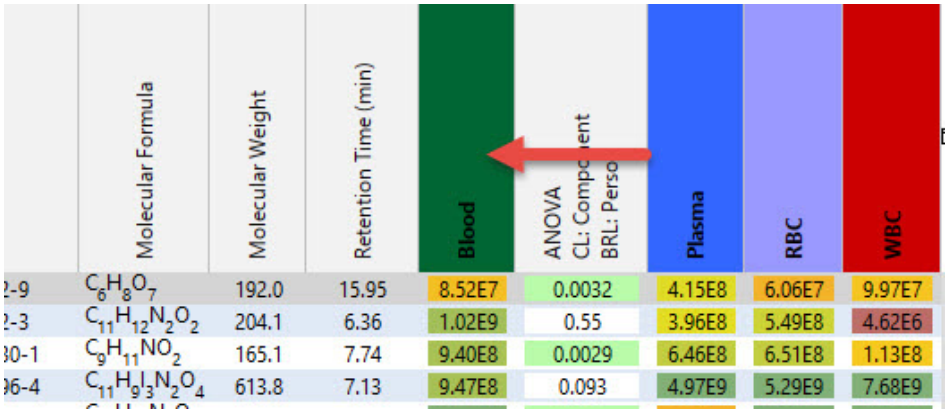
Moving columns

In all tables throughout Scaffold Elements, every column can be moved from one position to another for more comfortable access to the data that is summarized in them.

The user simply clicks on the header of the column that is be moved and drags it to the new

location. The new position will be retained when the user switches to another view and then returns.

Figure 4-19: Moving columns in tables



	Molecular Formula	Molecular Weight	Retention Time (min)	Blood	ANOVA CL: Compo ent BRL: Perso	Plasma	RBC	WBC
2-9	$C_6H_8O_7$	192.0	15.95	8.52E7	0.0032	4.15E8	6.06E7	9.97E7
2-3	$C_{11}H_{12}N_2O_2$	204.1	6.36	1.02E9	0.55	3.96E8	5.49E8	4.62E6
30-1	$C_9H_{11}NO_2$	165.1	7.74	9.40E8	0.0029	6.46E8	6.51E8	1.13E8
96-4	$C_{11}H_9I_3N_2O_4$	613.8	7.13	9.47E8	0.093	4.97E9	5.29E9	7.68E9

Column sorting feature

In all tables throughout Scaffold Elements, the user can use the tri-state column sorting feature and sort the display by clicking on any column header. For example, to sort the analytes based on their accession number, the user can click the Accession Number column header to initially select the column. Then to sort the analytes based on increasing alphabetical order, the user can click the Accession Number column header a first time. A second click will order the column on decreasing alphabetical order. To return to the default display, the user can click the Accession Number column header a third time.

Increasing and decreasing orders will be indicated by an up and down arrow respectively, shown in the header of the column that is being sorted, while the default order will have no arrow.

Multi selection of rows

In all tables throughout Elements the user can select multiple rows by using either the SHIFT or the CTRL key, depending whether the desired selection has contiguous rows or not, and the click of the mouse in a pretty standard fashion. Other functions can then be applied, such as assigning a star to the selected group of analytes in the Samples table, for example.

Mouse Right-click Context Menus

When the user hits the right-click button of the mouse while hovering over the Display Pane of a View, a menu with various options appears close to the working arrow. Depending on the selected view the list of options available in the menu varies. A description of the mouse right-click command is provided in "Description of Mouse Right Click Context Menu Commands" on page 267.

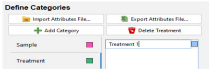
Organize View

There are two different context menus appearing in this view, one connected to the Define Categories Pane and the other to the Sample Organization tree table.

-

- When selecting a row in the Organize View Table, Right Click Menu A appears. The menu contains number of commands and a list of the Categories defined in the table. Each attribute group appears in its assigned color and provides in a sub-menu the list of attributes included in it.When selecting a row in the Categories tree table Right Click Menu A appears.The Edit commands opens up the Bulk Edit Sample Names and the Delete option deletes a selected attribute.

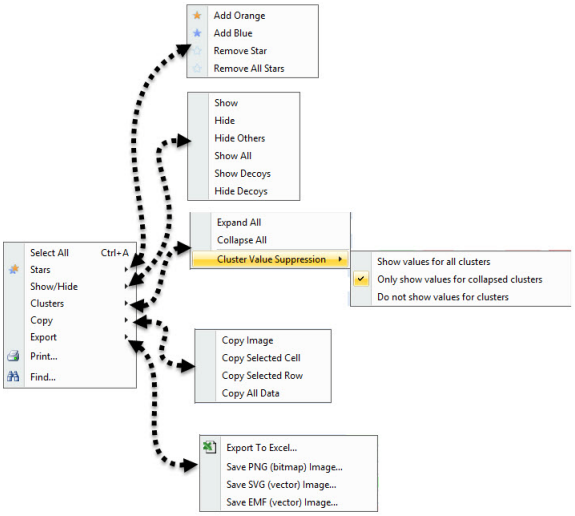
Right Click Menu A



Samples View

When hovering over the Samples table the Right Click Menu B appears. This menu has a number of sub-menus as shown in the picture.

Right Click Menu B



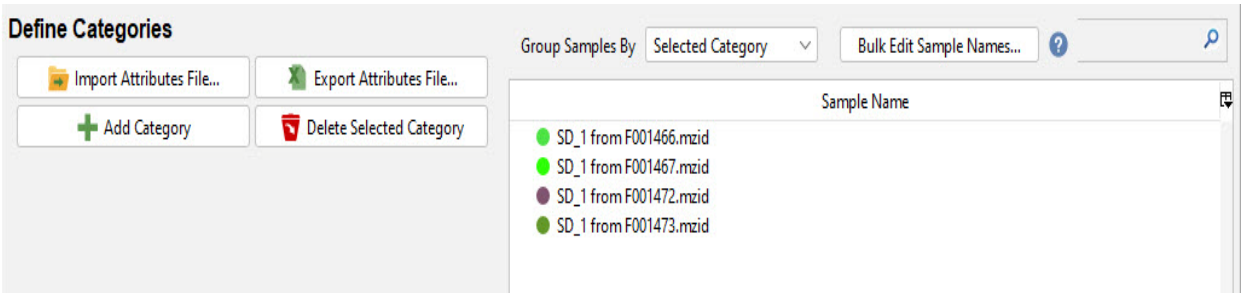
Chapter 5
The Organize View

The Organize view displays the loaded samples in a tree structure that can be customized to reproduce the organizational structure of the experiment to be analyzed in Scaffold Elements. It works in conjunction with the “Configure Sample Organization and Statistical Analysis” dialog to support data analysis to expose meaningful biological trends in the experiment.

Organizing data in Scaffold Elements

The Organize View in Scaffold Elements provides an easy method to organize even complex experiments involving multiple variables. A variable or factor in Scaffold Elements is called a Category, and each different level or value assumed by the variable or factor is called an Attribute. Through the Organize View, Categories, along with their Attributes are defined, and the appropriate Attributes are associated with the specific samples to which they apply.

For example, “Treatment Group” might be a Category, with Attributes “Control” and “Treated”. A time-course study measuring response to administration of a drug might have a Category called “Time” with Attributes “0 min”, “20 min”, “40 min” and “60 min”. Many Categories may be applied to the same data. Often many different attributes comprising many categories are recorded for clinical samples, such as age, sex, disease history, etc. All of these may be applied in Scaffold Elements, and the researcher may experiment with analyzing the data on the basis of any one or a combination of these categories.



Attributes may be defined and assigned either through the graphical user interface of the Organize View or by reading an Attributes File, which may be created in Excel and saved as a CSV, TSV or .TXT file or exported from a LIMS system.

Figure 5-1:

The Tools in the Organize View provide helpful ways to restructure the loaded samples to reflect the

-

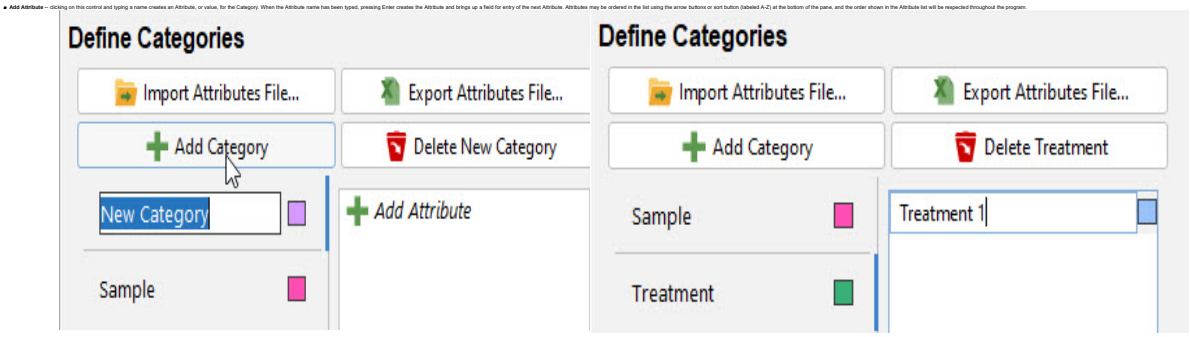
proper experimental design. This is done by defining Categories and their associated Attributes and assigning them to the samples. As Categories are added, additional columns are displayed, each corresponding to a Category. As samples are assigned Attributes, they are tagged with customizable labels and colors. Once the Attributes have been associated with the corresponding samples, the user can structure the experiment by creating a hierarchy of categories to view the data at different levels of summarization and can also apply a variety of statistical tests.

Tools in the Organize View

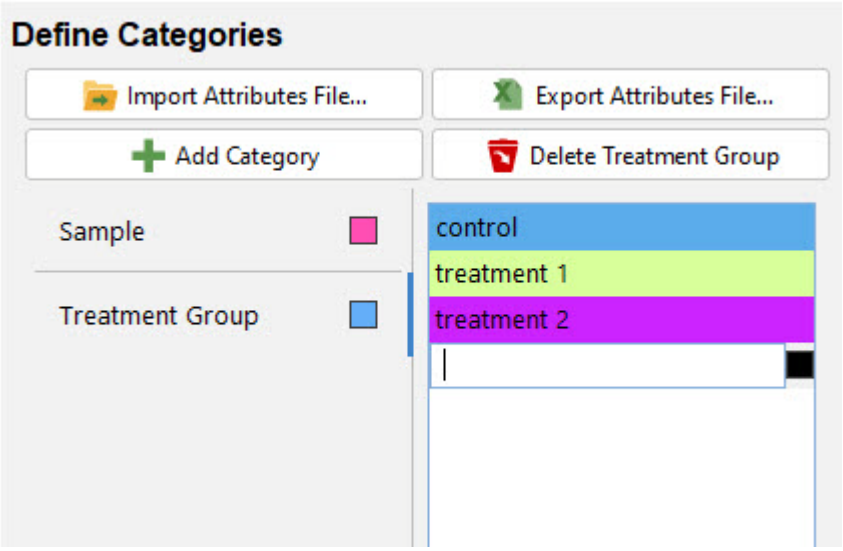
The Organize View consists of:

The Define Categories Pane -- which contains a number of tools to enable the user to provide the meta-data needed to organize the samples in accordance with the experimental design. It consists of:

- **The Import Attributes File... button**--which loads metadata that is already assigned to the samples and stored in a text file structured as a spreadsheet. It creates Categories and assigns Attributes to the currently loaded samples as specified in the file.
- **The Export Attributes File... button**--which saves the current Categories and Attribute assignments as an Attributes File. This is useful for saving Attributes created through the GUI or for exporting a skeletal Attributes File which can be completed in Excel and re-imported to create and assign Attributes.
- **The Add Category button** --which creates a new Category. When this button is clicked, a new Category is added at the top of the Category list below the button. By default, it is named “New Category”



-



- **The Delete Category button** -- clicking this button deletes the selected Category and all Attributes associated with it. To delete a single Attribute, select the Category in which it appears, then select it in the Attribute list and use the delete key or right-click and delete.
- **Configure Experimental Design and Statistical Analysis** -- when all samples have been organized according to their attributes, the user should click this button to open a dialog which allows for specification of the design of the experiment. This will establish a Summarization Hierarchy and allow the user to configure statistical analysis (see Experimental Design).

T**he Sample Organization Pane -** allows the user to assign attributes to the samples. It consists of:

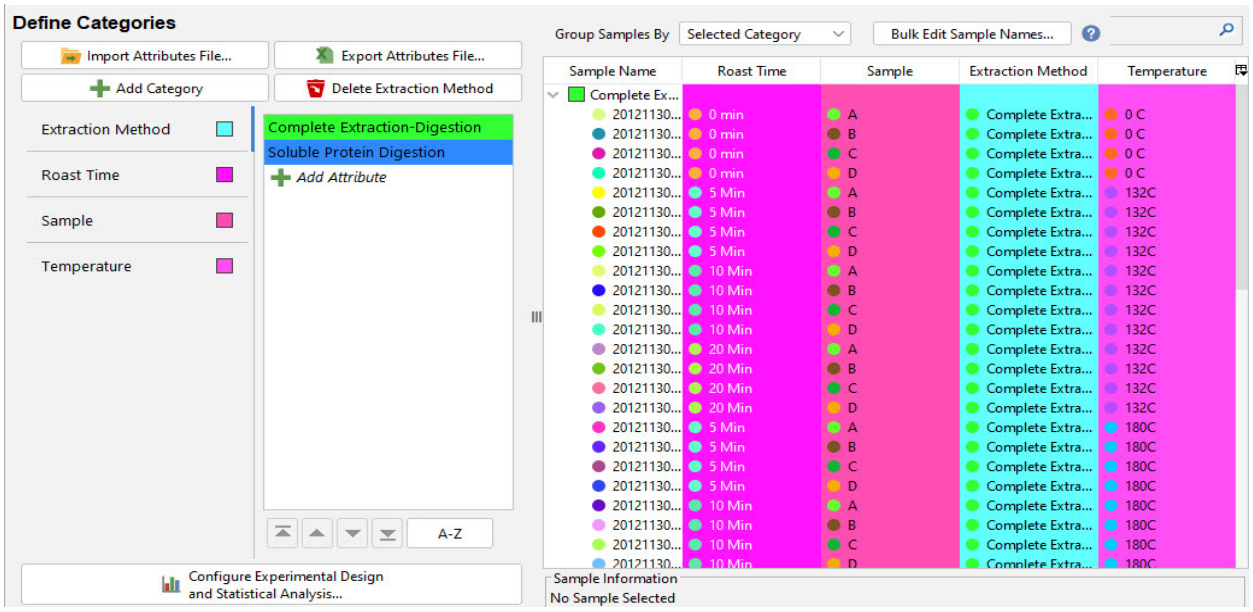
- **Group Samples By** -- a dropdown list which allows the user to determine how the samples will be displayed in the Sample Organization Table. Options are 1) to display the samples sorted by Sample Name, 2) to group the samples based on the Selected Category (the currently selected in the Category List under Define Categories), or 3) to display them hierarchically, organized according to the Experimental Design (see Experimental Design). Each of these options can be useful at different points in the process of organizing the experiment.
- **Bulk Edit Sample Names**... -- a button which brings up a dialog to assist the user in editing sample names to make them more useful and legible throughout the program. It provides a number of options for trimming the names or allows individual editing if the custom option is selected.
- The **Sample Organization Table** -- which lists the MS samples loaded into the experiment, organized as a tree structure and displays the Attributes associated with each sample.
- **Sample Information** -- displays sample information for the sample currently selected in the Sample Organization Table.
- The **Experimental Design** button which allows creation and editing of the Summarization Hierarchy.

-

Sample Organization Table

Scaffold Elements allows the user to derive a much deeper understanding of the experiment by creating new Categories and then assigning their Attributes to the appropriate MS samples. The Categories may be hierarchically organized using the Summarization pane, described in Experimental Design. Figure 5-2 shows the data after a series of attributes has been applied.

Figure 5-2: Organize View with added attributes

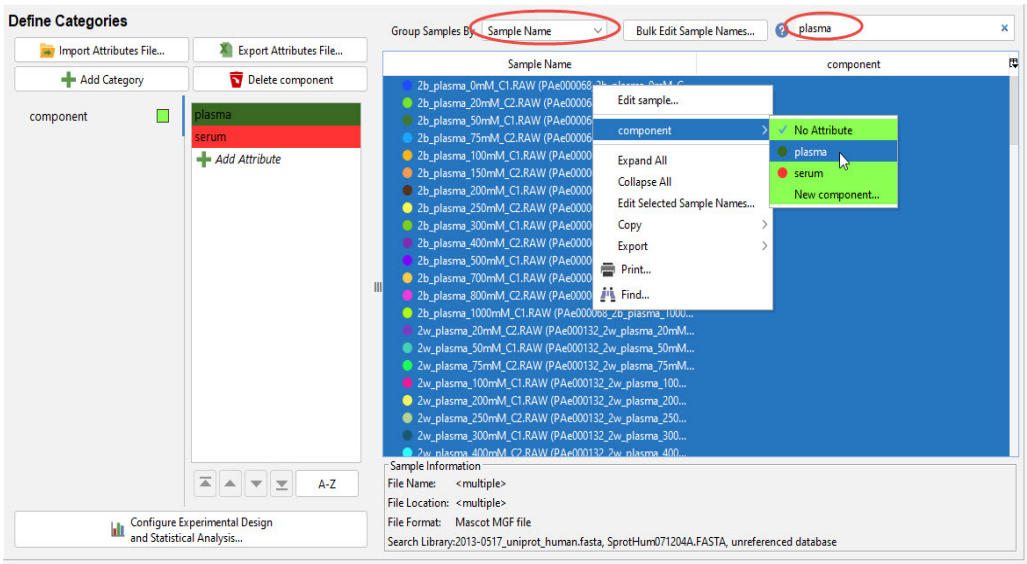


There are several methods by which Attributes may be assigned to samples.

- Drag and drop - Select the Category whose attributes are to be applied by clicking on it in the list under Define Categories. Select Group Samples By Selected Category Drag an individual attribute in the Sample Organization tree table to a sample name. Alternatively, select one or more samples and drag them to one of the Attributes, which appears in the table as the Attribute name and its color block.
- Right-click on Sample(s) - This option is often used in Display Samples by Sample Name mode, but also works in the other modes. Select one or more samples in the Samples column and right-click. Hover over an Attribute Group in the context menu that appears, then select an Attribute to assign to all selected samples.
- Search Box - A helpful method when organizing large experiments is to use the Search Box to display a subset of samples, select them all then either right-click and select an Attribute or drag the Attribute to the set of samples. Often the sample names contain substrings that indicate which attributes belong to which samples. In this case, the Search Box approach allows the user to leverage this information to quickly organize the samples.

-

Figure 5-3: Assigning an Attribute to multiple samples using the Search Box



Sample Organization tree table

When the Group Samples By option is set to Sample Name:

- The table shows the list of samples in alphabetical order in the first column. A colored dot next to the sample name indicates the color associated with that sample throughout the program. The sample name or color may be edited by right-clicking in the cell and selecting Edit sample...
- An additional column is displayed for each Category that has been created, and the cells in these columns show the Attribute associated with the specific sample in that Category. A colored dot indicates the color associated with that Attribute throughout the program.

When the Group Samples By option selected is Selected Category or Experimental Design:

- The table shows the list of Categories as collapsible folders:

  When the folders are collapsed a + sign appears to the left side of the folder. Clicking the + sign expands the folder showing the list of attributes in the group and the +sign becomes a -. Clicking the - sign collapses the folder.

- When the Attribute Group is expanded, the Attributes belonging to the group are listed with a colored dot assigned to each of them.

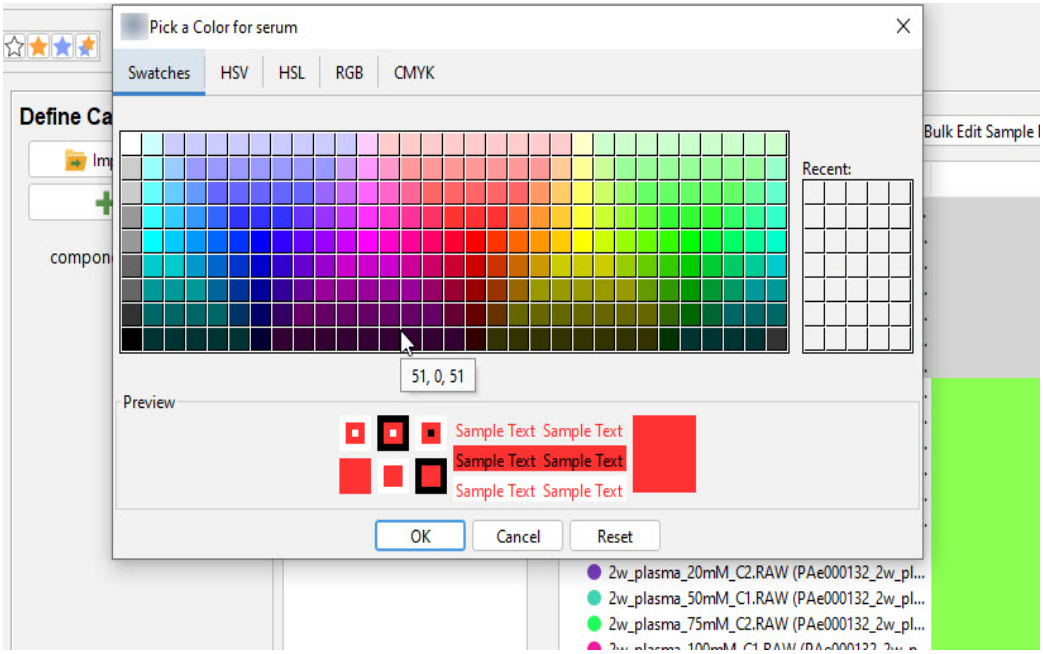
Right clicking on an Attribute Group folder or on an Attribute in the table:

- Right-clicking on an Attribute Group or Attribute in the tree table:

  This action displays a menu that allows the user to edit or delete the Attribute Group. The Edit Name option makes the attribute name editable and the Edit Color color option opens a dialog that allows the user to select a new color to be assigned to the Attribute.

-

Figure 5-4: Organize View - Edit Categories



- If Group Samples By Selected Category is selected, all of the samples are shown in folders grouped by the selected category. Each folder contains all samples with a specific Attribute of that Category. If Experimental Design is chosen as the grouping method, samples are organized into a hierarchical set of folders based on the Experimental Design (see Experimental Design).

Bulk Edit Sample Names

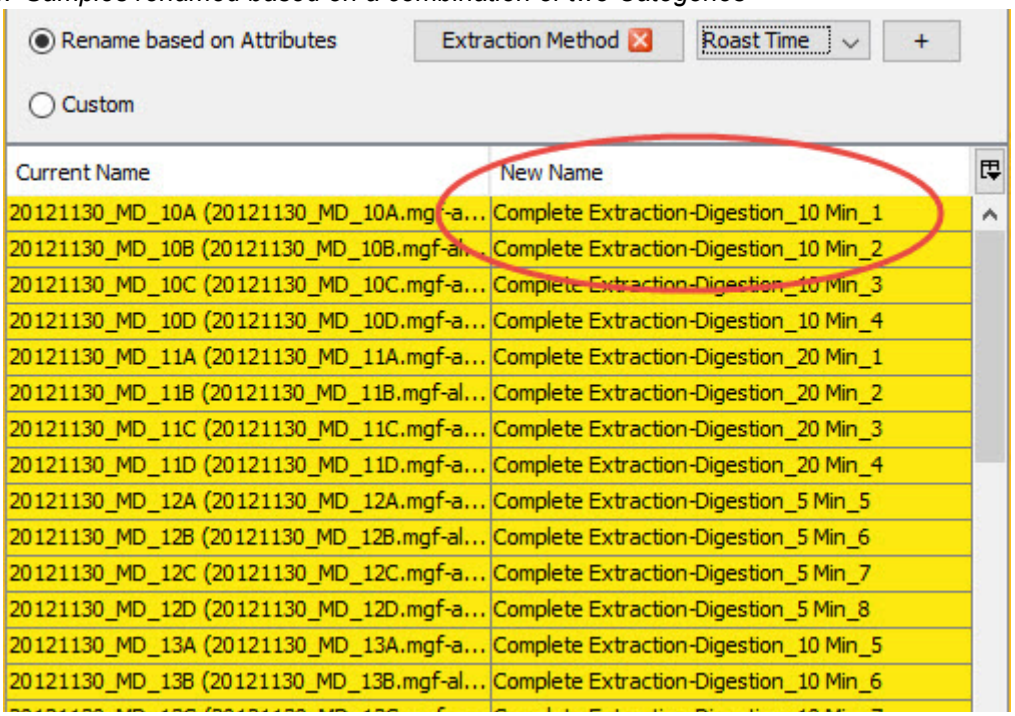
Often sample names are quite long and difficult to distinguish. This can cause problems when viewing the data in the other Views of Scaffold Elements. The user may wish to edit the sample names to make them shorter and/or more meaningful. Several tools are provided to assist in this effort. The names as they appear currently are shown in the left column of the table, and as they would appear if a proposed edit were applied on the right. Buttons at the bottom allow the user to Apply an edit, Cancel an edit or close the dialog (OK). The editing tools provided are:

- Remove prefix - if all sample names share a common prefix, it will appear in the text field. It may be edited if the user wishes to remove just a portion of the common prefix.
- Remove suffix - if all sample names end in a common suffix, it will appear in the text field. It may be edited to allow removal of a portion of the common suffix.
- Remove characters at beginning - the user may select a specific number of characters to be removed from the beginning of all sample names.
- Remove characters at end - the user may select a specific number of characters to be removed from the end of all sample names.

-

- Rename based on Attributes - the user may select a Category from the dropdown. Samples will be renamed to the Attribute name associated with that sample for that Category appended with a sequential number. Clicking the + button will add a second Attribute from another Category (see Figure 5-5).

Figure 5-5: Samples renamed based on a combination of two Categories



- Custom - allows selection and editing of individual names in the New Name column.

Import Attributes File...

A quick way to apply a number of Attributes to MS samples already loaded into Scaffold Elements is to read them from a formatted list of Attributes saved as a tab delimited text file, see Compiling an Attributes File. The file can be imported through the I**mport Attributes file...** button or by selecting the **Experiment > Load Attributes from File** command from the main menu.

If some Attributes have already been applied, importing an Attributes File will update assignments of existing Attributes listed in the file, but will not affect any Attributes that are not included in the file. Thus an Attributes File may be used to add to existing Attribute assignments, or to update them.

Compiling an Attributes File

The first line, or header line, in a Scaffold Elements attributes file begins with **Sample Name** followed by a list of Attribute Group names. If they do not already exist, these Categories will be created when the file is imported.

Each successive line must begin with the name of a sample loaded into the program followed by Attributes, each belonging to the Attribute Group listed above it in the header line. Note that the sample names must be precisely the same as the sample names loaded into Scaffold Elements.

One method of creating an Attributes File is to export a skeletal file containing the sample list from the

-

program and then add the Attribute information for each sample using Excel or a similar program. The list of loaded samples can be compiled by clicking the **Export Attributes File...** button or by selecting the **Export > Export Attributes File...** command from the main menu. Once the exported file is opened in Excel, it is easy to add attribute information to each sample in the list. The top row, or header line, will begin with SAMPLE NAME, and the names of the desired Categories, should be added. The list of samples must be the first column in the file, and the Attributes should be added in the subsequent columns (see Figure 5-6below).

Figure 5-6: Example of a Scaffold Elements Attributes text file

	Sample Name	Biosample	Category	Ethnicity	Anticoagulant	HPLC
2	CAK-20-400-900.RAW (F002703)	PAe000817	Lab-1	b1	edta	20-400-900
3	CAS-40-900-1200.RAW (F002698)	PAe000810	Lab-1	b1	serum	40-900-1200
4	CAH-40-900-1200.RAW (F002760)	PAe000862	Lab-1	b1	heparin	40-900-1200
5	CAH-20-900-1200.RAW (F002757)	PAe000862	Lab-1	b1	heparin	20-900-1200
6	CAC-10-400-900.RAW (F002743)	PAe000859	Lab-1	b1	citrate	10-400-900
7	CAC-40-900-1200.RAW (F002751)	PAe000859	Lab-1	b1	citrate	40-900-1200
8	CAH-20-400-900.RAW (F002756)	PAe000862	Lab-1	b1	heparin	20-400-900
9	CAS-20-1200-200.RAW (F002696)	PAe000810	Lab-1	b1	serum	20-1200-200
10	CAK-40-400-900.RAW (F002706)	PAe000817	Lab-1	b1	edta	40-400-900
11	AAS-10-400-900.RAW (F002734)	PAe000797	Lab-1	b3	serum	10-400-900
12	CAS-40-400-900.RAW (F002697)	PAe000810	Lab-1	b1	serum	40-400-900
13	CAC-40-1200-2000.RAW (F002752)	PAe000859	Lab-1	b1	citrate	40-1200-2000
14	CAC-20-400-900.RAW (F002747)	PAe000859	Lab-1	b1	citrate	20-400-900
15	CAS-10-900-1200.RAW (F002692)	PAe000810	Lab-1	b1	serum	10-900-1200
16	CAK-40-1200-2000.RAW (F002708)	PAe000817	Lab-1	b1	edta	40-1200-2000
17	CAH-40-400-900.RAW (F002759)	PAe000862	Lab-1	b1	heparin	40-400-900
18	CAH-10-400-900.RAW (F002753)	PAe000862	Lab-1	b1	heparin	10-400-900
19	CAH-10-1200-2000.RAW (F002755)	PAe000862	Lab-1	b1	heparin	10-1200-2000
20	CAK-20-1200-200.RAW (F002705)	PAe000817	Lab-1	b1	edta	20-1200-200
21	CAC-20-1200-200.RAW (F002749)	PAe000859	Lab-1	b1	citrate	20-1200-200
22	AAS-40-400-900.RAW (F002740)	PAe000797	Lab-1	b3	serum	40-400-900
23	CAK-10-1200-2000.RAW (F002702)	PAe000817	Lab-1	b1	edta	10-1200-2000
24	CAH-40-1200-2000.RAW (F002761)	PAe000862	Lab-1	b1	heparin	40-1200-2000
25	AAS-20-400-900.RAW (F002737)	PAe000797	Lab-1	b3	serum	20-400-900
26	AAS-10-900-1200.RAW (F002735)	PAe000797	Lab-1	b3	serum	10-900-1200
27	CAS-20-400-900.RAW (F002694)	PAe000810	Lab-1	b1	serum	20-400-900
28	AAS-20-1200-200.RAW (F002739)	PAe000797	Lab-1	b3	serum	20-1200-200
29	CAS-40-1200-2000.RAW (F002699)	PAe000810	Lab-1	b1	serum	40-1200-2000
30	CAH-10-900-1200.RAW (F002754)	PAe000862	Lab-1	b1	heparin	10-900-1200
31	CAS-20-900-1200.RAW (F002695)	PAe000810	Lab-1	b1	serum	20-900-1200
32	CAC-10-400-900.RAW (F002744)	PAe000859	Lab-1	b1	citrate	10-400-900
33	AAS-40-900-1200.RAW (F002741)	PAe000797	Lab-1	b3	serum	40-900-1200
34	CAS-10-400-900.RAW (F002691)	PAe000810	Lab-1	b1	serum	10-400-900

If the Scaffold Elements file already contains attribute data, such as Category and Biosample, these Categories do not need to be added again.

- *Important: Open the file in Excel, add the Attributes and then export from Excel as comma- or tab-delimited text file.*

Experimental Design

Supported experimental designs

Scaffold Elements supports label-free quantification and statistical analysis for three types of experimental designs

Basic Design

Experiments of this design consist of two or more biological classes of MS samples, between which variation is considered to be caused by experimental conditions (e.g. control and treatment classes). Each biological class is made up of one or more biological replicates which represent identical experimental conditions, but differ from one another because of biological variation (e.g. multiple organisms raised under identical experimental conditions would be biological replicates of one another). A biological replicate, in turn, consists of one or more technical replicates which originate from the same biological source. Finally, a technical replicate may be fractionated, in which case it is a set of MS samples which correspond to different portions of a sample which contain (ideally, but not exactly) distinct subsets of the total content of the originating sample. A non-fractionated technical replicate is simply a single MS sample.

Figure 5-7: Example experimental design

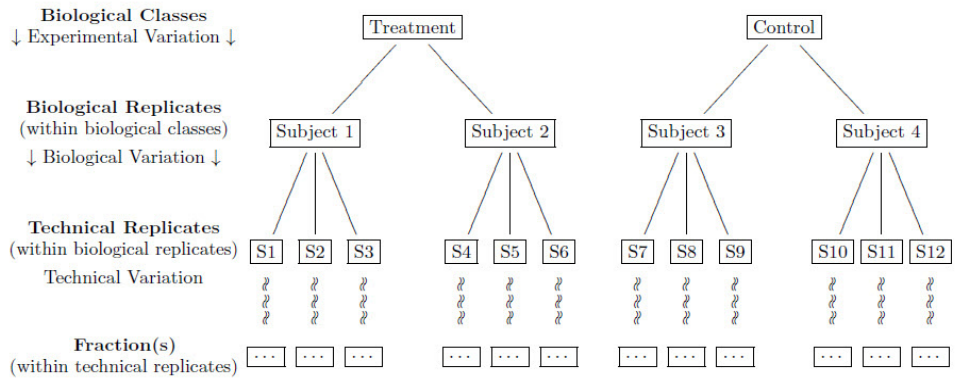


Figure 5-7 shows an example experimental design which can be easily represented in Scaffold Elements with two categories in a simple hierarchy, for example:

Group -> {Treatment, Control}

Subject-> {Subject 1, Subject 2,Subject 3,Subject 4}

can be summarized with the following hierarchy:

Group

Subject

-

(MS sample)

Repeated Measures

In a repeated measures experiment, samples are obtained at different times or under different conditions from the same biological entities. The goal is to analyze how each individual's levels change in response to the varying conditions. Each individual may have its own baseline value, but the goal is to analyze whether there are patterns in the changes from these baseline levels in response to the changing conditions.

Some examples of repeated measures studies are time-course studies and crossover studies. Time-course studies are used, for example, for measuring the response of individuals to a drug treatment. Initial baseline levels are measured, then measurements are taken at a series of time points to ascertain the pattern of response to the drug. In crossover studies, a set of individuals are exposed to a series of different treatments, with each individual receiving each treatment, although not necessarily in the same order.

The Samples Hierarchy in experiments of this type may contain biological replicates, technical replicates and/or fractions, as described in Basic Designabove.

Scaffold Elements requires that the experiment is complete, meaning that a sample is provided for each individual at each time point or in each condition.

Two-way Design

A two-way analysis compares the effects of two independent variables. It can help in determining whether there is an interaction between these two variables. For example, a study might compare the reaction of males and females to the administration of a drug. A two-way design would allow the researcher to test whether males and females respond differently.

When a two-way ANOVA test is applied to an experiment, three different measures are produced, each of which tests a different hypothesis. One measure assesses the degree to which the two factors interact, the second measures the effect of the category designated as the primary factor, and the third measures the effect of the category designated as the secondary factor. In Scaffold Elements, the user may select which of these measures should be displayed in the test result column.

The Samples Hierarchy in experiments of this type may contain biological replicates, technical replicates and/or fractions, as described in Basic Designabove.

Scaffold Elements currently requires that a two-way study be balanced, meaning that each combination of factors is represented by the same number of samples.

Randomized Block Design

Randomized Block is a specific type of Two-way analysis in which samples are divided into groups called blocks. Blocking compensates for situations in which known factors (e.g. age, sex) other than treatment group status are likely to affect what is being observed in the study[1]. The Randomized Block ANOVA measures the treatment effect while minimizing the effect of the blocking category; unlike the two-way ANOVA, it does not provide any assessment of the effect of the blocking category. Scaffold Elements only supports complete randomized block designs, i.e. those which contain one value per cell in the Design

1. https://www.statsdirect.com/help/analysis_of_variance/randomized_blocks.htm

Matrix.

Specifying the Design of an Experiment

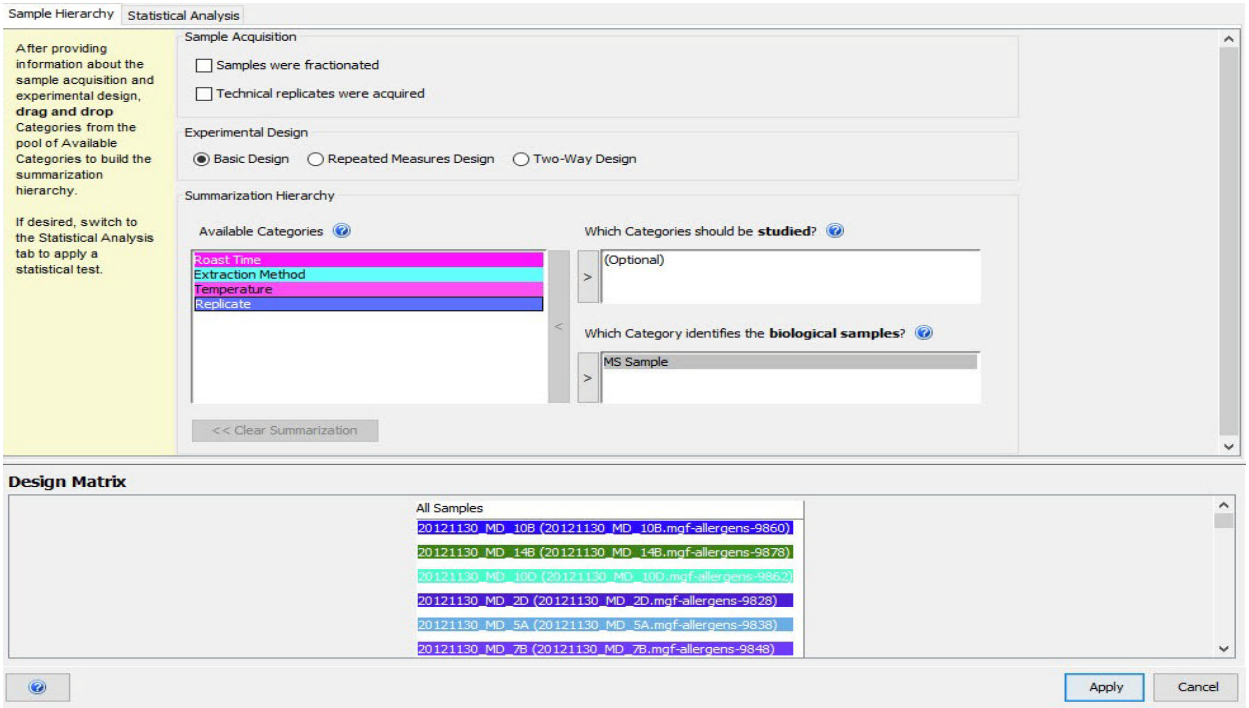
After samples have been organized into Categories (see Specifying the Design of an Experiment), it is important to organize the Categories to reflect the design of the experiment. This is accomplished through the Configure Sample Organization and Statistical Analysis dialog.

The Configure Sample Organization and Statistical Analysis Dialog

This dialog may be opened by:

- Clicking the Configure Experimental Design and Statistical Analysis... button at the bottom of the Organize View
- Selecting the menu item Experiment>Quantitative Analysis...
- Clicking on the Quantitative Analysis icon in the toolbar
- Selecting Edit from the list in the Summarization dropdown

Figure 5-8: Configure Sample Organization and Statistical Analysis dialog, *initial state*



The Configure Sample Organization and Statistical Analysis dialog includes two tabs:

- Sample Hierarchy Tab -- allows the user to specify the type of experiment to be analyzed and the roles of the various Categories in the analysis.

- Statistical Analysis Tab -- presents the various statistical tests available for analyzing the experiment as it has been specified in the Sample Hierarchy tab, and allows the user to select the test and specify its parameters.

Sample Hierarchy Tab

The upper portion of the Sample Hierarchy Tab consists of three sections:

- Sample Acquisition - this portion consists of two checkboxes:
  - Samples were fractionated - should be checked if the samples were fractionated,e.g. if positive and negative modes for each sample are in separate samples and should be combined during analyte identification.
  - Technical replicates were acquired - should be checked if technical replicate samples were gathered, e.g. if biological samples were aliquoted and the aliquots were analyzed as separate MS Samples. If samples are designated as technical replicates, their analyteanalyte-level quantitative values are first normalized and then summed to give a total value for the biological sample.
- Experimental Design - this section provides options for defining the basic structure of the experiment.
  - Basic Design - this option should be selected if the user simply wishes to view the MS results without performing any statistical analysis, or if a simple comparison based on a single Category is to be carried out (see Basic Design). This allows performance of, e.g., a T-Test or ANOVA.
  - Repeated Measures - this option should be selected if samples from each biological subject have been analyzed under different conditions or at different time points (see Repeated Measures).
  - Two-way Design - this option should be selected if the data is to be analyzed on the basis of two Categories (see Two-way Design). For example, a study might assess the differential effect of a treatment on males and females.
- Summarization Hierarchy - this section allows the user to specify how quantitative values should be summarized based on the Categories. The Available Categories are shown on the left. On the right are a series of boxes used to assign the Categories to the different analysis levels required by the experimental design. Different boxes are displayed depending on the experimental design type and whether or not there are technical replicates and fractionation.

  The Summarization Hierarchy determines how the data may be viewed in the Samples View as well as how it will be analyzed in statistical tests. Once Categories have been assigned to different levels in the Summarization Hierarchy, quantitative values may be “rolled up” or summarized to any of the levels for display and analysis.

Figure 5-9: The Samples View with a Summarization Hierarchy defined

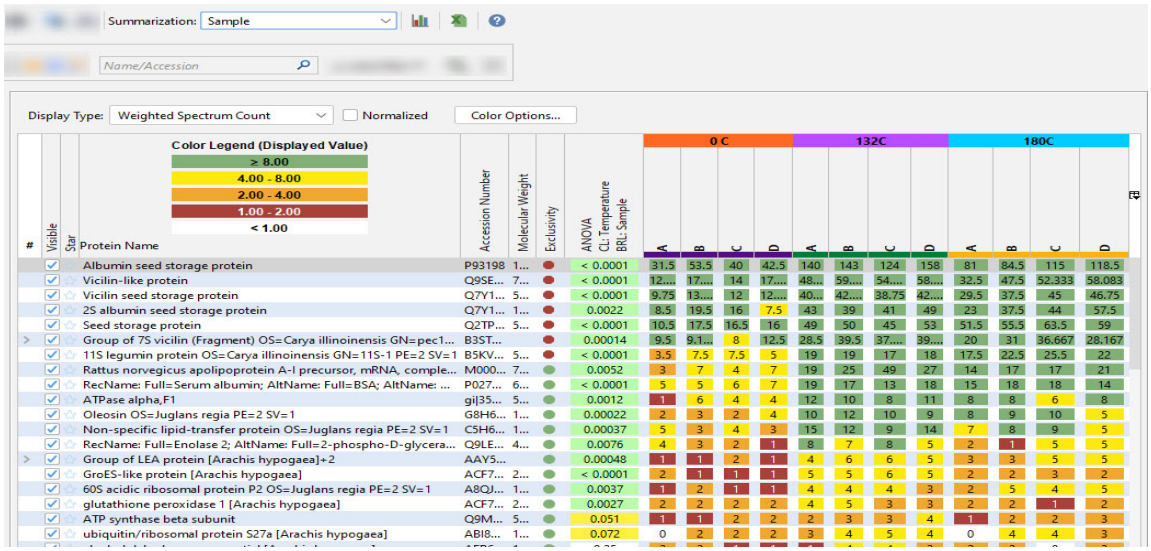
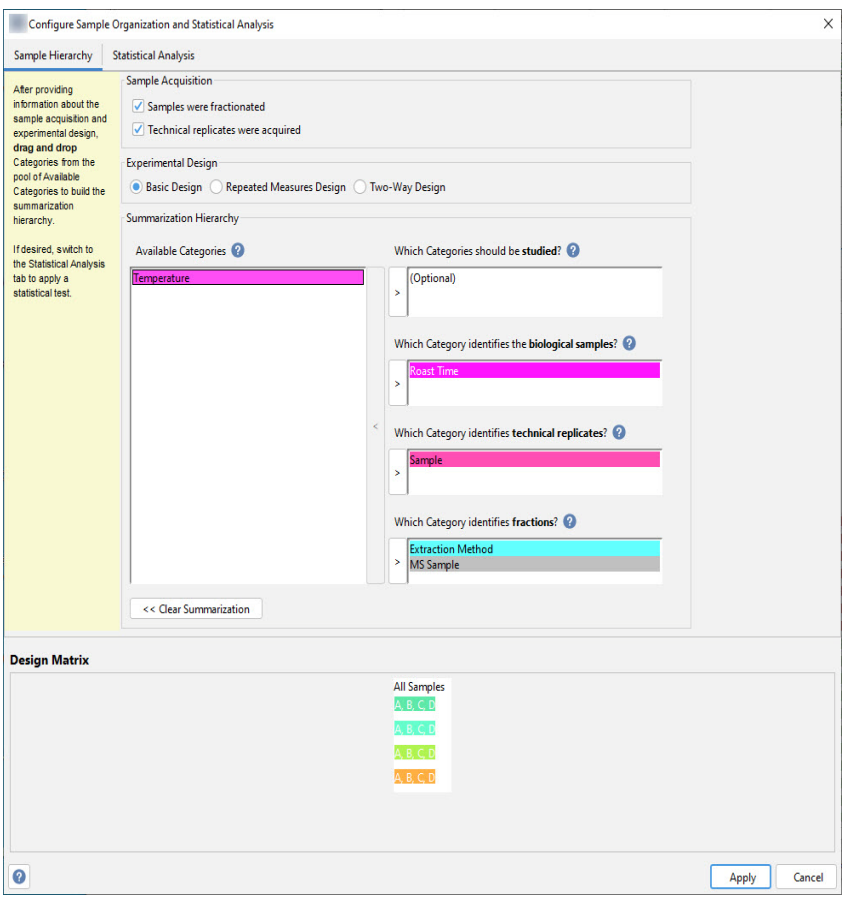


Figure 5-10: The Sample Hierarchy dialog when Basic Design is selected, showing fractions and technical replicates



Each Category may be moved from the Available Categories list on the left to an appropriate box on the

right either by dragging and dropping or by using the left and right arrows on the boxes. To add a Category to a box, select the Category in the Available Categories list, then click the right arrow on the box. To return a Category to the Available list, select the Category in a box and click the left arrow on the Available Categories box.

Design Matrix

As Categories are moved from the Available Categories List to their assigned roles in the experiment, A table is constructed in the lower pane of the Configure Sample Organization and Statistical Analysis dialog. Samples are placed into rows and columns to indicate how they will be grouped for evaluation in statistical testing.

The Design Matrix can help the user to visualize the experiment and verify that the experiment has been set up correctly. When viewed from the Statistical Analysis tab, the column and row headers also contain check boxes. Using these boxes, the user may select which rows and columns should be included in the test. Unchecking a box excludes that row or column from consideration when the test is applied.

When a statistical test has been applied, if the summarization level is set to the level representing Biological Samples, colored bars appear in the column headers in the Samples View to indicate the Comparison Groups used in the statistical test. In the Samples Report, the comparison groups are indicated by numbers in parentheses in the column headers.

For the Basic Design:

The user should specify:

- Which Categories Should be Studied: if no Categories are moved into this box, it will not be possible to apply a statistical test and no summarization above the level of the biological samples will occur. If one Category is selected for study, statistical comparisons between groups of samples corresponding to the different Attributes of that Category may be made, and values may be summarized to the Category level. If more than one Category is selected, statistical comparisons will operate on groups representing each possible combination of attributes for those Categories, and summarization may proceed up through the levels specified.

  For example, if both Extraction Method and Roast Time were selected for study, ANOVA would compare Complete Digestion for 0 min., Complete Digestion for 5 min., Soluble Digestion for 0 min, Soluble Digestion for 5 min. etc. Data could be viewed at the MS Sample level, the Replicate level, the Roast Time level or the Digestion Method level.

Figure 5-11: Selection of two Categories to be studied

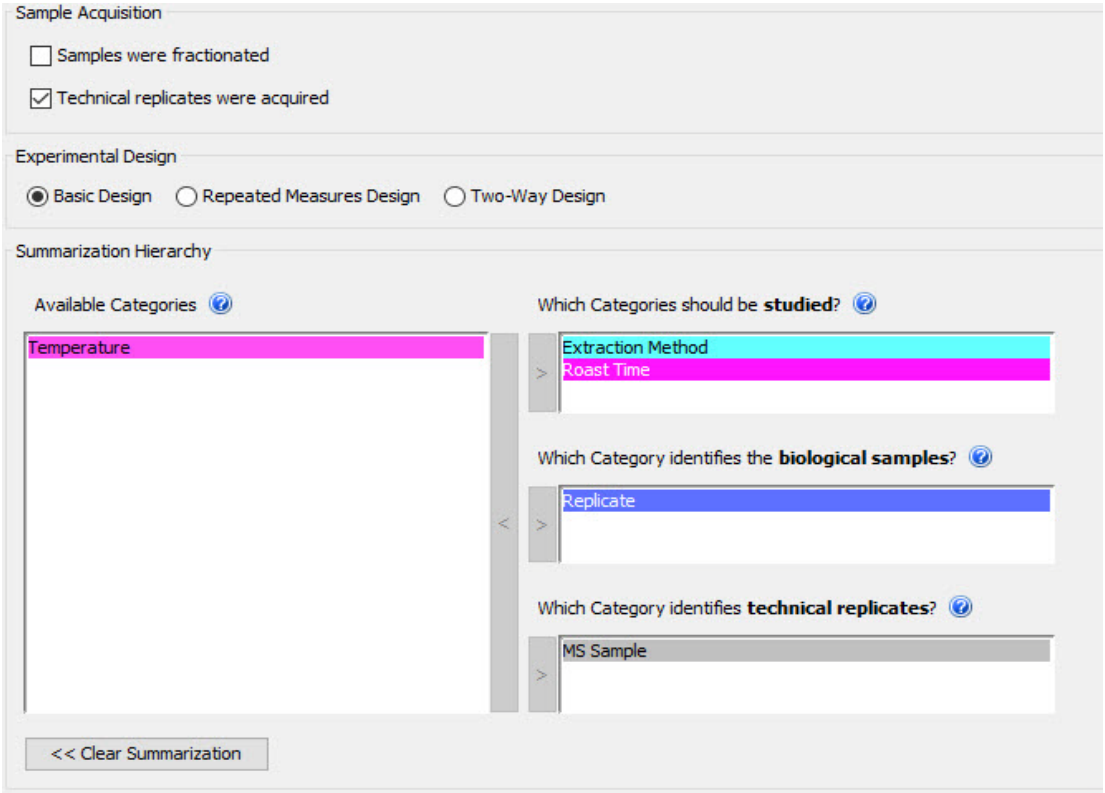
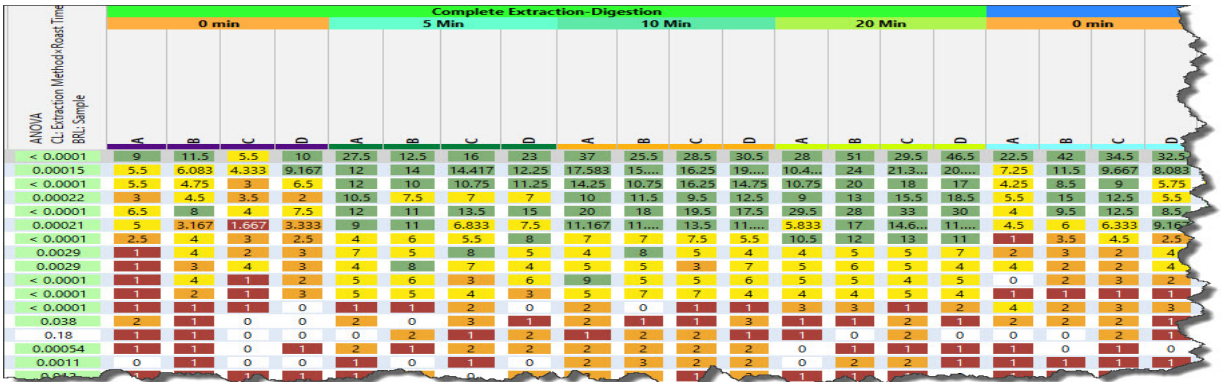


Figure 5-12: The resulting Samples View, shown at the Replicate level



ANOVA	CL:Extraction Method×Roast Time																			
	Complete Extraction-Digestion																			
	0 min				5 Min				10 Min				20 Min				0 min			
BRL:Sample	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
< 0.0001	9	11.5	5.5	10	27.5	12.5	16	23	37	25.5	28.5	30.5	28	51	29.5	46.5	22.5	42	34.5	32.5
0.00015	5.5	6.083	4.333	9.167	12	14	14.417	12.25	17.583	15....	16.25	19....	10.4...	24	21.3...	20....	7.25	11.5	9.667	8.083
< 0.0001	5.5	4.75	3	6.5	12	10	10.75	11.25	14.25	10.75	16.25	14.75	10.75	20	18	17	4.25	8.5	9	5.75
0.00022	3	4.5	3.5	2	10.5	7.5	7	7	10	11.5	9.5	12.5	9	13	15.5	18.5	5.5	15	12.5	5.5
< 0.0001	6.5	8	4	7.5	12	11	13.5	15	20	18	19.5	17.5	29.5	28	33	30	4	9.5	12.5	8.5
0.00021	5	3.167	1.667	3.333	9	11	6.833	7.5	11.167	11....	13.5	11....	5.833	17	14.6...	11....	4.5	6	6.333	9.16
< 0.0001	2.5	4	3	2.5	4	6	5.5	8	7	7	7.5	5.5	10.5	12	13	11	1	3.5	4.5	2.5
0.0029	1	4	2	3	7	5	8	5	4	8	5	4	4	5	5	7	2	3	2	4
0.0029	1	3	4	3	4	8	7	4	5	5	3	7	5	6	5	4	4	2	2	4
< 0.0001	1	4	1	2	5	6	3	6	9	5	5	6	5	5	4	5	0	2	3	2
< 0.0001	1	2	1	3	5	5	4	3	5	7	7	4	4	4	5	4	1	1	1	1
< 0.0001	1	1	1	0	1	1	2	0	2	0	1	1	3	3	1	2	4	2	3	3
0.038	2	1	0	0	2	0	3	1	2	1	1	3	1	1	2	1	2	2	2	1
0.18	1	1	0	0	0	2	1	2	1	2	2	1	1	0	2	0	0	0	2	1
0.00054	1	1	0	1	2	1	2	2	2	2	2	2	0	1	1	1	1	0	1	0
0.0011	0	1	0	0	1	0	1	0	2	3	2	2	0	2	2	1	1	1	1	1

- Which Category identifies the biological samples: the Category that identifies the biological subject should be placed into this box. This Category will be used as the blocking level in statistical analysis. This means that the values as they appear when rolled up to this level are used in statistical test calculations. In the Design Matrix, the biological subject defines the cells or blocks. More than one Category may be entered into this box, in which case each combination of the values in these Categories will define a biological sample.
- Which Category identifies technical replicates: if each biological sample has been divided and the resulting sub-samples have been analyzed in the mass spectrometer separately, the sub-samples are

technical replicates. The Category which names these sub-samples should be entered here. In some cases, the sub-samples may have undergone different treatments or have been obtained separately from the same biological subject, but in Scaffold Elements, they should be considered as technical replicates if the intent is to use them as multiple measurements of the same biological subject.

- Which Category identifies fractions: this field appears only if the fractionation check box is checked. This indicates that the MS Samples should be combined and treated as if they were a single MS Sample.

Clicking Apply finalizes the creation of the summarization level pull down list appearing in the summarization pane in the Scaffold Elements main window.

For the Repeated Measures Design:

- The first Category to be specified is the Time or Repeated Measure Category. The Time Category defines which of the repeat groups a sample represents. For instance, the Time Category could be Time Point, with values of 0 hr, 1hr, 2 hr and 3 hr. It need not represent time, however. For example, in a study that measures each subject's reaction to various treatments, it might be Treatment.
- Other Categories to be specified are similar to those described in "For the Basic Design:".

For the Two-Way Design:

Two categories should be selected for study. One will be designated as the Primary Analysis Category, while the other will be termed the Secondary Analyis Category. Even though one is designated as primary, the user may choose which category to assess with a Two-Way ANOVA test without changing the sample hierarchy. As a result, the user should specify:

- Which Category should be considered as the Primary Analysis Category. This is generally the treatment or condition that is the main focus of the experiment.
- Which Category identifies the Secondary Analysis Category. Often this will be a category that defines a condition which is to be controlled for in the experiment.

Other Categories to be specified are similar to those described in "For the Basic Design:"

Summarization Level

The level of summarization at which the data is to be grouped can then be selected from the summarization drop-down list. The user can choose the level from the most detailed (MS sample) to any higher level shown in the summarization list. Selecting 'Biosample' in this example and then looking at the Samples View, causes the data to be rolled up as shown in Figure 5-13.

Figure 5-13: Data grouped by Extraction Method>Roast Time>Replicate, summarized at the Roast Time attribute level

ANOVA CL: Extraction Method×Roast Time BRL: Sample	Complete Extraction-Digestion				Soluble Protein Digestion			
	0 min	5 Min	10 Min	20 Min	0 min	5 Min	10 Min	20 Min
< 0.0001	36	79	121.5	155	131.5	258.5	181.5	168.5
0.00015	25.083	52.667	68.333	76.417	36.5	84.5	67.167	61.833
< 0.0001	19.75	44	56	65.75	27.5	63.5	49.5	43.5
0.00022	13	32	43.5	56	38.5	85.5	53.5	63.5
< 0.0001	26	51.5	75	120.5	34.5	71.5	54.5	53.5
0.00021	13.167	34.333	47.667	48.833	26	58	35.333	36.667
< 0.0001	12	23.5	27	46.5	11.5	22.5	21.5	19.5
0.0029	10	25	21	21	11	35	27	60
0.0029	11	23	20	20	12	30	20	19
< 0.0001	8	20	25	19	7	3	4	0
< 0.0001	7	17	23	17	4	7	6	3
< 0.0001	3	4	4	9	12	29	14	19
0.038	3	6	7	5	7	12	5	6
0.18	2	5	6	3	3	8	8	7
0.00054	3	7	8	3	2	5	3	4
0.0011	1	2	9	5	4	6	5	4
0.012	4	5	7	3	4	[illegible]	[illegible]	[illegible]

The Summarization Bar allows the user to combine samples at various levels of categorization. The drop-down list displays the Categories in the Summarization hierarchy. The Samples View table will display samples combined at the level of the selected Attribute Group, with values rolled up to that level appropriately. The last item in the list is the command Edit..., which, when selected, opens the Edit Experimental Design dialog.

Available Categories

This column initially lists all of the Categories that have been created in the Organize View. These categories may be assigned various roles in the experiment by moving them into the boxes to the right. To assign an available category to a specific role, click on the category and either:

- drag the category into the appropriate box at the right.
- click on the right arrow button in the appropriate box at right.

To return a category to the Available Categories list, select it in the box to which it has been assigned and either:

- drag it back to the Available Categories list.
- click on the left arrow in the Available Categories box.

Categories to be Studied

The category or categories that will define the groups to be compared in ratios or statistical tests. If no Categories are moved into this box, it will not be possible to apply a statistical test and no summarization above the level of the biological samples will occur. If one Category is selected for study, statistical comparisons between groups of samples corresponding to the different Attributes of that Category may be made, and values may be summarized to the Category level. If more than one Category is selected, statistical comparisons will operate on groups representing each possible combination of attributes for those Categories, and summarization may proceed up through the levels specified. For example, if both Condition and Sex are selected, statistical tests will compare Male Control, Female Control, Male Treated and Female Treated.

The Time Category

In a repeated measures experiment, the same subjects are measured at various time points or under different conditions. The Time Category defines which of the repeat groups a sample represents. For instance, the Time Category could be Time Point, with values of 0 hr, 1hr, 2 hr and 3hr. It need not represent time, however. For example, in a study that measures each subject's reaction to various treatments, it might be Treatment.

Primary Analysis Category

In a two-way experiment, the Primary Analysis Category should be the grouping that is the primary focus of the experiment. For example, in an experiment that compares analyte levels with and without drug treatment, but wants to consider the possibility of a differential response to the drug in males and females, the Primary Analysis Category would be set to Treatment, while the Secondary Analysis Category would be set to Sex. Note that in a Two-Way ANOVA, however, the user may select whether to assess the Primary Factor effect, the Secondary Factor effect, or the Interaction effect, so the choice of Primary vs. Secondary Analysis Category is not extremely important.

Secondary Analysis Category

In a two-way experiment, the Secondary Analysis Category is a second grouping that may have an effect on the outcome and should be considered along with the Primary Analysis Category in comparisons and statistical tests. Often it is a grouping that may represent a variable that should be controlled for in testing the Primary Factor effect. Note that in a Two-Way ANOVA, however, the user may select whether to assess the Primary Factor effect, the Secondary Factor effect or the Interaction effect, so in cases where there is not a clear Primary factor, the two categories to be studied may be presented in either order.

Biological Samples

This level indicates the Category that defines a biological sample or subject. There may be more than one sample representing one biological sample if technical replicates or fractions are collected.

Technical Replicates

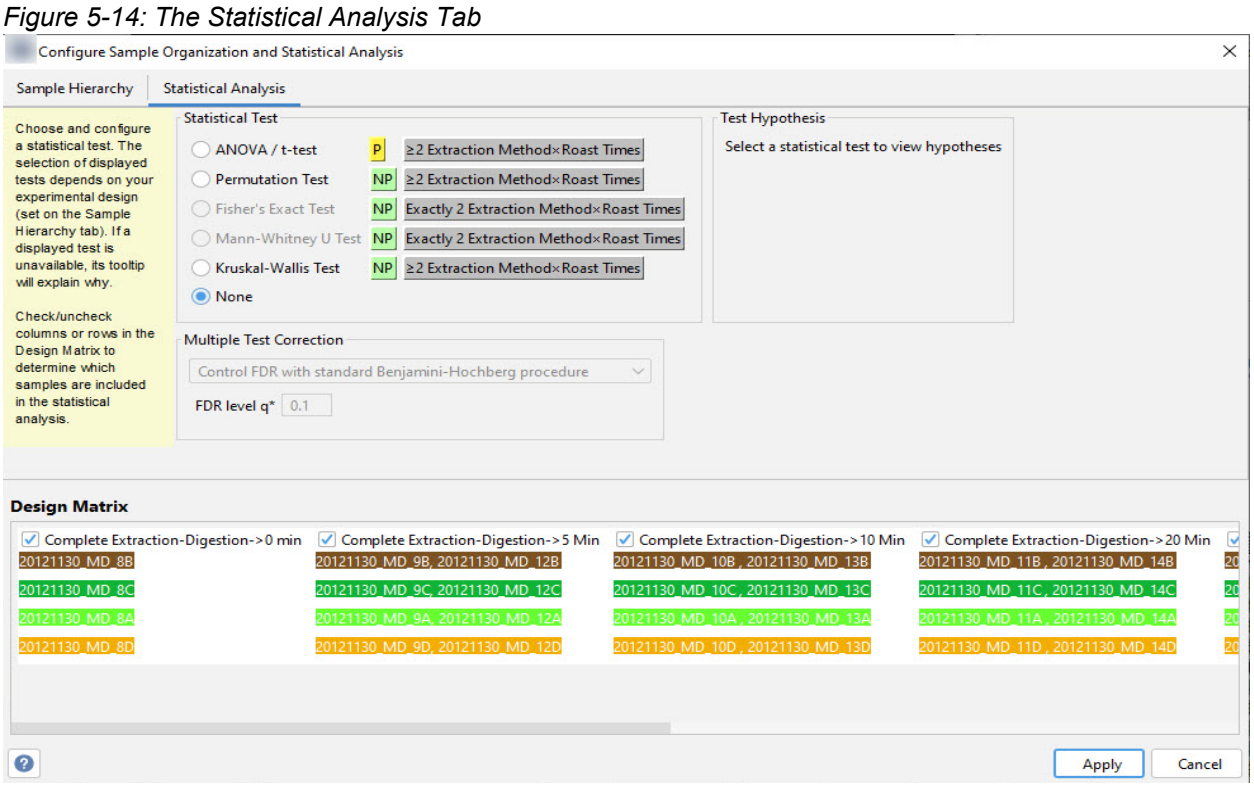
If the Technical replicates were acquired box is checked, the user must specify a category that represents the technical replicates. If each biological sample has been divided and the resulting sub-samples have been analyzed in the mass spectrometer separately, the sub-samples are technical replicates. The Category which names these sub-samples should be entered here. In some cases, the sub-samples may have undergone different treatments or have been obtained separately from the same biological subject, but in Scaffold Elements, they should be considered as technical replicates if the intent is to use them as multiple measurements of the same biological subject. Technical replicates may be the MS Samples if fractionation has not been performed.

Fractions

Fractions are generally the MS Samples in cases in which the same biological samples have been analyzed in different ways and the results should be considered as a single sample for analyte identification.

Statistical Analysis Tab

Figure 5-14: The Statistical Analysis Tab



The Statistical Analysis Tab consists of several sections:

Instructions -- Helpful text displayed in the box with a yellow background.

Statistical Test - The tests appropriate for the selected experimental design are shown. If a test may not be performed, the test name is grayed and its selection is disabled. An explanation is provided.

Unchecking some boxes in the Design Matrix may make additional tests available. For example, the Mann-Whitney U Test requires exactly two categories. If the data has three, the test is unavailable, but unchecking the box in the header of one category allows the user to compare the other two categories using this test.

Multiple Test Correction - When testing the significance of a large number of analyte inferences, it is advisable to apply a multiple test correction to the statistical test. The dropdown allows selection of a correction method.

The **FDR Level q*** allows the user to select a significance level.

Test Hypothesis - Displays the null hypothesis being tested and the alternative hypothesis, which would be accepted if the test result is significant. In the case of a two-way analysis, several hypotheses may be tested. A dropdown allows the user to select which of the possible hypotheses should be tested, and the test hypothesis text adjusts accordingly.

Hypothesis options for the Two-Way ANOVA

- Interaction Effect - measures whether the Primary and Secondary analysis categories are related. A significant result for the Interaction Effect for a analyte means that the Primary and Secondary categories are not independent, but rather that the combination of these factors has an effect on the level of the analyte.
- Primary Factor Effect - measures whether the Primary Analysis Category has a significant effect when controlling for the Secondary Analysis Category.
- Secondary Factor Effect - measures whether the Secondary Analysis Category has a significant effect when controlling for the Primary Analysis Category.

Hypothesis options for the Randomized Block Design

- Treatment Effect - measures whether the Primary Analysis Category has a significant effect when controlling for the Secondary or Blocking Category.
- Block Effect - measures whether the Secondary Analysis Category (which is the Blocking Level in a Randomized Block experiment) has a significant effect.

-

Chapter 6
The Samples View

The Samples View provides an overview of the experiment, allowing the user to visualize the data summarized and analyzed to reflect the design of the experiment that generated them.

The organizes all of the information about the analytes identified in the experiment to provide an overview of the experimental results, as well as analyte-level quantitative information.

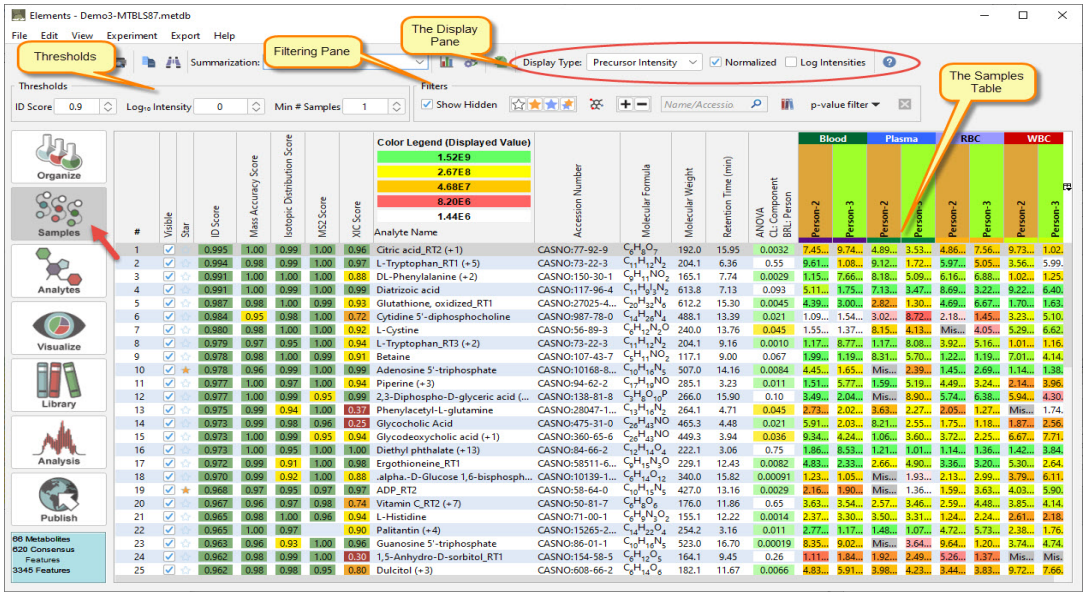
The Summarization Bar, found in the main Scaffold Elements window, allows the user to specify the level of summarization at which quantitative values should be reflected in the Samples View. The levels available depend on the sample organization established in and the experimental design set in the Configure Sample Organization and Statistical Analysis dialog (see Specifying the Design of an Experiment).

Samples View Table

In order to provide the user with an overview, the Samples View opens automatically in the Display pane of the Scaffold Elements main window when the application finishes loading data or when the user opens an Scaffold Elements *.METDB file. When more than one sample has been loaded, a pop-up dialog asks if the user wants to open the Organize View to organize the experiment. A checkbox allows the user to turn this prompt off permanently.The user may select a View other than the Samples View as the default upon loading or opening through Edit>Preferences>User Interface>Views.

The function of the Samples View Table is to show at a glance, for each identified analyte, the quantitative levels among the various Mass Spectrometry (MS) samples in the experiment. A variety of Display Type options are available.

Figure 6-1: The Scaffold Elements Samples View Table Tab



The main elements of the Table tab are:

- The Samples Table
- The Samples Table Display Bar

Samples Table

The Samples View presents an overview of the experiment. It provides the user with a list of detected analytes, labeled with hypothesized identifications when Scaffold Elements has been able to match the features to library entries. It also provides scoring information and other characteristics to facilitate assessment of the reliability of the identifications, and quantitative information indicating the level of expression of each analyte in each sample.

In addition, the Samples View provides tools such as filters, flexible summarization and statistical analysis that are designed to help answer the fundamental questions underlying the experiment. Through the Samples View, a researcher can identify the analytes that distinguish various classes of samples, such as treatment groups, tissue types, time since treatment, etc. Other Elements Views then provide additional resources for further examination of the preliminary results presented in the Samples View.

The level of data summarization is chosen from the Summarization Bar in the Scaffold Elements main window. The Display Type pull down menu determines the type of quantitative values reported in the table.

The analytes list can be filtered using the tools of the Filtering control bar.

General characteristics of the Samples Table:

- Samples Table Features
- Initial Sorting of Columns
- Summarization Level in the Samples Table
- Color Legend

Samples Table Features

Like any table in Scaffold Elements, the Samples Table makes use of the features and tools described in the Display pane section.

The following columns, initially ordered as follows, appear in the Samples Table:

- **# --** Order number of each row at the current ordering conditions.
- **Visible --** Shows a list of selected check boxes. Deselecting a box hides the corresponding row unless the “Show Hidden” check box in the Filters pane is selected. See Filters Pane.
- **Star** -- Initially shows a grayed out star for every row. Clicking the star activates it, tagging the analyte group. Clicking repeatedly causes the star to loop through its four possible states. The color of the star goes from gray to orange, to blue and to orange and blue then back to gray. For more information see “Tagging Analyte of Interest, the star function” on page 99.
- **Five Search Score columns** -- Report the maximum score value for the ID score and the four Match Scores. For more information about the scores, see “Elements Scoring

Algorithms" on page 213.

- (Optional) **RT Match --** appears only when a search has been conducted using the Search option: "**Match on mass and retention time**". This column consists of check boxes indicating whether or not the identification of the analyte relied on matching the retention time as well as the mass accuracy. If the box is not checked, no retention time information was available in the library, as an analyte identification that matches on mass but fails to match on retention time is rejected. Note that if even one ion matches on retention time, the box is checked.
- **Analyte Name** -- Name of the analyte group.
- **Accession Number** -- The analyte identification number. The particular type of ID shown in this column depends on the IDs available in the spectral library used for the search and on the options selected by the user in the **Preferences dialog > Accession Number tab**, see Accession Number"Accession Number" on page 52. Furthermore, the accession number is linked to the database selected through **Edit>Preferences>Web Links**. By default, clicking on the Accession Number opens a browser to the ChemSpider page for the selected analyte.
- **Molecular Formula** -- Provides the molecular formula of the selected analyte.
- **Molecular Weight** -- Provides the theoretical molecular weight.
- **Retention Time** -- provides the retention time of the selected analyte feature. Te units used to measure RT can be switched by going to **Edit > Preferences > Units**

The rest of the columns represent, depending on the selected level of summarization, either the MS samples or the Categories defined in the The Organize View and included in the hierarchical summarization.

The order of the columns can be changed by the user, see "Display pane" on page 67.

Initial thresholding

When the Samples View first opens, default thresholds are applied to the Tools for Limiting the Analyte List as follows:

- **ID Score:** 0.7
- **Log_{10} Intensity:** 0
- Reproducibility threshold or **Min # Samples:** 1

An analyte group must meet ALL specified thresholds in order to be included in the Analytes List. For more information about thresholds see "Elements Scoring Algorithms" on page 213.

Initial Sorting of Columns

When the Samples View first opens, the analytes are sorted by:

1. Decreasing ID score
2. Decreasing alphabetical order of the accession number

The analytes included in a group and the analyte groups included in a cluster are sorted as follows:

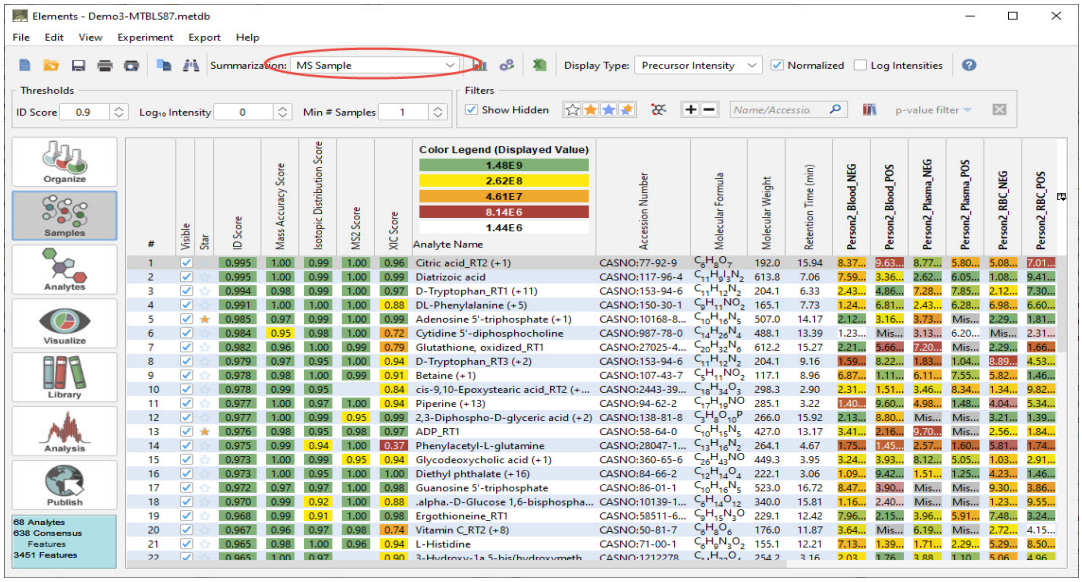
1. From the highest scoring analyte to the lowest across all samples
2. From the analyte that appears in most samples to the one that appears in the fewest samples
3. From the shortest to the longest analyte name

Each column is provided with a tri-state sorting feature. Clicking on any column header will reorder the table according to the values in that specific column. The first click sorts the column in ascending order, the second in descending order and the third returns the column to the original order.

Summarization Level in the Samples Table

After a new experiment is completely loaded, the Scaffold Elements Samples View appears with the Samples Table initially summarized at the lowest level of summarization, which is the MS sample levelFigure 6-1.

Figure 6-2: Scaffold Elements Samples View - The Samples Table initial summarization level



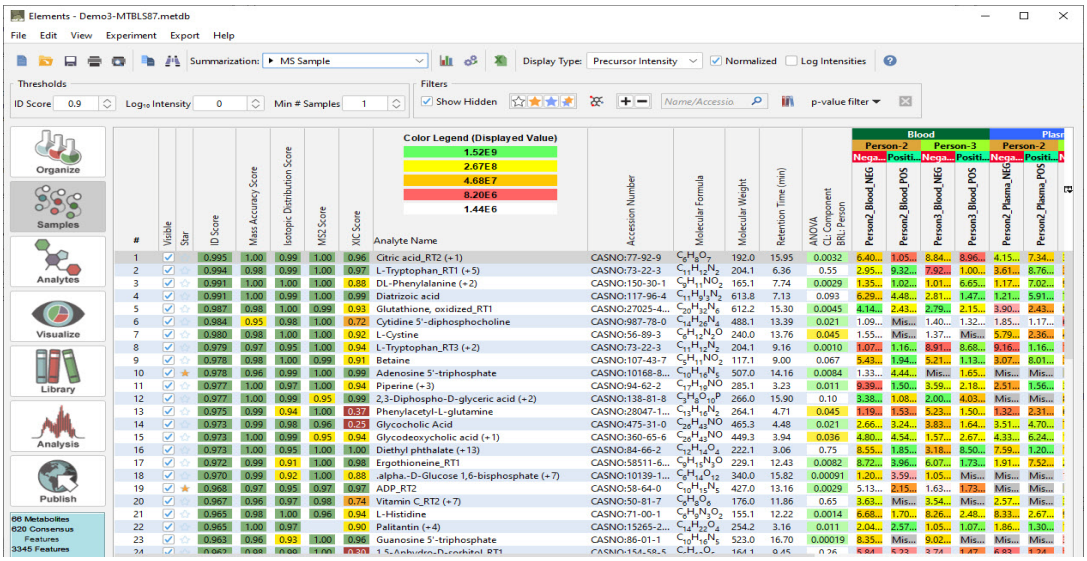
The user can add Categories and classify the MS samples within them through the Organize

View. It is then possible to set up the desired Summarization Hierarchy using any of the defined Attribute Groups by choosing **Edit...** in the Summarization Bar.

The selected Summarization Hierarchy and the choice of the summarization level will be reflected in the Samples Table column headers. In Figure 6-4 the selected level is still MS Sample but a more complex hierarchy has been set up in the Summarization Pane.

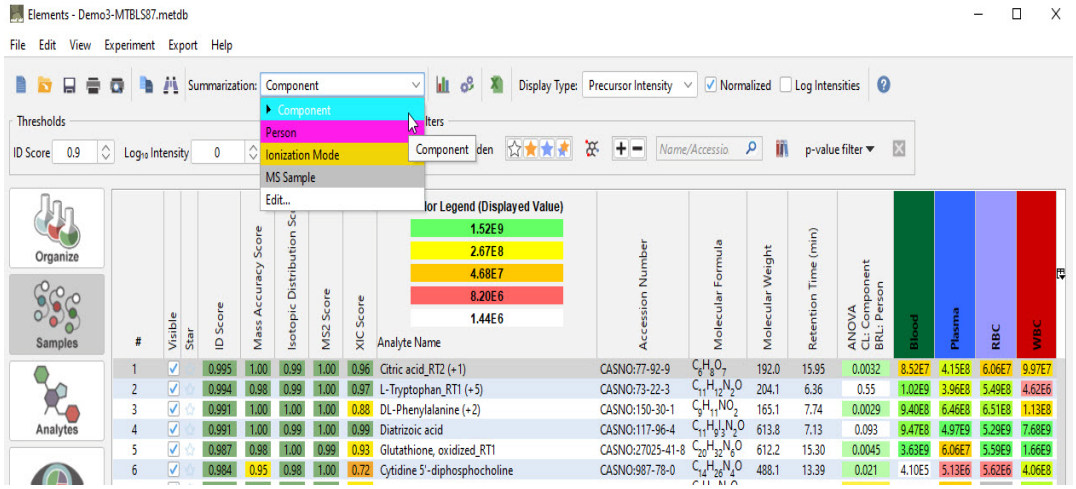
The headers of the samples columns show the Attributes for all Categories in the Summarization Hierarchy. They are colored according to the values assigned to the Attributes through the Organization View.

Figure 6-3: Samples View - The Samples Table with new levels of summarization added



Selecting a higher level of summarization will hide the lower ones.

Figure 6-4: Scaffold Elements Samples View - The Samples Table, highest level of summarization



Rolling up of quantitative values to higher summarization levels

The quantitative values shown in the Samples table depend on the selected Display Type and the chosen summarization level.Quantitative values are combined (or rolled up) to higher levels of summarization differently depending on the Display Type selected and the experimental design, see Rolling up of quantitative values.

Color Legend

Located at the top of the Samples Table in the analytes column header, the color legend defines the color coding associated with the selected Display Type. The color legend can be customized through the Edit Coloring for Display Type dialog opened by the Color Options button located in the Samples Table Display Bar.

Tools for Limiting the Analyte List

Analyte lists can be very long and extremely detailed and for that reason difficult to examine. The Analytes List in Scaffold Elements is organized into groups and clusters that can be expanded or collapsed for ease of inspection, thus reducing the number of independent rows in the list.

Scaffold Elements also includes tools that allow filtering the list to make it easier to view the analytes that are of particular interest. Thresholds help the user filter out less than optimal candidates, and various filters allow uninteresting analytes to be hidden.

Elements also reports analytes that have not been identified during the search, when the option **Report Unknown Analytes** is selected in the Search Tab of the Workflow dialog.

The following tools are available to adjust the display of the Analyte List:

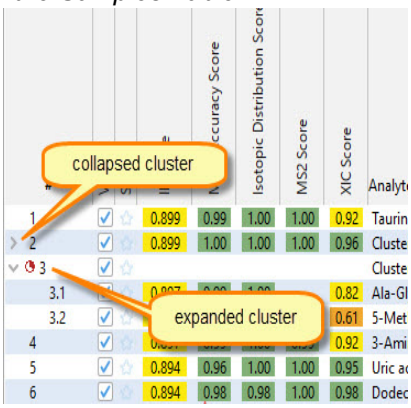
- “Initial thresholding” on page 94
- “Tagging Analyte of Interest, the star function” on page 99
- “Hidden Analytes” on page 100
- “Applying Confidence Thresholds to the Analyte List” on page 100
- “Applying filters to the Analyte List” on page 100
- “Rolling up of quantitative values to higher summarization levels” on page 97
- “Normalized check box” on page 104

- Tagging Analyte of Interest, the star function

Representation of Analyte groups and clusters

Scaffold Elements initially shows all analytes that pass the default thresholds as defined in section Initial thresholding. If the option to treat all features with the same retention time as a single analyte (see “Forming Consensus MS1 peak groups” on page 234), some of the analytes will be grouped into clusters. When a row includes a cluster of analytes, the row number is preceded by a small clickable expansion icon and a pie-icon.

Figure 6-5: Number column in the Samples Table



- Clicking the expansion icon expands or collapses the cluster.When a cluster is expanded the list of its analytes is visible. Each row in the list is numbered as a subset of the cluster number. For example, for a cluster appearing in the third row of the samples table and made up of two analytes, the analytes are numbered as 3.1 and 3.2. The numbers appear as a hierarchical structure in the # column.
- The pie icon graphically represents the degree of shared evidence of pairs of analytes in the cluster.

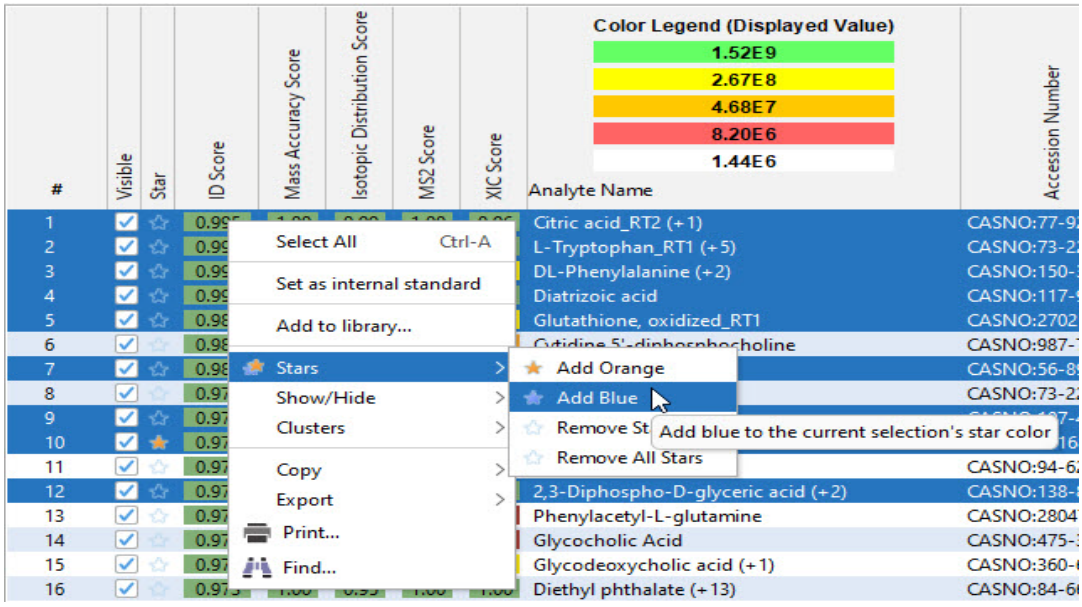
Tagging Analyte of Interest, the star function

The user can mark analytes in an experiment that are of special interest by clicking the Star icon ☆ in the **Star** column for the analyte. Three different colored stars, blue, orange and a combination of the two colors may be applied by clicking multiple times on the same star or by selecting the star option in the right click menu. By using a combination of different stars it is possible to create four different sets of analytes of interest. The user can then bring these items to the top of the display by clicking twice on the **Star** column header. To return to the default display order, click the column header twice more.

Sets of analytes can be starred at the same time by selecting multiple analytes and using the star option in the right click menu.

Star filters are included in the "Filtering control bar" of the Scaffold Elements Main Window. Clicking a specific colored star in the filters bar removes all analytes tagged with that star color from the table.

Figure 6-6: Samples Table- Starring analytes



Hidden Analytes

The user can easily remove analytes displayed in the Samples Table that are of no interest and/or are contaminants. An entire row in the Samples Table can be eliminated by simply clearing the Visible option in that row. This can be done for a single analyte by clicking the Visible check box, or for a group by using the right click menu.

The check box Show Hidden in the Filtering pane in the Scaffold Elements Main window toggles the display of hidden analytes

Applying Confidence Thresholds to the Analyte List

Through the Thresholds pane, see "Thresholds Pane" on page 63, the user can define a desired level of confidence in the identification of the analytes present in the list. A higher level of confidence means a shorter list of hypothetical analyte identifications in the Samples View Table and vice-versa.

Note that if in setting up the search the user has selected to include unidentified analytes and wants to see them in the analyte list, the unidentified compounds may be added to they Samples View Table through the "Show Unknown" option in the View menu even though they do not meet the score thresholds.

Applying filters to the Analyte List

Elements provides a number of different filters in the Filters Pane in the main Elements window. For a more detailed description of the options available, see Filters Pane"Filters

Pane” on page 58.

The filters in the Filters Pane can be combined. In order to be displayed, an analyte must meet minimum thresholds and must not match any filtered state.

Applying filters to the Analyte List

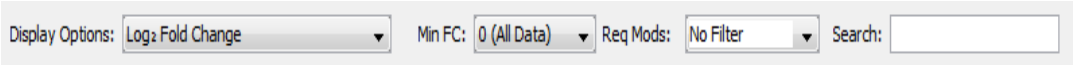
A number of filter options are offered in the Filtering control bar located in the Scaffold Elements Main Window. The filters affect the number of visible rows in the Samples Table.

- The Show Hidden check box toggles whether rows that have been tagged as not visible can be seen in the list.
- The Star Filter filters out all the rows that are not tagged as selected in the filter.
- The P-Value Filter filters out all analytes for which the statistical test result is not significant based on the selected criteria.

Samples Table Display Bar

Through the Samples Table Display Bar, the user can specify the type of values (for example, the Number of Assigned Spectra) that are displayed in the Samples Table for every analyte group. The bar also contains filtering options for limiting the display to only those analytes that meet specific criteria.

Figure 6-7: Scaffold Elements Display tab bar



The Samples Table Display bar contains the following features:

- Display Type
- Normalized check box and/or Ref: pull down list
- Log Intensities check box
- Color Options button
- Search Box

Display Type

The Display Type drop down list offers a range of sample statistics. Once a particular quantitative value is selected from the list all the related numbers are reported in the table cells for each analyte and MS sample or chosen level of summarization. If a display type is not available in the currently experiment, it will be shown in gray and will be disabled.

•

The coloring reflects the selected Display Type. For each type the coloring can be customized by selecting Color Options from the View Menu.

Log_2 Fold Change and Precursor Intensity missing value tags

When selecting a Display type that represents the log_2 Fold change of a quantity, or a rolled-up log_{10} intensity value, three different missing values tags might appear in the samples table's cells:

- **Missing Values** -- the log of a quantity that is zero, which ultimately refers to a analyte that has not been detected in a particular MS sample or group of samples belonging to the selected level of summarization.
- **No Values** -- the log of the ratio between two missing values.
- **Missing Ref.** -- the log of the ratio between a value and a missing reference value.

Rolling up of quantitative values

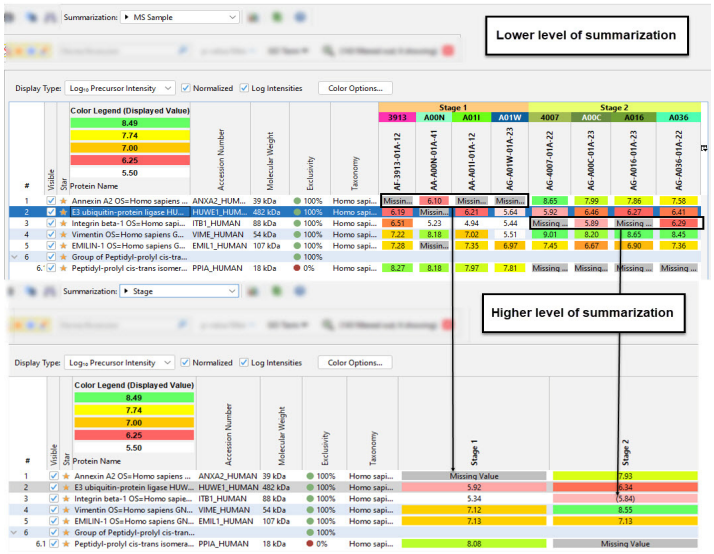
The quantitative values shown in the Samples table depend on the Display Type selected and the level of summarization chosen from the Summarization Bar. At the lowest level of summarization, the values shown relate to the amount of analyte present in each of the loaded MS samples in the Scaffold Elements experiment. Each MS sample corresponds to a column in the Samples table. Changing the level of summarization groups the MS samples according to the categorization created through the The Organize View and hierarchically ordered using the Experimental Design. When transitioning from one level to the next, the columns in the lowest level are subsumed into a new column representing quantification at a higher level of summarization. The methods by which values are rolled up from one level of summarization to the next depend on the selected Display Type.

Rolling up of Precursor intensity

Precursor intensity values appearing in the Samples table are rolled to the upper level of summarization by taking the median of the values in the corresponding lower level group. If more than 50% of the values are missing in the group that is to be rolled up, the QRILC algorithm is used to impute missing values (see "Rolling up Values" on page 220).

At the lowest level of summarization, missing values are highlighted using the Missing Value tag. At a higher level of summarization, when QRILC is used, the values are reported in parentheses. They are tagged as Missing Value when the lower level group of samples does not include any value at all, as shown in the example in Figure 6-8 .

Figure 6-8: Rolling up quantitative values: Log_{10} Precursor Intensity



Note: The **Log_2 Fold change (Precursor Intensity)** is calculated using the Log_{10} Precursor Intensity values with the Log_{10} value of the selected reference subtracted.The result of the subtraction is then transformed into the $Log_2$of the difference.

Normalized check box

When the **Normalized** check box is checked, the values shown in the Samples Table for the selected Display Type will appear normalized.

When the chosen Display Type is the Log_2 Fold Change (precursor intensity) and Normalization is selected, the display type Log_{10} Precursor Intensity will also appear normalized and vice-versa.

When the chosen Display Type is the Log_2 Fold Change (weighted spectrum count) and normalization is selected, the display type Weighted Spectrum Count will also appear normalized and vice-versa.

The normalization algorithms applied are described in section Quantitative Method.

Ref: pull down list

When Display Types containing a fold change are selected, the **Ref:** pull down list appears between the **Display Type** control and the **Normalized** check box. From this list, the user may select the Attribute or combination of factors to be used as the reference or denominator for the fold change calculation. The pull down list includes all of the Attribute Groups available in the summarization list, plus a list of all possible combinations of factor levels available at the selected summarization level.

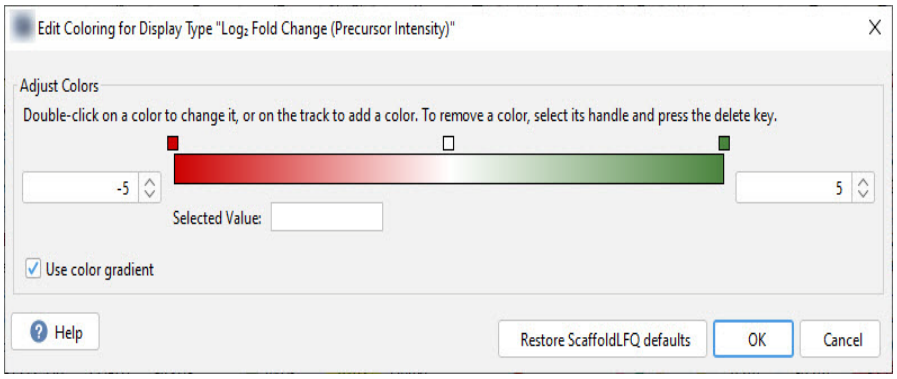
Log Intensities check box

When the Precursor Intensity display type is selected, a check box appears which allows the user to choose to display the values in Log_{10} format. When this is checked, base 10 logs of the appropriately summarized precursor intensity values are shown, and the name of the display type changes to Log_{10} Precursor Intensity.

Color Options button

Selecting the Color Options button opens the Edit Coloring for Display Type dialog. This dialog offers coloring adjustments tools for each Display Type.

Figure 6-9: Edit Coloring for Display Type dialog



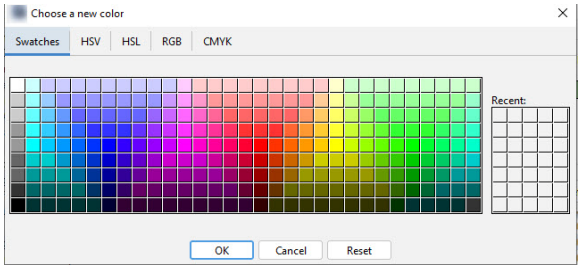
The Adjust Colors pane, included in the dialog, allows the user to create a custom color legend for the currently selected Display Type. Various features are available to set the intervals for a specific color as well as the definition of the overall range.

Sliding one of the colored squares located above the legend changes the range of the selected color. The color ranges can also be set by typing a value in the selected value box.

It is also possible to use a color gradient by clicking the color gradient check box.

Double clicking a color on the legend opens the **Choose a new color dialog** where the user can pick a different color to be added to the legend using either swatches, HBS or RGB methods. Double clicking a specific colored square also opens this dialog and allows the user to change the color of the selected square.

Figure 6-10: Choose a new color dialog



At the bottom of the Adjust Color pane, the **Restore Elements Defaults** button resets the legend to the default Display Type colors.

Search Box

This box allows the user to search the analyte list using the analyte description or the analyte accession number. To allow the use of regular expressions, the user must make the option available by selecting the box **Use regular expressions in search fields** in the Preferences, General tab.

Chapter 7
The Analytes View

Scaffold Elements' Analytes View includes a number of tools that can be used to conduct a more detailed examination of the consensus features that lead to a specific putative analyte identification listed as a row in the Samples table.

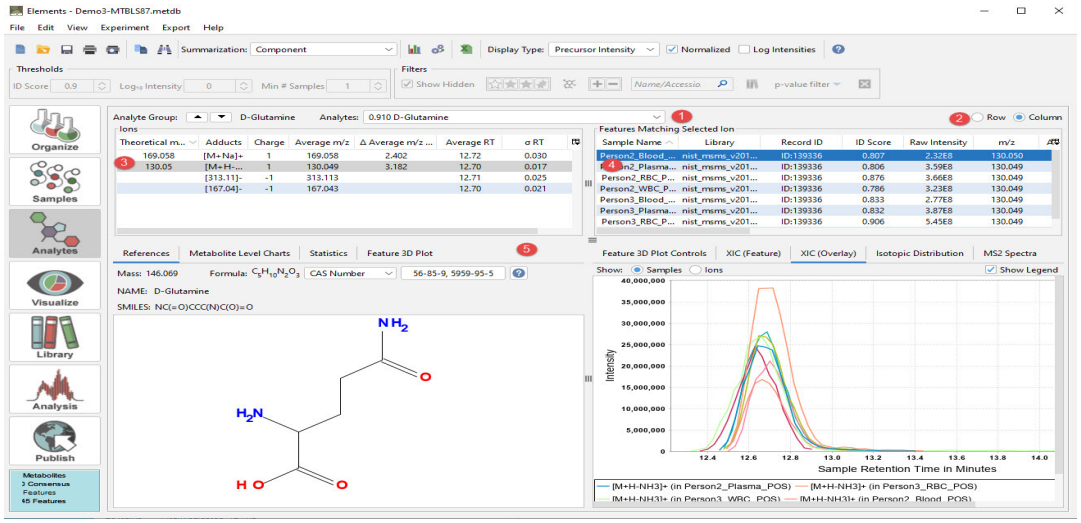
The view consists of five major sections, see Figure 7-1:

1. "Analytes List bar" on page 108.
2. "Layout" on page 108.
3. "Ions pane" on page 109.
4. "Features Matching Selected Ion pane" on page 109.
5. "Visualization pane" on page 110.

The three panes can be expanded or contracted by clicking on the boundary between panes and pulling it in the desired direction.

Note that all graphs and tables contained in the View include the features and tools described in "Display pane" on page 67.

Figure 7-1: Scaffold Elements Analytes View: 1. Analytes List bar, 2. Layout, 3. Ions pane, 4. Feature Matching Selected Ion pane, 5. Visualization pane.



-

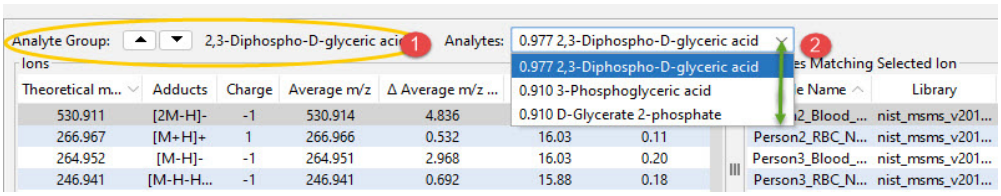
Analytes List bar

Positioned at the top of the Analytes View, this bar contains two lists, see Figure 7-2:

1. Analyte Group: list
2. Analytes: pull down list

Through these lists it is possible to explore all of the analyte groups appearing in the Samples table without switching views.

Figure 7-2: Analytes List bar



Analyte Group: list

This list has two functional arrows and a text box that shows the currently selected analyte group. Clicking the up or down arrow scrolls through the Analytes List from the Samples View. The other panes in the Analytes View are updated accordingly, as is the selected analyte group in the Samples View.

Analytes: pull down list

When the selected analyte group contains more than one analyte, the **Analytes:** pull down list shows the list of compounds which are members of the group. The user may select any of these analytes and explore its supporting evidence in the other panes of the view.

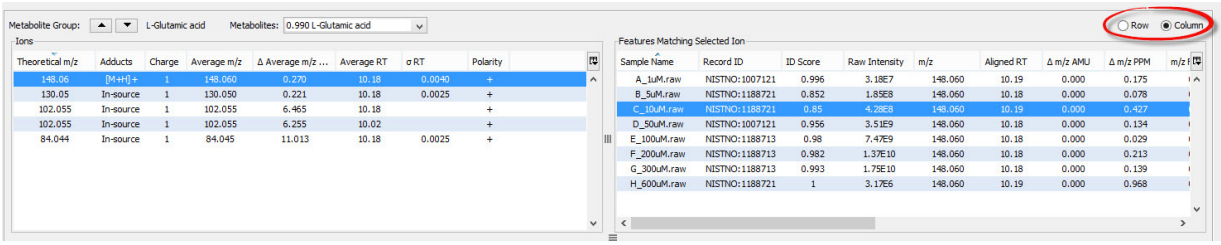
Layout

There are two possible layouts for viewing the Ions pane and the Features Matching Selected Ion pane available in this view:

- **Row layout** - The two panes are stacked one on top of each other with the Ions pane placed on top, see Figure 7-1.
- **Column layout** - The two panes are placed side by side with the Ions pane placed on the left side, see Figure 7-1.

-

Figure 7-3: Analytes View Column layout



Ions pane

The **Ions** pane contains a table that lists all of the consensus features that eluted from the LC column at approximately the same retention time (RT) and that also matched the analyte selected from the Analytes pull down menu.

Each row in the table represents a consensus feature matched to an entry in the searched spectral library that represents an adduct or loss ion of the selected analyte.

The columns in the table provide the following information:

- **Theoretical m/z**--Theoretical mass of the ion calculated from its formula
- **Adducts** --Type of adduct or loss ion matched to the consensus feature
- **Charge**--Charge of the matched ion
- **Average m/z** -- m/z value that characterizes the consensus feature. It is the average of the m/z values of the features making up the consensus feature
- Δ **Average m/z (PPM)**--Difference between the Average m/z and the theoretical m/z expressed in parts per million
- **Average RT**--Average retention time of the consensus feature
- σ **RT**--Retention time standard deviation
- **Polarity**-- Polarity (ionization mode) of the scans

When a row is selected, it is highlighted in blue and the Features Matching Selected Ion pane is updated to show the features related to the selected ion

Features Matching Selected Ion pane

This pane contains a table that lists the features in each of the MS samples in the experiment that comprise the consensus feature selected in the Ions Pane. For each feature, the table displays information that characterizes the feature or that helps in evaluating the quality of its match with the corresponding library entry. Note that only the best ID score matches are listed in the table.

The columns in the table provide the following information:

-

- **Sample Name** -- Name of the sample that includes the selected feature
- **Library** -- Name of the library that provided the highest scoring match for the feature
- **Record ID** -- Spectral library record ID for the specific match
- **ID Score** -- Overall score that is computed using a weighted average of the individual scores, see "Analyte ID Score" on page 213
- **Raw Intensity** -- Trapezoidal approximation of the area under the XIC curve
- **m/z** -- Intensity weighted average of the raw (m/z, RT, I) data points comprising the feature, see "Agglomerative Point Clustering Feature Finding Algorithm" on page 223
- **Aligned RT**-- Retention time after alignment
- **Δm/z (AMU)** -- Difference between theoretical and observed m/z in AMU
- **Δm/z (PPM)** -- Difference between theoretical and observed m/z in PPM, calculated as follows:

$$\text{Delta (PPM)} = \frac{|\text{Theoretical m/z} - \text{observed m/z}|}{\text{Theoretical m/z}} \cdot 10^6$$

- **m/z FWHM** -- m/z Full Width at Half Maximum of the trace, see definition FWHM
- **RT FWHM** -- RT Full Width at Half Maximum of the XIC
- **Charge** -- Analyte feature charge. The sign of the charge is determined by the instrument setting, the number is determined by Elements' Isotopic Clustering Algorithm.
- **Isotopic Distribution Score** -- See "Isotopic Distribution Score" on page 214
- **Mass Accuracy Score** -- See "Mass Accuracy Score" on page 214
- **MS2 Score** -- See "MS2 Score" on page 216
- **XIC Score** -- See "XIC Score" on page 217
- **Attributes** -- List of attributes assigned to the sample in the Organize View

Visualization pane

The lower half of the Analytes View, or Visualization Pane, is split into two sub-panes, one on the right and one on the left side of the view. The pane contains a total of seven tabs distributed over the two sub-panes. The tabs can be moved from one sub-pane to the other for visualization convenience, see "Moving tabs" on page 119. When selected, a tab appears in the forefront. Each tab includes a variety of tools that facilitate the visual inspection of the data and help the user to asses the validity of the identification.

- References tab
- XIC tab

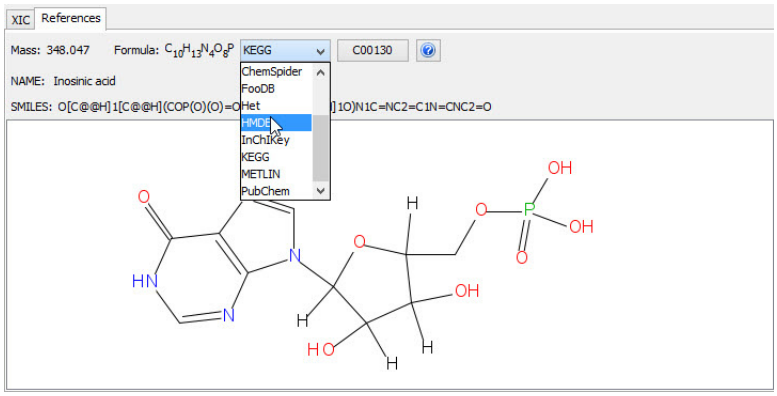
-

- Isotopic distribution tab
- MS2 Spectra tab
- Analyte Level ChartsTab
- Feature 3D viewer (two interconnected tabs)

References tab

This tab lists the name, mass (AMU), chemical formula, SMILES, molecular structure and accession number of the analyte associated with the selected ion. The type of accession number shown can be selected from a pull down list, and the accession number is reported on a button. When the button is clicked, it opens the default browser to a page in the ChemSpider or PubChem database (as selected in Edit>Preferences> Web Links) that shows further information about the analyte, see Figure 7-4.

Figure 7-4: Visualization Pane: References tab



The lower section of the tab shows a drawing of the molecular structure of the identified analyte. A context menu is available when the user right clicks over the drawing board. The menu provides the following functionalities:

- **Copy** - Copies the molecular structure appearing on th board
- **Save** - Saves the molecular structure in four possible file formats:
  - **MDL MOL**[1] **file** - A file format for holding information about the atoms, bonds, connectivity and coordinates of a molecule.
  - **SMILES** - The simplified molecular-input line-entry system (SMILES), a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings.
  - **CDK source code fragment** - Chemistry Development Kit code fragment.

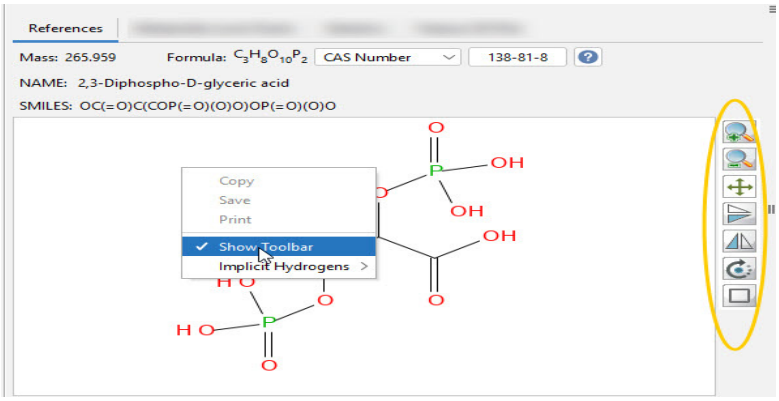
1. The MOL format was created by MDL Information Systems (MDL), which was acquired by Symyx Technologies then merged with Accelrys Corp., and now called BIOVIA, a subsidiary of Dassault Systemes

-

- **Chemical Markup language** - Chemical Markup Language (ChemML or CML), an approach to managing molecular information using tools such as XML and Java. CML uses XML's portability to help CML developers and chemists design inter-operable documents.

- **Print** - Sends the image to a local printer
- **Show Tool Bar** - Activates a graphical tool bar on the left side of the drawing board. The bar includes the following tools:
  - Zoom In
  - Zoom out
  - Move Molecule
  - Flip Horizontal
  - Flip Vertical
  - Rotate Molecule
  - Resize to fill viewing area
- **Implicit Hydrogens** - Has two options:
  - **On All** - Tags all the atoms in the molecular structure.
  - **Off** - Hides implicit Hydrogens and Carbons

Figure 7-5: Visualization Pane: context menu and Drawing tool bar



When the mouse is hovered over the molecular structure, a hand cursor appears and the wheel on the mouse acts as a zooming tool. Left clicking the mouse over the structure activates the “Move molecule tool” visualized as a cross.

XIC tab

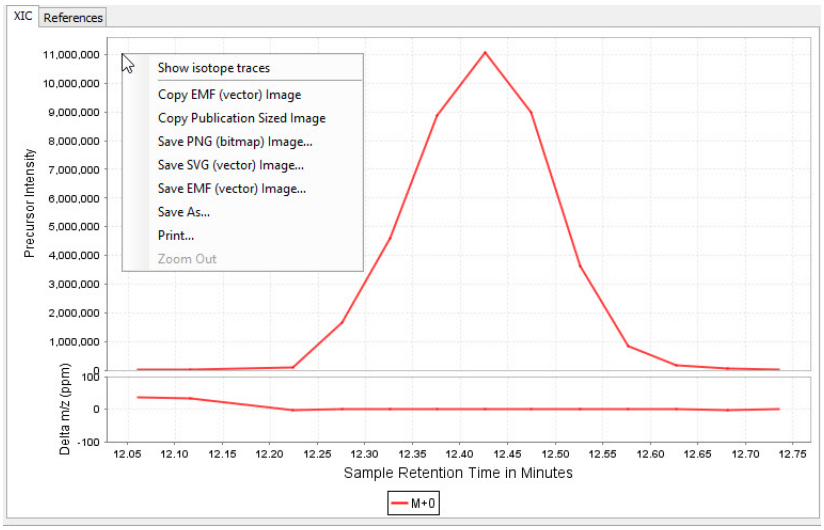
The tab contains two graphs stacked vertically one on top of the other:

1. **Plot of the Extracted Ion Chromatogram (XIC)** - which shows the Precursor intensity as a function of the retention time (RT). The plotted intensity is calculated by summing

the intensities of all of the features appearing in the m/z dimension at a particular retention time.

2. **Plot of Δm/z as a function of RT** - where Δm/z represents the difference between the m/z of the main feature and the intensity weighted m/z average of other features found around it. The plot shows if the intensity appearing in the XIC at a specific RT comes from feature aligned with the m/z of the main feature or not. The context menu available for this graph contains an option that allows the display of XICs for heavier (+1 and +2) isotope traces.

Figure 7-6: XIC tab with context menu

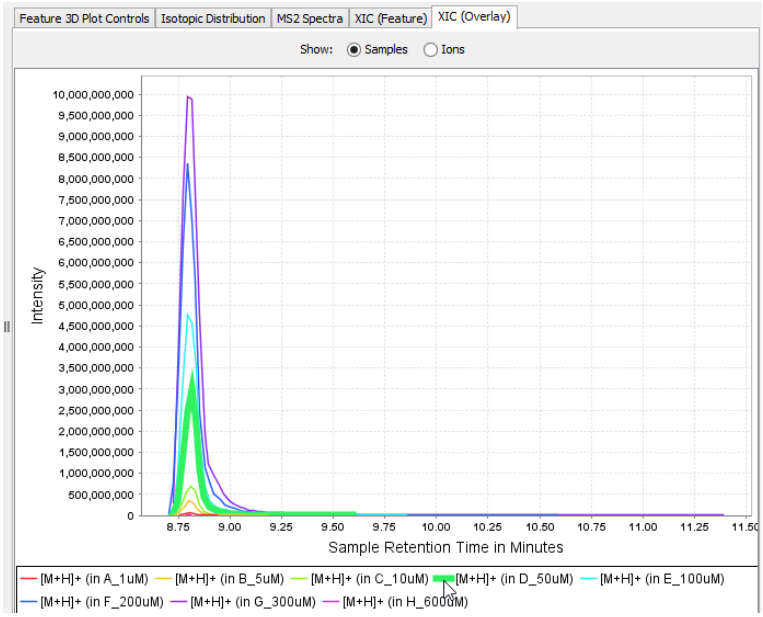


XIC Overlay tab

The XIC (Overlay) plot is a powerful visualization feature that facilitates manual validation of cross-sample consensus feature formation and within-sample multiple-ion association. The XIC Overlay tab operates in two modes. Either Samples or Ions mode is selected through a radio button at the upper left of the tab. In Samples Mode, the tab displays the XICs of the monoisotopic features of the currently selected ion in the various samples plotted together. Note that the unaligned retention time values are used.

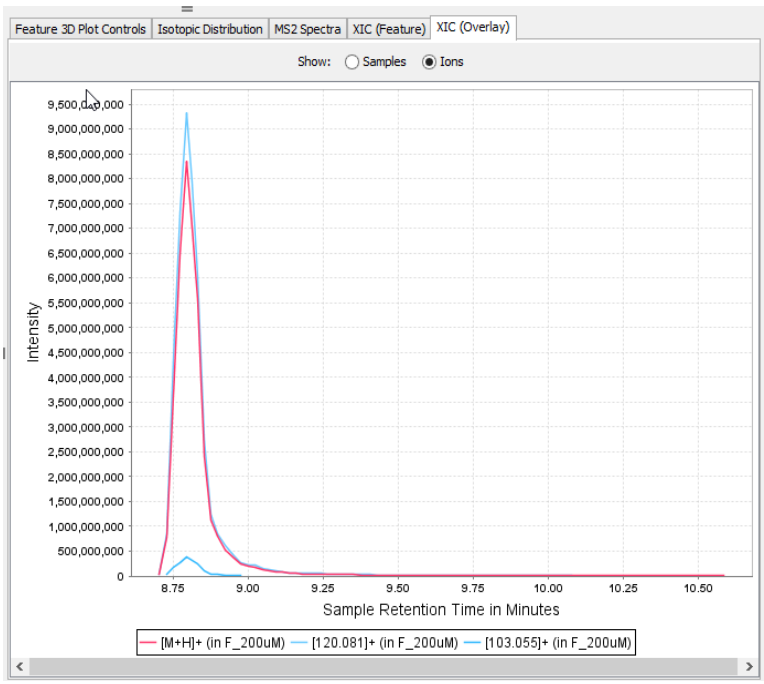
-

The XIC (Overlay) tab in Samples Mode



In Ions Mode, the XICs of all ions of the current analyte in the sample that is currently selected in the **Features Matching Selected Ion** table are plotted together.

XIC (Overlay) tab in Ions Mode



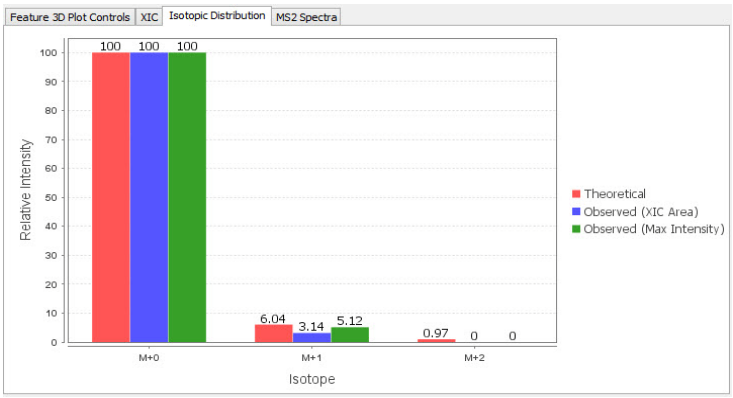
-

Isotopic distribution tab

The isotopic distribution tab shows, for the specific feature selected in the **Features Matching Selected Ion** table, a comparison between the observed abundances of the isotopic forms and their expected theoretical abundances.

The comparison is shown using a bar chart where the relative abundances in the theoretical model (red), are compared with the observed values, as measured using the XIC area under the curve (blue), and the Observed values measured using the max intensity (mz, RT, I) . The three values are plotted for the [M+1], [M+2] and [M+3] isotopes.

Figure 7-7: Isotopic Distribution tab



Both measures of the observed iosotopic abundances are used in calculating the Isotopic Distribution Score (see "Isotopic Distribution Score" on page 214).

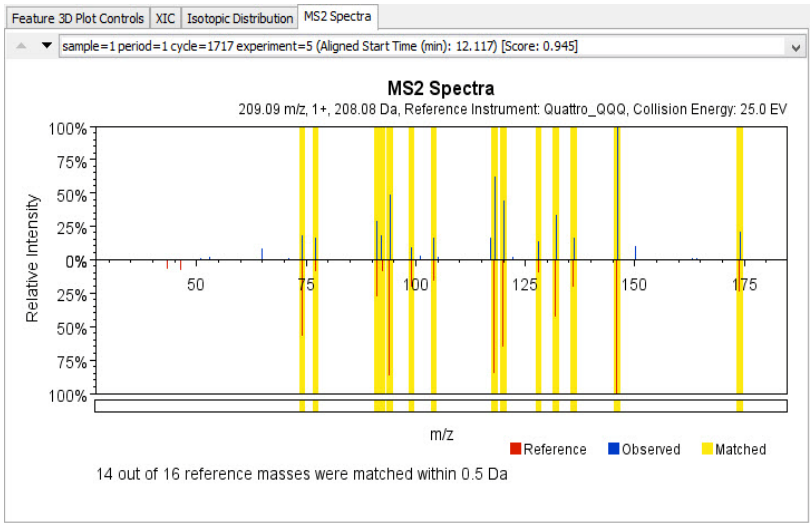
MS2 Spectra tab

This tab provides a visual representation of the observed MS2 spectra associated with the feature selected in the **Features Matching Selected Ion** table. If the matching entry in the spectral library contains an MS2, it is also displayed in the same graph, allowing the user to evaluate the quality of the MS2 match.

When there is more than one MS2 spectrum matching a feature, the user may scroll through the list of spectra using the up and down arrows or the pull down list at the top of the tab. Note that the Features Matching Selected Ion pane table lists only the highest MS2 match score for each sample.

-

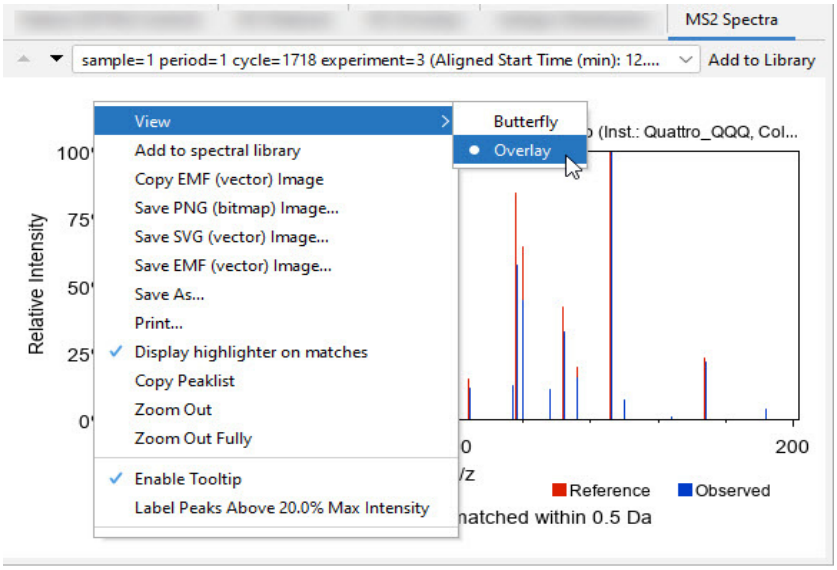
Figure 7-8: MS2 Spectra tab - Butterfly Plot



Scaffold Elements offers two different formats for comparing the observed MS2 and the library spectrum. The default view is the Butterfly plot, in which the intensities of the peaks in the observed MS2 are plotted in the positive direction, while the peaks of the library MS2 are plotted in the negative direction on the same axis. The Observed peaks are shown in blue, the library spectrum's (Reference) peaks are red, and when the two match, may be highlighted in yellow. Highlighting may be toggled through the Display Highlighter on Matches option in the context menu which is activated by a right-click in the spectrum tab.

The other display mode for comparing the MS2 spectra is the Overlay Plot, which is also activated through the View option in the context menu.

Figure 7-9: MS2 Spectra tab - Overlay Plot



In the Overlay Plot, both spectra are plotted together on the same axis with the Reference

peaks shown in red and the Observed peaks shown in blue.

The right click context menu provides further tools to facilitate visual inspection of spectral matches. Options include exporting peak lists, and turning on labeling of peaks with intensities at least 20% of the maximum intensity in the spectrum.

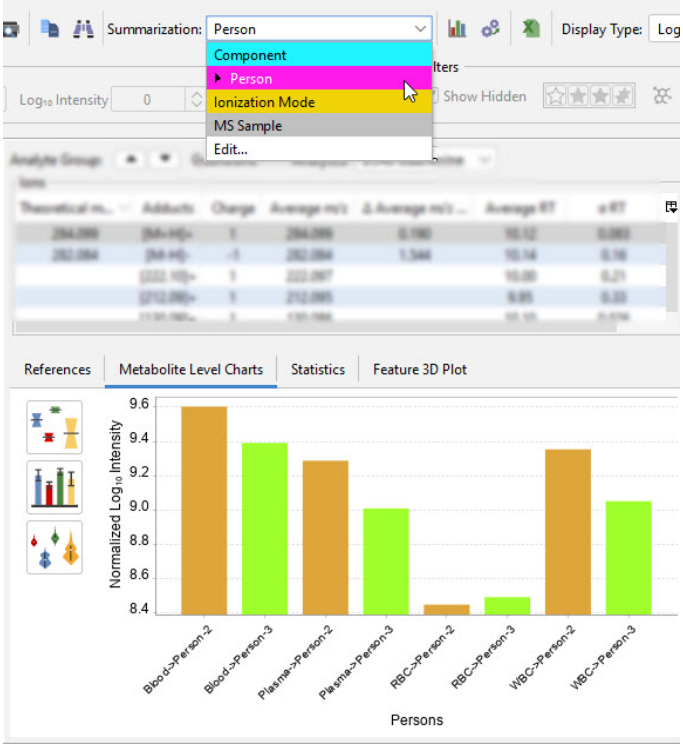
Add Spectrum to Library

An important feature of the MS2 Spectra Tab is the **Add to Library** button. This button allows the user to add the spectrum to a personal library

Analyte Level ChartsTab

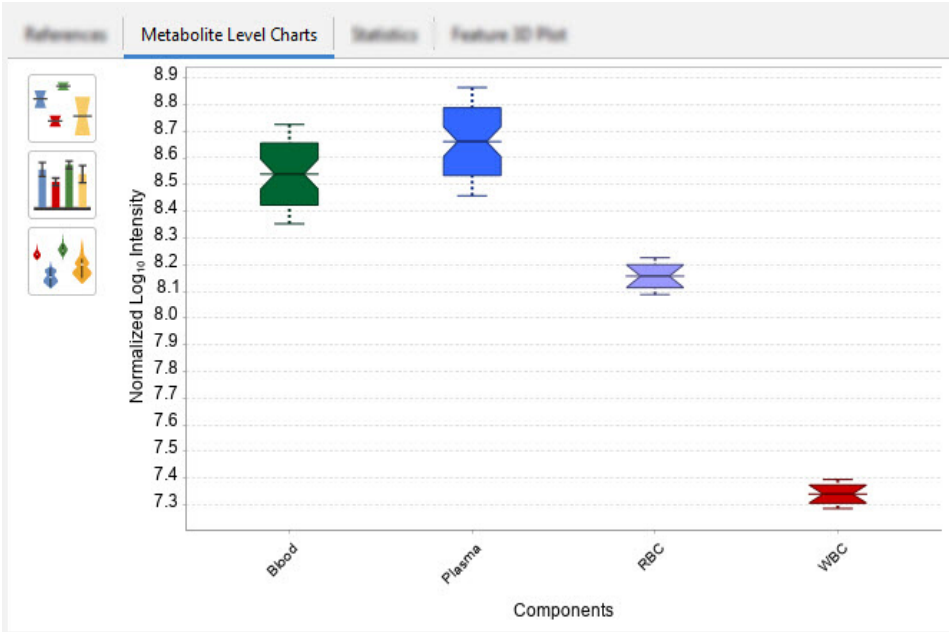
This tab provides charts showing the quantitative distribution of the data according to the selected level of summarization. Two buttons on the left side of the chart switch the plot to a box plot, a bar chart or a violin plot. The summarization level can be changed through the summarization pull down list.

Figure 7-10: Quantitative chart tab



-

Figure 7-11: Quantitative Box Plot



Statistics Tab

Statistics Table - The contents of this table depend on which statistical test has been applied to the experiment. It presents all of the values relevant to the calculation of that particular test. For example, for the ANOVA/t-test, columns displayed are: Sum of Squares, degrees of freedom, Mean Square, F-statistic, and significance of F-statistic.

Interaction Chart - This chart is designed to help the user interpret interaction effects between primary and secondary variables. The chart consists of a series of graphs with one line for each Attribute in the secondary comparison Category. Each of these plots a series of points, one point for each Attribute in the primary comparison Category, representing the average quantitative value of all technical replicates with the indicated combination of Attributes.

If there is no interaction between the primary and secondary variables, the lines would be roughly parallel, with the slope of the lines indicating the effect of the primary variable. If the lines are not parallel, but do not cross, there is an interaction effect but it is still possible to draw conclusions about the effect of the primary variable with the understanding that the secondary variable affects the size of that effect.

If, on the other hand, the lines cross, it means that the two variables exhibit significant interaction and it is impossible to draw general conclusions about the effect of the primary variable.

Feature 3D viewer

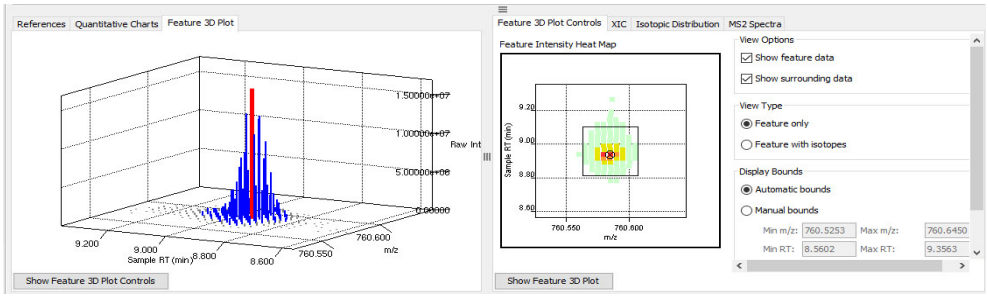
This viewer helps the user to visually inspect the peak-picked features. It includes two tabs with interdependent functionality located in the Visualization Pane at the bottom of the view.

-

1. **Feature 3D Plot tab** - This tab contains a 3-D graphical environment where the user can examine the shape of an identified feature. By clicking and moving the mouse over the Feature 3-D plot the user may rotate the point cluster representing the feature. All of the data points belonging to the analyte feature are drawn in blue while the single point corresponding to the (m/z, RT, I) that characterizes the feature is colored in red.
2. **Feature 3D Plot controls tab** - This tab includes the Feature Intensity Heat Map, which shows a top down 2-D view of the feature. The heat map is color coded according to the data point's intensity using a color gradient in which red highlights the most intense points. Clicking and dragging the mouse over the heat map selects an area and at the same selects the corresponding area in the Feature 3D Plot, and upon release zooms in within both graphs. Clicking back over the heat map zooms out. A number of view options are also included on the right side of the tab for toggling the visualization of the observed feature, its related isotopic peaks and/or surrounding data.

Note: If the user has selected the option to "Save Indexed Feature Files" (see Feature Finding tab) **during loading and has not moved the METDB file or otherwise broken its connection to the saved files, the Feature 3D Viewer offers the option to view not only the selected feature, but also the surrounding raw data, and an option is also offered to allow browsing of a specified m/z range. If the files have been moved but are still accessible, the user may reestablish the connection by using the "Set Data Source" dialog accessed through the Experiment menu. Using this dialog, the user may also choose to view only the selected features or to disable the 3D Viewer entirely. This may increase the responsiveness of the GUI when viewing large files.**

Figure 7-12: The Feature 3D Viewer tabs.



Moving tabs

For convenience, any tab can be dragged from the right pane to the left pane or vice versa. To do so the user should left click over the tab to be moved and drag it to the other pane while holding down the left mouse button.

-

Figure 7-13: Visualization pane: moving tabs

-

Chapter 8
The Visualize View

The Visualize View offers a variety of graphical tools to help the user discover quantitative trends and relationships between analytes and samples.It consists of three tabs: Principle Component Analysis, Quantitation, and a Heat Map of the filtered analytes list shown in the Samples table.

- "Quantitative Pane" on page 122, which provides a Volcano Plot to help identify which analytes exhibit significant differential expression, a Quantitative Scatterplot to show the relationships between values in different samples or categories, and a plot that is intended to help the user assess the quality of the quantitation in the experiment.
- "Principal Component Analysis tab" on page 126, which helps identify the underlying sources of variation in the data set.
- "Heatmap Tab" on page 129, which provides a graphical environment where a Heat map based on the findings listed in the Samples table is provided.

-

Quantitative Pane

The Quantitation tab consists of four plots, which are visible under different circumstances: the Volcano Plot appears when a quantitative test comparing two groups has been applied; the Quantitative Scatterplot appears when the selected summarization level contains two or more groups; the Quantitative Trend Chart apears whenever there is precursor intensity data and in the experiment.

Volcano Plot

The Volcano Plot makes it easy to identify analytes that exhibit significant quantitative differences among the samples. In a volcano plot, the y-axis represents the $-\log_{10}$ of the p-value and the x-axis represents the $\log_2$ fold change, calculated by comparing the rolled up intensity value of one category to that of another. As a result, the plot is only available when the following conditions are met:

1. The summarization hierarchy is arranged so that exactly two Attribute values have been selected for comparison in the statistical analysis, allowing calculation of a fold change.
2. A Statistical Test has been applied.

In the Volcano Plot, points with statistically significant p-values are colored green while points with p-values that would have been significant had a FWER not been applied are colored yellow. This corresponds to the coloring of the Statistical Test Result column in the Samples View.

The analytes that are most likely to be of biological significance are those which are statistically significant and also show large positive or negative fold changes. A few of these analytes are marked in Figure 8-1.

Plot Actions

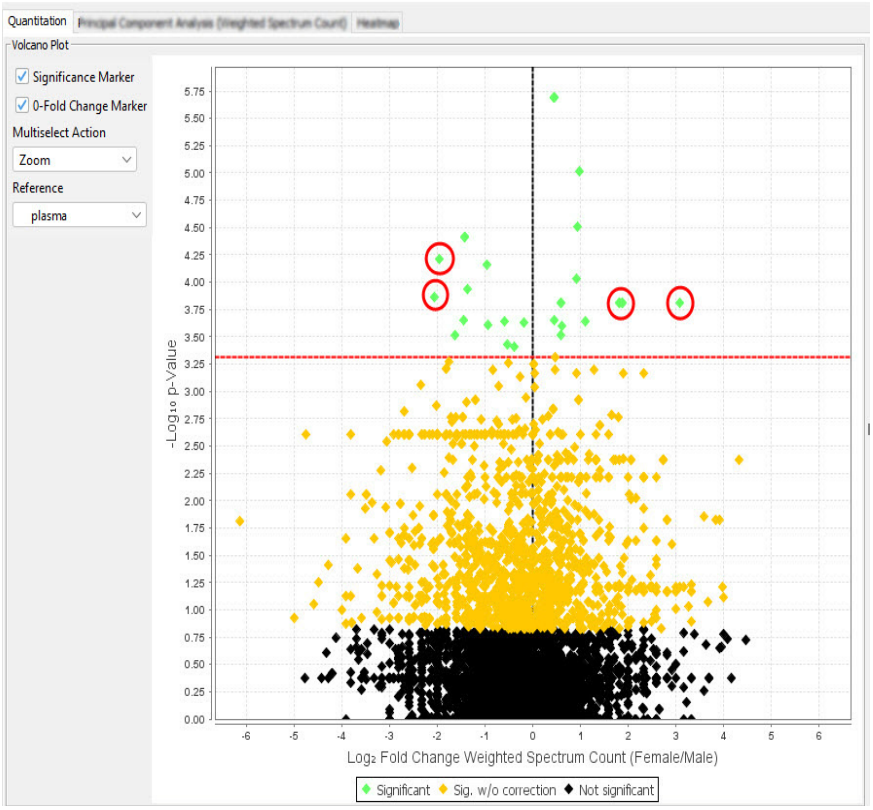
A check box at the left of the plot toggles display of a horizontal dotted line that marks the significance threshold.

Another optional line, controlled by its own check box, marks the point at which there is no difference between the compared values, i.e. the zero fold change line.

The **Multi-Select Action** pull down menu determines the behavior of the pane when the user selects a rectangular area in the plot by holding down the left mouse button and dragging the cursor. Depending on the option chosen, the graph may zoom in on the selected area or the analytes in the selected area may be tagged with stars. When stars are added or removed through the multi-select mechanism, a pop-up message informs the user of the action. The selected analytes will display the modified star status in the Samples View. This provides a convenient mechanism for filtering the analytes list based on the plot.

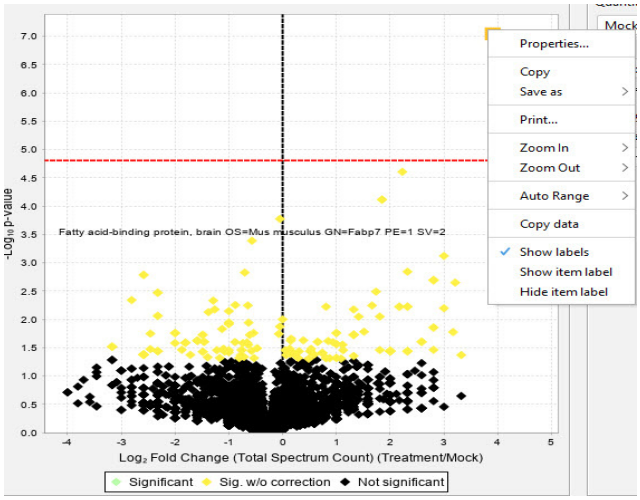
-

Figure 8-1: IThe Volcano Plot



To **Label Points** in the graph, the user right-clicks to bring up a context menu. To label an individual point, select the point and choose **Show item label**. This can be done repeatedly to label a set of points. To label all points, turn off any individual labels and click **Show labels**. containing options to label only the currently selected point, label all points, or remove all labels. Clicking **Hide item label** removes the label of the currently selected point. Clicking **Show labels** when it is already checked clears all labels.

Figure 8-2: Labeling Points in a Chart



-

This menu also offers options that allow for copying the chart, saving the Image in various formats, adjusting the chart properties or copying the data displayed in the chart in order to recreate the graphic or analyze the data using a different program.

Quantitative Scatterplot

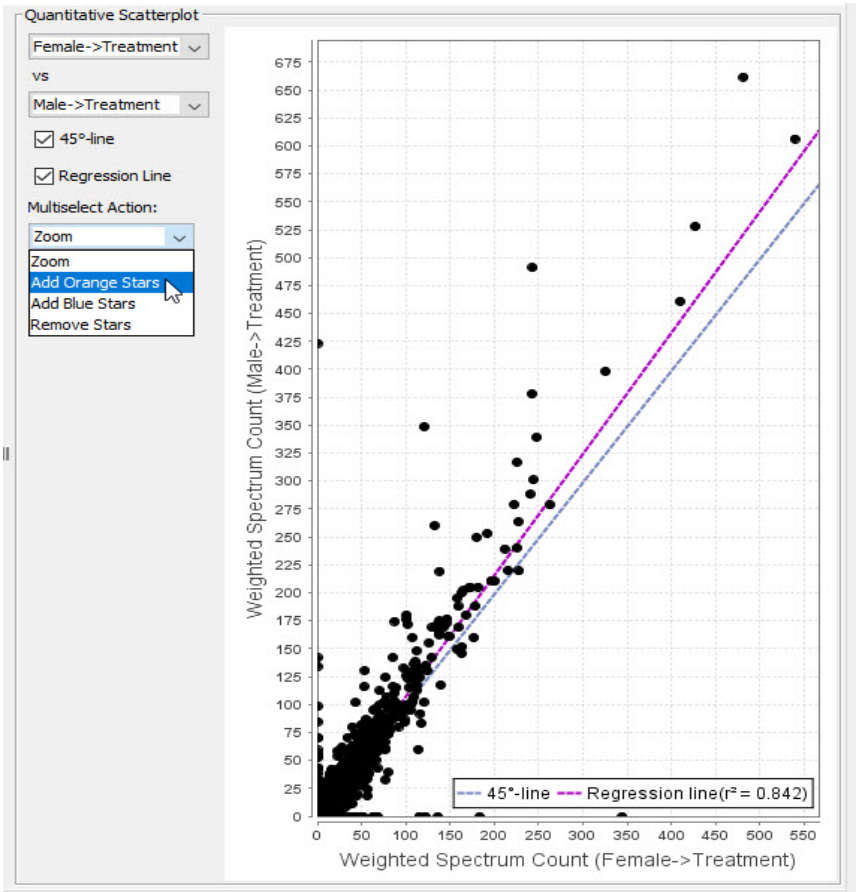
In the Quantitative Scatterplot, the quantitative values of analytes in one category are plotted against the corresponding values in another category, where the categories represent Attributes rolled up to the Summarization Level in the summarization hierarchy. The two categories are selected from pull down lists to the left of the graph.

Examination of the Quantitative Scatterplot assists the user in assessing the relationship between the two categories, and in identifying outliers, which may be analytes that are especially important in distinguishing the two groups.

Two lines may be displayed on the graph, each activated by a check box to the left of the plot. Points would be expected to cluster along the 45 degree line if the two categories are completely correlated. The regression line shows the result of performing a linear regression calculation of y on x. The correlation coefficient is shown in the legend when the regression line is displayed.

The Quantitative Scatterplot also features Labeling of Points and a Multi-select Action pull down. In the Quantitative Scatterplot, the color of the point indicates the star status of the analyte in the Samples View. Orange indicates the analyte has an orange star, blue a blue star, and purple indicates the analyte has both orange and blue stars. Setting the star status through the Quantitative Plot allows the user to return to the Samples View and filter on characteristics recognized in the graph.

Figure 8-3: The Quantitative Scatterplot



Quantitative CVs Chart

The Quantitative CVs chart contains two distinct types of plots, which, in combination, provide insight into the reliability of the quantitative values calculated in the experiment. The Quantitative CVs chart displays the relationship between mean analyte intensity and Coefficient of Variation (CV) for each group of samples at the currently selected Comparison Level (see "Experimental Design" on page 85).

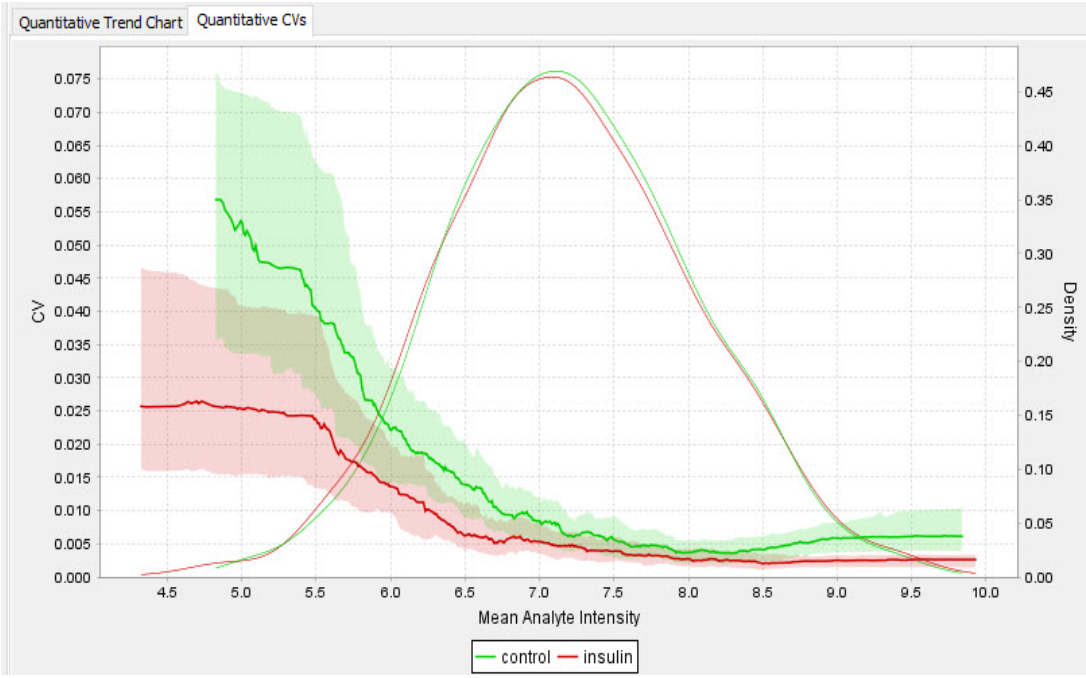
Shaded areas indicate the 50% confidence interval for CV, with the thick line showing the median value, computed for a sliding window of at least 50 analytes. The window size increases with increasing numbers of analytes. Note that the median line will appear flat if there are very few analytes in the experiment. The CV level is indicated in the y-axis displayed on the left of the chart.

A second set of plots is also displayed in the same figure. These plots show the distribution of analyte intensities within each group. The intensity levels are indicated in the y-axis on the right of the chart. The intensities plotted here are computed with a Gaussian kernel density estimate with bandwidth set by Silverman's rule.

-

The chart is built from the unfiltered, thresholded set of analytes in the experiment. Values are gathered from the level of summarization directly below the Comparison Level, and analytes are ignored if any values at that level are missing or imputed.

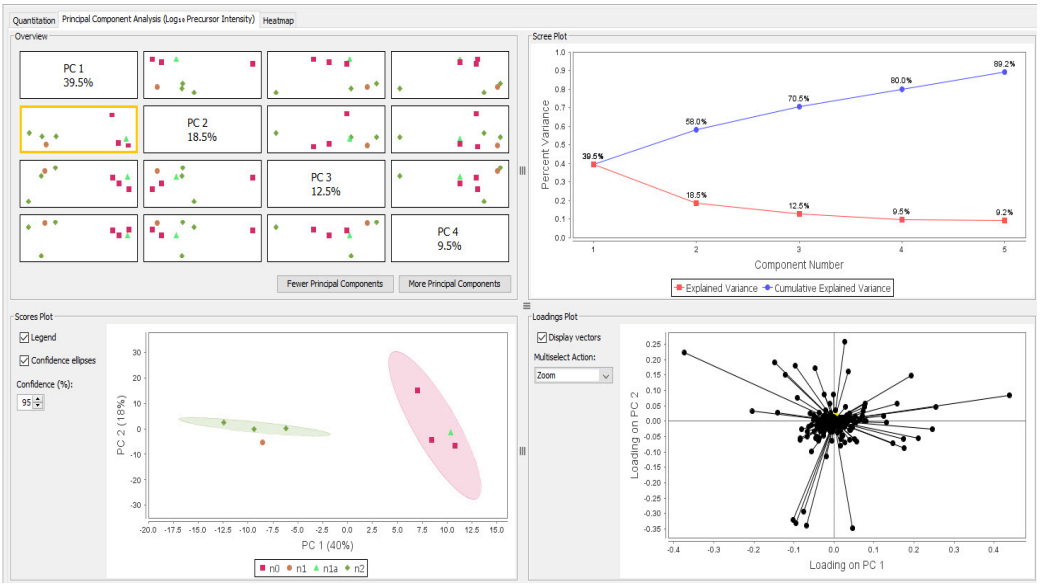
Figure 8-4: The Quantitative CVs Chart



Principal Component Analysis tab

Principal Component Analysis (PCA) is a tool used to identify the underlying sources of variation in a data set. PCA looks for patterns of expression among the analytes that can be used to group samples in meaningful ways. When used in combination with the flexible summarization offered in Scaffold Elements, this provides a powerful tool for exploring the

biological meaning of quantitative differences observed in an experiment.



The PCA tab consists of four Plots:

The Overview

The Overview consists of a series of graphs where one Principal Component is plotted against another. The points in these graphs represent samples and the X and Y coordinates are the values computed from the corresponding Principal Component functions. Samples tend to cluster in different ways depending on the Principal Components applied. Clicking on a graph in the Overview selects that combination of Principal Components for display in greater detail in the Loadings and Scores Plots below.

The Scree Plot

The Scree Plot graphs the percentage of variance explained by each Principal Component. The lower (red) plot shows the percentage of variance explained by the individual Principal Components, while the upper (blue) plot is cumulative, so the first point shows the variance explained by PC1, the second by PC1 and PC2, etc.

The Scores Plot

The Scores Plot graphs one Principal Component against another. The points in the Scores Plot represent samples (rolled up to the Biological Replicate Level specified in the summarization hierarchy) and the X and Y coordinates represent the values of the Principal Components.

The Scores Plot has several associated controls, which are found on its left. Check boxes allow the user to toggle display of the legend and sample names on or off. Another option is to show confidence ellipses for the Attributes at the Comparison Level in the summarization hierarchy. A **confidence ellipse** is a colored ellipse that represents the area in which we can

-

expect a sample with a certain Attribute to appear, with a certain level of confidence. The confidence level is adjustable through the Confidence (%) spinner. Note that if an Attribute is represented by two or fewer values, no ellipse is displayed.

The Scores Plot allows Labeling of Points through the right-click context menu and zooming in by dragging the mouse over an area of interest.

The Loadings Plot

In the Loadings Plot, each point representsone analyte. The coordinates of each analyte are a measure of the contributions of that analyte to each of the Principal Components.in the plot. For example, if the plot displays PC1 on the x-axis and PC2 on the y-axis, points far to the left and right represent analytes that contribute strongly to Principal Component 1. The analytes near the top and bottom contribute strongly to PC2.

Options available to the left of the Loadings Plot allow the user to toggle on or off the display of analyte names and vectors. The vectors connect each analyte point to the origin, and the slope of the vector corresponds to the relative contribution that analyte to each Principal Component.

Points in the Loadings Plot may be labeled through the right-click context menu The Loadings Plot also features a Multi-select Action pull down. As in the Quantitative Scatterplot, the color of the point indicates the star status of the analyte in the Samples View. Orange indicates that the analyte has an orange star, blue a blue star, and purple indicates the analyte has both orange and blue stars. Setting the star status through the Loadings Plot allows the user to return to the Samples View and filter on characteristics recognized in the graph.

Further information about PCA

For details about how PCA is calculated in Scaffold Elements, see “How PCA is Performed in Scaffold Elements” on page 260.

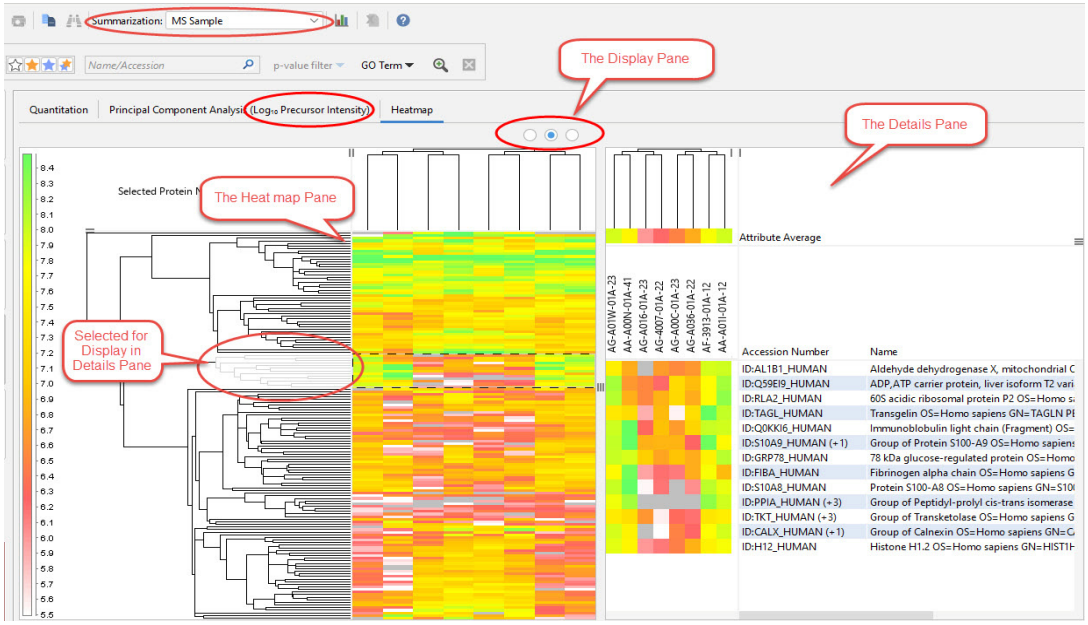
Heatmap Tab

Heat maps are an efficient method of visualizing complex data sets organized in two dimensional tables or matrices. Through the application of two independent procedures to a data matrix, heat maps make patterns more visible to the eye. The first procedure reorders columns and rows according to a “closeness” criteria which groups together in space highly similar data. The other procedure translates a numerical matrix into a color image[1].

In order to produce a figure that is meaningful, Scaffold Elements restricts the heatmap to a maximum of 1000 analytes. As a result, it may be necessary for the user to filter the analyte set before accessing the heatmap. One method of accomplishing this is to use the star filters. For example, one might select the first 1000 analytes in the Samples View, right-click and choose Stars > Add Orange, then click on the empty star icon in the toolbar to filter out all unstarred analytes. Alternatively, one could filter on statistical significance or some other criterion.

When the analyte list includes less than 1000 analytes, Scaffold Elements constructs a heatmap from the data appearing in the Samples table and displays it in the Visualize View Heatmap tab.

Figure 8-5: Visualize View: Heatmap tab



The Heatmap tab includes the following three components, each containing a number of graphical tools:

- The **Heatmap Landscape pane** -- which shows the overall heat map.

1. Key, M., “A tutorial in displaying mass spectrometry based proteomic data using heat maps.” BMC Bioinformatics 2012 13(Suppl 16):S10. doi:10.1186/1471-2105-13-S16-S10

- The Heatmap Details pane --Which shows a selected portion of the heat map and includes labels for the displayed columns and rows.
- The Heatmap Display controls--which lists the Display type selected from the Samples View and three toggle buttons

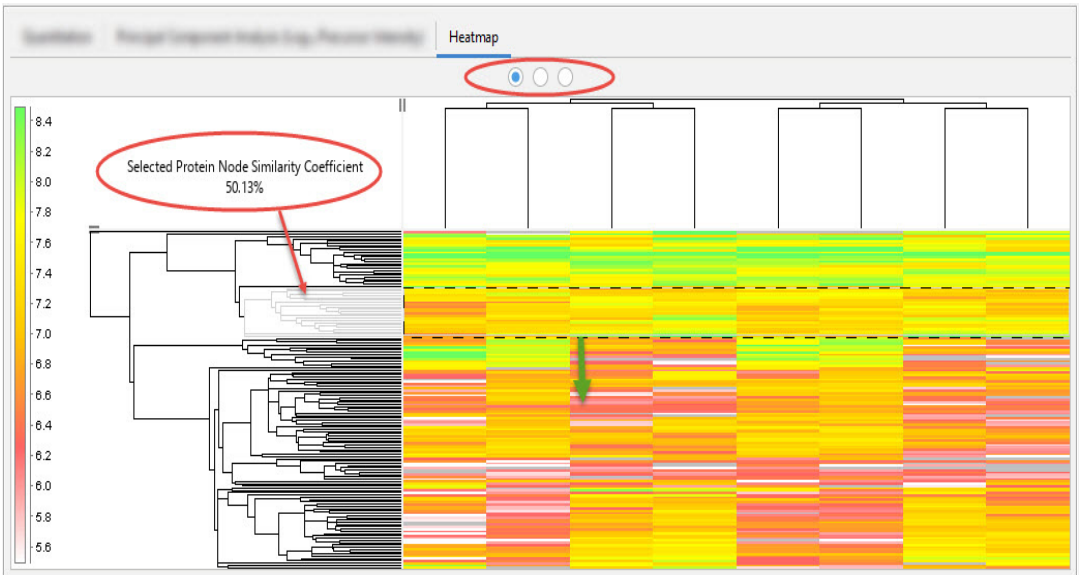
Heatmap Landscape pane

The heat map shown in this pane is created using the data summarized in the Samples Table, see Figure 8-6. As in the Samples table the rows in the Heat map represent analyte groups, but not analyte clusters as at times are shown in the Samples table. Each column contains data from every MS sample or selected level of summarization as chosen from the Summarization Bar in the Scaffold Elements main window.

The Display Type listed in the Heatmap Display controls of this tab, determines the type of quantitative information used to reorder the data.

More information about the way the Heat map is constructed can be found in the appendix section Terminology

Figure 8-6: Heatmap Landscape pane



The Heat map includes a colored landscape, color coded according to the legend shown on the left side of the Heat map pane. The type of color coding depends on the Display type chosen in the Samples View, which can be customized at will through the Color Options button.

Grays represent missing values.

Dendrograms representing the output of the hierarchical clustering are shown on the left and top sides of the Heat map landscape. The root of each dendrogram represents a single object

or cluster of size 1.

Clicking and dragging the mouse over the Heatmap landscape allows the user to select a section of the map that is then shown in larger detail in the Heatmap Details pane

Sections of the map can also be highlighted and selected by clicking over the different nodes in either dendrogram or in both, thus allowing the user to select desired sets of analyte groups and/or sets of samples. When doing so the calculated node distance is shown on the top left side of the pane.

Context menu

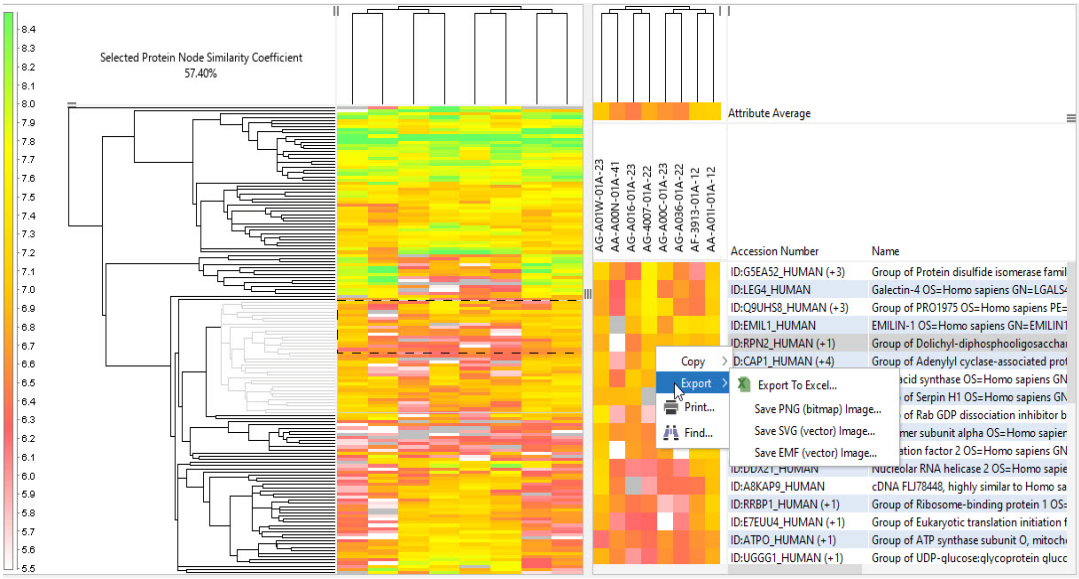
A right-click of the mouse while hovering over the Heatmap landscape provides a context menu with zoom and export options, see Figure 8-7.

Hot keys for zooming in and out of the heat map are also available:

- ZOOM IN: CTRL + NumPad +
- ZOOM OUT: CTRL + NumPad -

The Export PNG (Bitmap) Image... command opens an Export Preview dialog of the current heat map with options for toggling the inclusion or exclusion of some of the components of the exportable picture like, for example, any dendrogram or the colored landscape.

Figure 8-7: Heat map Context menu



Heatmap Details pane

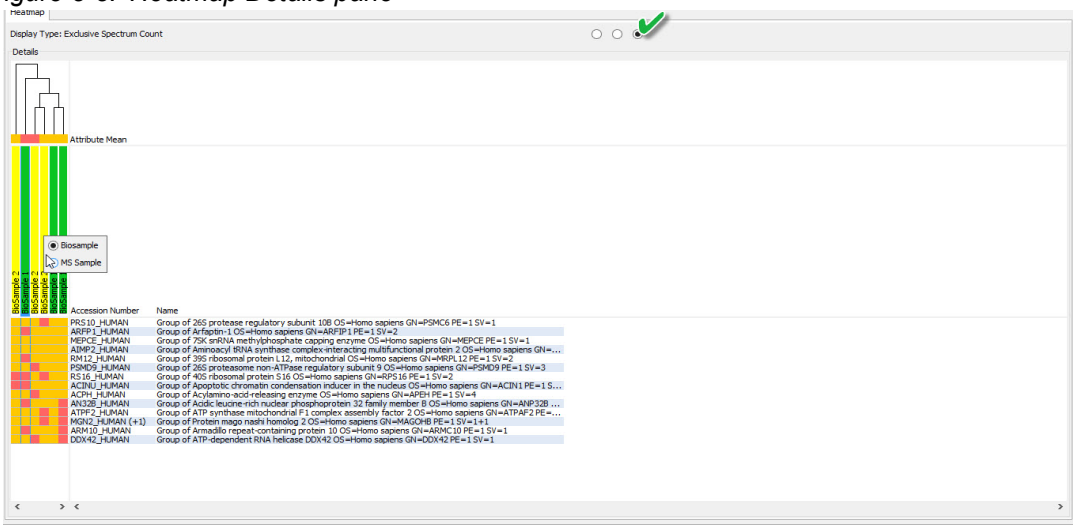
When sections of the Heat map are selected the Heatmap details pane is populated with the selected section of the map and the related information about the analyte groups associated with each row and the related MS sample or selected level of summarization as chosen from the Summarization Bar available in the Scaffold Elements main window.

When the user right-clicks on the column headers, a context menu allows the selection of the

level of summarization he/she wants to see represented within the column headers.

The table that represents the selected portion of the Heat map has the common properties of all Scaffold Elements tables, see Display paneWhen no selection is active the pane will appear empty with instructions to help the user populate the pane.

Figure 8-8: Heatmap Details pane



Heatmap Display controls

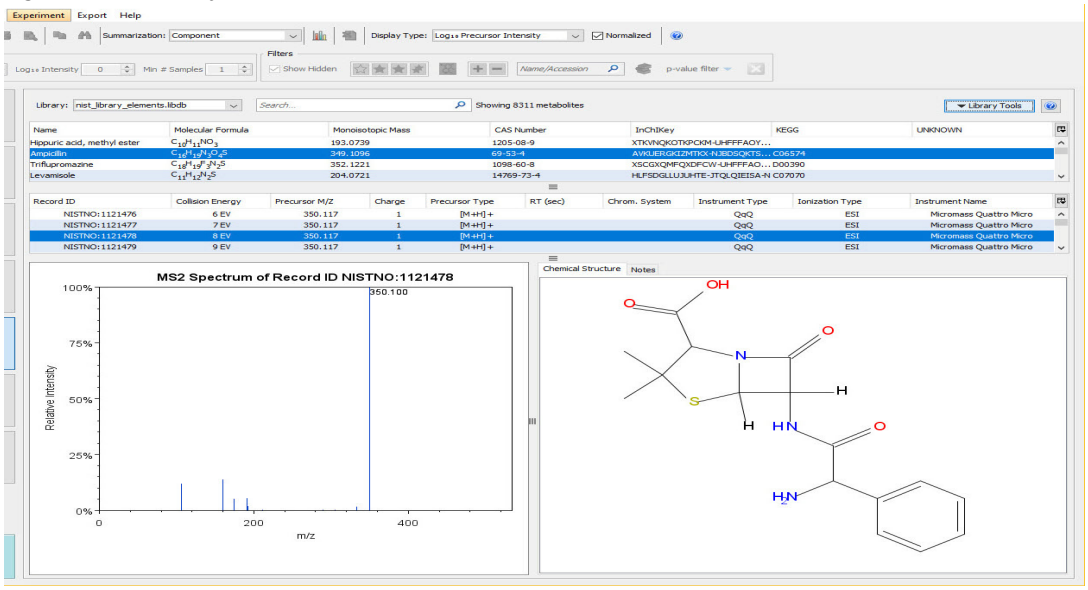
These controls are located in the top section of the Heatmap tab. The three toggle buttons allow the user to determine which of the two heatmap panes display:

- The Heatmap Landscape pane only (left button)
- The Heat map and the Details pane are shown together (middle button)
- The Heatmap Details pane only (right button).

Chapter 9
The Library View

Scaffold Elements performs analyte identification by searching against one or more analyte databases, also known as spectral libraries, which include MS and MS2 spectral information. When spectral libraries are loaded into the application, the user may visualize and inspect their contents using the Library View, see Figure 9-1.

Figure 9-1: Library View



Editable and Protected Libraries

Libraries may be editable or non-editable. Those which are imported from external sources, such as NIST, METLIN, HMDB, MoNA, etc. are non-editable, as are custom libraries created from tab-delimited text files. This prevents corruption of these libraries. Some libraries, such as METLIN, are further protected through encryption in accordance with their licensing requirements and copying or exporting their contents is not allowed.

Personal libraries created by saving Elements experimental results are editable. In the following description, fields which may be edited for editable libraries are marked with an "*". Analytes or records may be deleted from an editable library using the right-click context menu.

Copy to Library

Analytes or entries may be copied from any unprotected library to any editable library. This allows the user to create a subset library. For example, a user may wish to curate a subset of a larger library that contains only verified high-quality spectra, or only analytes of a certain

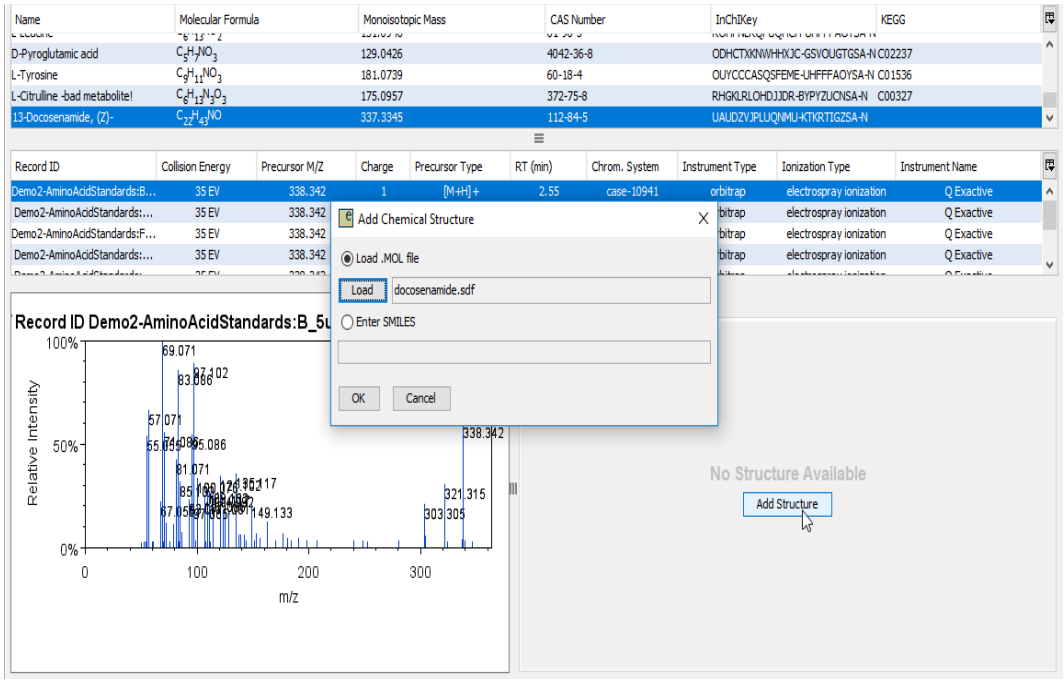
type. An editable library must be created through the Library Manager (see "Library Manager" on page 49). Analytes may be copied into this library by selecting one or more analytes in the Analytes pane, right-clicking and choosing "Copy Analyte(s) to Library...". Specific records may be copied by selecting one or more records in the Records pane, right-clicking and choosing "Copy Record(s) to Library...".

When a copy operation is selected, the **Add Records to Library** dialog appears. The user must select an editable library (or create one through the Library Manager button) and click **Add** to copy the records to the selected library.

Add Chemical Structure

In an editable library, it is possible to add or replace the chemical structure associated with an analyte in the library. If an analyte has no specified chemical structure, an **Add Structure** button appears in the Chemical Structure Tab in the Library View. Clicking the button opens a dialog that offers the option to read a structure from an SDF or MOL file, or to enter a SMILES structure.

Figure 9-2: Add Chemical Structure



In editable libraries only, when a chemical structure is present, the context menu accessed by right-clicking in the Chemical Structure Tab includes options for removing or modifying the chemical structure. From this menu, the User may remove the structure or select **Change Structure**, which opens a dialog similar to the **Add Chemical Structure** dialog and allows a new structure to be entered or uploaded from a file.

Components of the Library View

At the top of the Library View, is a Tool Bar, which allows the user to select a library, search or filter it, and perform various library-related operations. Beneath the Tool Bar are three

-

adjustable panes. The panes may be expanded or contracted by clicking on the border between panes and pulling up or down.

Note that each of the tables in the view includes all of the features and tools described in Display pane.

The components of the Library View are:

1. Tool Bar
2. Analytes Pane
3. Record Pane
4. Visualization Pane

Tool Bar

At the top of the Library view display pane, this bar contains the following tools:

- **Library:** - pull down list with which the user can select which of the spectral libraries already loaded into Scaffold Elements to browse.
- **Search...** - Search text box that searches the analyte table in the Analytes Pane for names and formulas.
- **Library Tools** - Drop-down menu that provides access to a variety of tools for managing libraries.
  - **Library Manager** - opens the Library Manager dialog, see Library Manager .
  - **Download Libraries** - opens a browser to a page on the Proteome Software website where a number of spectral libraries in the LIBDB format are available for download.
  - **Export to Skyline** - exports a transition list as a text file which may be loaded into Skyline, see .
  - **Copy Operations** - these options will be disabled if the library is protected:
    - Copy Selected Analyte(s) to Library... - opens the Add Records to Library dialog, populated with all records for all selected analytes.
    - Copy Selected Records to Library... - opens the Add Records to Library dialog, populated with only the selected records for the selected analyte.
    - Advanced Copy to Library... - opens the Advanced Copy to Library dialog to allow selection of analytes and records for copy based on their specific characteristics.
  - **Delete Operations** - these options will be disabled if the library is not editable:
    - Delete Selected Analyte(s) - removes all analytes which are currently selected in the Analytes Table in the Library View from the library.

-

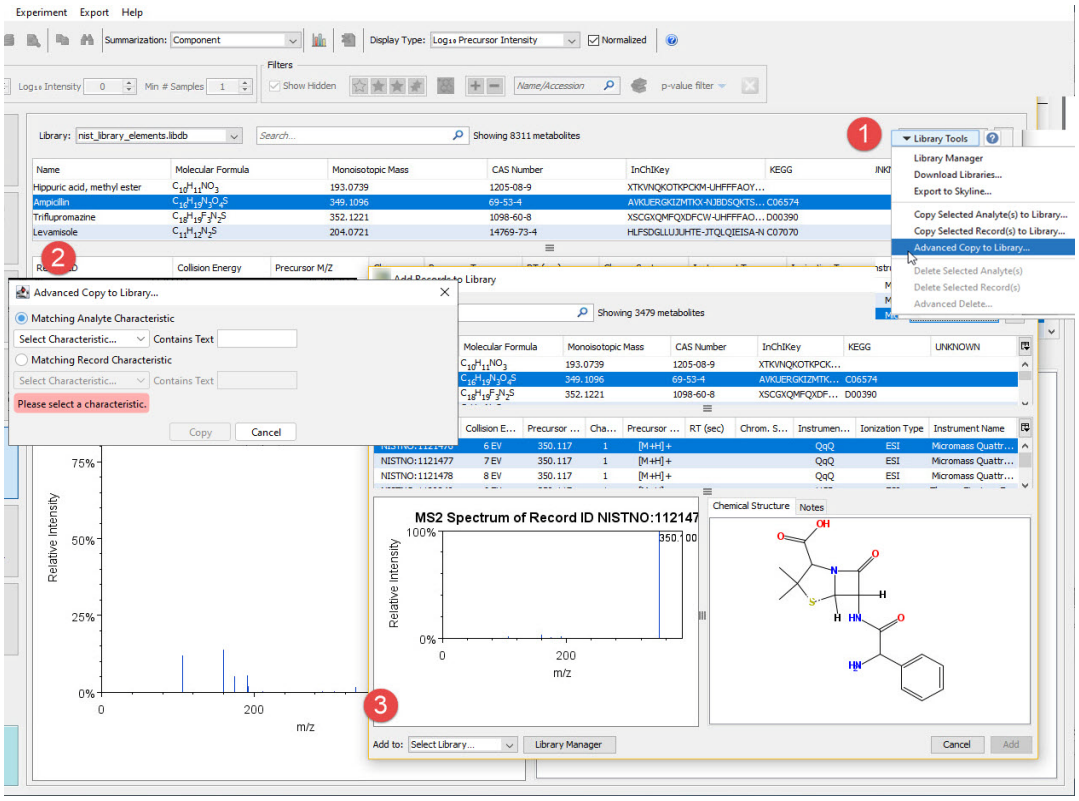
- Delete Selected Record(s) - removes all records which are currently selected in the Records Table in the Library View from the library.
- Advanced Delete - opens the Advanced Delete dialog to allow selection of analytes and records for deletion based on their specific characteristics.

Advanced Copy and Delete

The Advanced Copy and Delete options in the Library Tools allow the user to select a set of analytes or records based on specific criteria to include in or remove from a personal spectral library. Selecting the Matching Analyte Characteristic displays criteria for selecting analytes, including Name, Molecular Formula and Monoisotopic Mass. Name and Molecular Formula are text filters, and will accept a substring of the name or molecular formula to be matched. Monoisotopic Mass allows specification of a mass range.

Selecting the Matching Record Characteristic radio button offers a set of criteria for selection of analytes based on the characteristics of their records. All records which match the specified criteria are selected, regardless of which analytes they represent. Criteria at the record level include Record ID (range), Collision Energy (range and units), Precursor M/Z (range), Charge (range), Precursor Type (text), Retention Time (range), Chromatographic System (text), Instrument Type (text) and Instrument Name (text).

Figure 9-3: Using the Advanced Copy Option: 1) Select the option from the Library Tools; 2) Specify the criteria for selecting records/analytes; 3) Review the selected list and Add to a library.



-

Analytes Pane

This pane contains a table listing the analytes in the selected spectral library. Each row in the table corresponds to an analyte and lists some of its properties. The columns in the table provide the following information:

- ***Name** -- of the analyte. In an editable library, double-click in cell to edit.
- **Molecular Formula** - its molecular formula
- **Monoisotopic mass** - its mass
- A series of columns displaying the various identifiers for the analyte which are available in the library, such as:
  - **CAS number** - unique numerical identifier assigned by Chemical Abstracts Service (CAS) to every chemical substance described in the open scientific literature
  - **InChiKey** - IUPAC International Chemical Identifier, textual identifier for chemical substances, designed to provide a standard and human-readable method for encoding molecular information.
  - **KEGG -** identifier for the analyte in the KEGG Database (Kyoto Encyclopedia of Genes and Genomes). KEGG is a database resource for understanding high-level functions and utilities of the biological system. [1]It provides pathway information, genomic information, etc.

The table can be searched using the search tool available in the .

Record Pane

This pane contains a table with a list of the records in the library for the selected analyte. Each row provides the following information about each record:

- **Record ID** -- Spectral library record ID for the selected analyte.
- **Collision Energy**-- Energy needed to achieve optimum fragmentation efficiency. It is express as a percentage of the available energy in the system or Normalized Collision Energy[2].
- **Precursor m/z** -- Observed m/z for the precursor ion in MS1.
- **Charge** -- Of the precursor ion.
- **Precursor Type** -- Adduct type
- **RT (<units>)** -- Retention Time - the unit is selected through Edit>Preferences>Units>Retention Time Unit and will be either minutes (min) or seconds (sec)

---

1. http://www.genome.jp/kegg/kegg1a.html
2.

-

- **Chrom. System --** An identifier created by the user to describe the chromatographic system used in the experiment from which the record was created (see **Chromatographic System** in Loading Data**)**
- **Instrument Type** -- Instrumentation used to produce the recorded results.
- **Ionization type** -- Ionization method, such as MALDI or ESI.
- **Instrument Name** -- Type of instrumentation used to obtain the identification.

Visualization Pane

This pane is split into two graphical sub-panes:

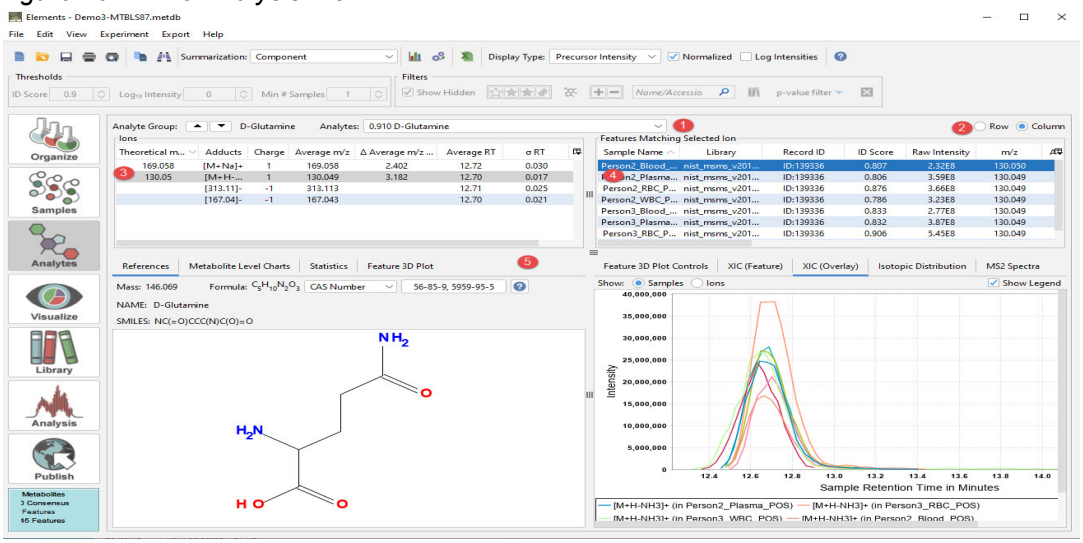
- **MS2 spectrum pane** --Shows the spectrum of the MS2 (when present) for the selected record. Additional visualization options are available in the right click context menu, see Description of Mouse Right Click Context Menu Commands
- **Chemical Structure tab**- shows the graphic representation of the molecular structure of the selected analyte. When hovering the mouse over the molecular structure a little hand appears and the wheel on the mouse acts as a zooming tool. Left clicking the mouse over the structure activates the “Move molecule tool” visualized as a cross.
- ***Notes tab -** displays the source of the library (if available) and any notes or comments related to the entry. May be edited in editable libraries.

Chapter 10
Analysis View

The Analysis View provides graphical tools to help the user visualize the samples alignment in retention time.

The view includes two tabs, one for each ionization mode. If the data is acquired in only one mode, the corresponding mode tab will be active while the other tab will be grayed out.

Figure 10-1: The Analysis View



Each tab contains two graphs, a legend and a check-box.

Total Ion Current (TIC) Chromatogram

The top graph in the tab plots the MS1 TIC for each MS sample in the experiment as a function of retention time. The TIC represents the summed intensity across the entire range of masses being detected at every point in the analysis. The chromatograms are drawn in the colors assigned to their corresponding samples, as described in the legend to the right of the graph.

The chromatograms can be displayed either aligned or unaligned by clicking the “Use aligned retention times” check box.

RT Deviation from Reference plot

The lower graph in the tab displays the differences between the raw retention times and the aligned retention times for each sample. This indicates the degree of adjustment that was

necessary in order to align the sample to the reference at any given point.

When the “Use aligned retention times” box is checked, the differences are graphed relative to the reference retention times. When the box is unchecked, each sample’s deviations are plotted relative to that sample’s retention time. This allows a comparison between the upper plot and the lower, since they are always plotted on the same time scale.

The colors of the plots correspond to the colors assigned to the samples, as described in the legend to the right of the graph.

Raw samples legend

Provides information about the color assignment of each raw data file in the experiment.

Use aligned retention times check box

This check box changes the Y axis in both graphs. When the box is unchecked, values are plotted as a function of unaligned Retention Time. The unaligned version of the TIC Chromatogram provides a comparison of the chromatography in the various samples, and the RT Deviation plots below indicate the degree of adjustment that would be required at each point in order to align the chromatograms.

When it is checked, values are plotted against the Reference Retention Time. The TIC Chromatogram provides a view of the aligned chromatograms, and the RT Deviation plot indicates the degree to which each sample was adjusted in order to achieve alignment.

-

Chapter 11
The Publish View

The Scaffold Elements Publish View displays detailed information about the data loaded and the analytic methods applied in the current experiment. This is usually required for publication.

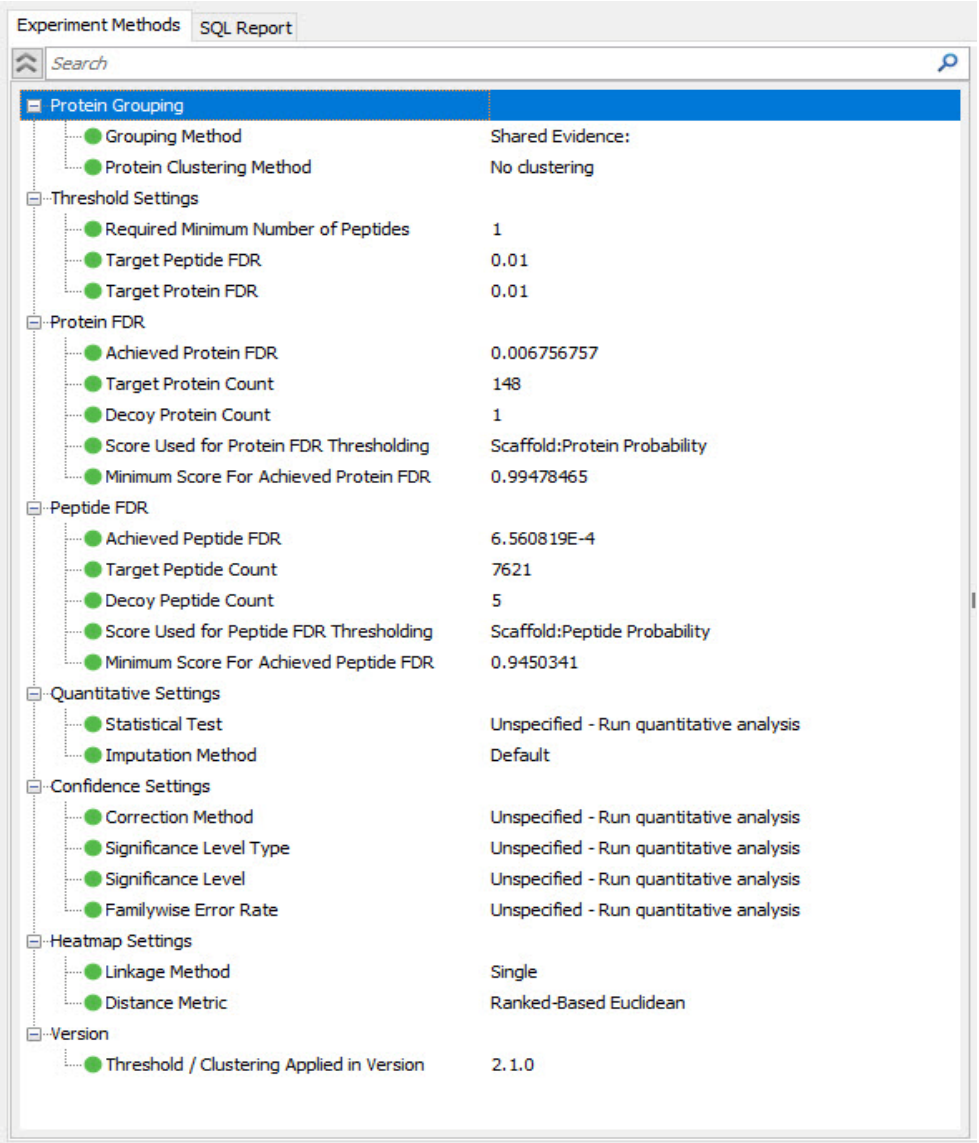
The Publish View has two tabs:

- Experiment Methods Tab
- SQL Export tab

Experiment Methods Tab

The left side of the Experiment Methods Tab contains a tree that lists the parameters characterizing the current experiment and their values. This information must be included in pubications.

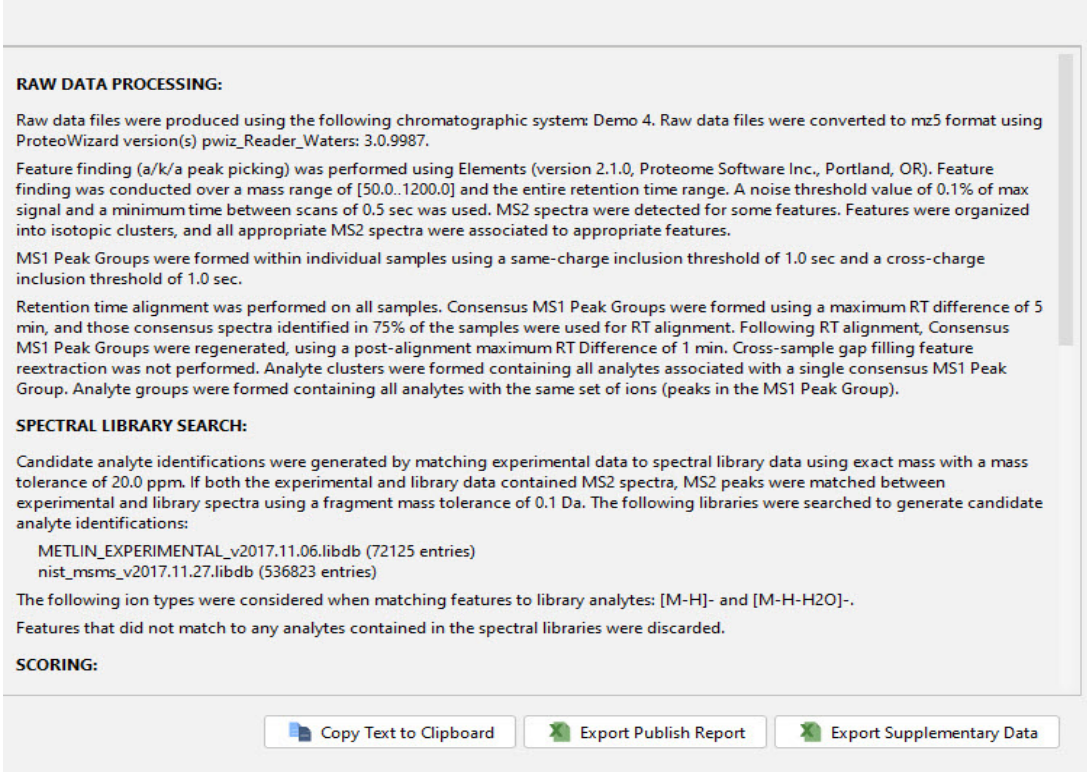
Figure 11-1: Experiment Information Pane tree



Analysis Overview

The right side of the Experiment Methods tab of the Publish View provides a draft version of the analysis parameters in text format to help the user in writing the Methods section of a publication or poster.

Figure 11-2: TheAnalysis Overview (Experimental Methods)



RAW DATA PROCESSING:

Raw data files were produced using the following chromatographic system: Demo 4. Raw data files were converted to mz5 format using ProteoWizard version(s) pwiz_Reader_Waters: 3.0.9987.

Feature finding (a/k/a peak picking) was performed using Elements (version 2.1.0, Proteome Software Inc., Portland, OR). Feature finding was conducted over a mass range of [50.0..1200.0] and the entire retention time range. A noise threshold value of 0.1% of max signal and a minimum time between scans of 0.5 sec was used. MS2 spectra were detected for some features. Features were organized into isotopic clusters, and all appropriate MS2 spectra were associated to appropriate features.

MS1 Peak Groups were formed within individual samples using a same-charge inclusion threshold of 1.0 sec and a cross-charge inclusion threshold of 1.0 sec.

Retention time alignment was performed on all samples. Consensus MS1 Peak Groups were formed using a maximum RT difference of 5 min, and those consensus spectra identified in 75% of the samples were used for RT alignment. Following RT alignment, Consensus MS1 Peak Groups were regenerated, using a post-alignment maximum RT Difference of 1 min. Cross-sample gap filling feature reextraction was not performed. Analyte clusters were formed containing all analytes associated with a single consensus MS1 Peak Group. Analyte groups were formed containing all analytes with the same set of ions (peaks in the MS1 Peak Group).

SPECTRAL LIBRARY SEARCH:

Candidate analyte identifications were generated by matching experimental data to spectral library data using exact mass with a mass tolerance of 20.0 ppm. If both the experimental and library data contained MS2 spectra, MS2 peaks were matched between experimental and library spectra using a fragment mass tolerance of 0.1 Da. The following libraries were searched to generate candidate analyte identifications:

METLIN_EXPERIMENTAL_v2017.11.06.libdb (72125 entries)
nist_msms_v2017.11.27.libdb (536823 entries)

The following ion types were considered when matching features to library analytes: [M-H]- and [M-H-H2O]-.

Features that did not match to any analytes contained in the spectral libraries were discarded.

SCORING:

Report Buttons:

- Copy Text to Clipboard - copies the contents of the Analysis Overview pane for pasting into a text editor.
- Export Publish Report - exports the contents of the Experiment Methods table as a CSV file which can be opened in Excel.
- Export Supplementary Data - exports a CSV file similar to the Samples Report suitable for submission as a supplementary data table.

SQL Export tab

The experiment files created by Scaffold Elements, *.sfdb files, are essentially SQLite databases.

The SQL export tab is a SQLite graphical interface where a Scaffold Elements experiment file can be searched as a database using SQLite commands. This allows a User to create custom tables exportable to Excel.

A depiction of the schema of a *.SFDB file is shown in the Appendix Structure of Scaffold Elements files (*.metdb) in the Scaffold Elements User's Guide. The default SQL query which appears in the SQL pane when the program is opened displays all tables in the database.

Figure 11-3: The SQL tab

Figure 12: The SQL tab

The SQL Export tab contains four different panes:

- The SQL pane
- The Saved Queries pane
- The Results pane
- The Icon pane

The SQL pane

Through the SQL pane it is possible to directly explore the information stored in a Scaffold Elements file using SQLite queries.

- The SQL text box --where the user can enter, copy and paste SQL queries.
- The SQL Icon pane -- which contains the Run query button and a text box and a save button to save queries.

The results of the queries are shown in The Results pane. The saved queries are listed in the The Saved Queries pane

Example:

List of tables available in *.SFDB files.

SELECT name FROM SQLite_master WHERE type='table' ORDER BY name;

The Saved Queries pane

When a query is saved, with a name selected by the user, it will appear in this pane from where it is conveniently available to be launched again whenever needed.

The Results pane

When the run query button is pressed, if there are no errors, a table with the results of the query will appear. To export the results the User needs to right click the mouse and select the menu option **Export > Export to Excel** and save the table to a tab delimited file that can be easily opened in Excel.

The Icon pane

The icon pane contains an icon to save new queries to a file that can later be retrieved and an icon to import previously saved queries.

SQL Export tab

-

Chapter 12
Metabolomic Flux Analysis

In this type of experiment, isotopic labels are used to trace the transformation of metabolites through metabolic pathways by comparing the stoichiometric ratios of different labeled species of the same metabolite at different time points or in different samples.

First , unlabeled samples are loaded and searched as usual in Scaffold Elements. These samples will be used for identification of metabolites.

Once the unlabeled samples have been analyzed, the user may specify labeled samples which have undergone similar chromatographic processing and which are expected to contain the same metabolites, although some of them will be found in labeled forms.

Scaffold Elements uses the unlabeled samples to provide the locations in retention time and m/z of the identified metabolites. It then searches in the corresponding locations in the labeled samples for the presence of isotopic variants of those metabolites. Results are presented in an export, which can then be analyzed in Excel to provide information about the changes in the metabolites from sample to sample.

Launch Isotopic Flux Experiment

This dialog is launched either by clicking on the Flux icon in the toolbar or through the Experiment>>Flux Analysis... option in the Experiment menu. This dialog allows the user to select labeled sample files, equate these with one or more unlabeled samples in the current Scaffold Elements experiment, and set various parameters to be used in the flux analysis.

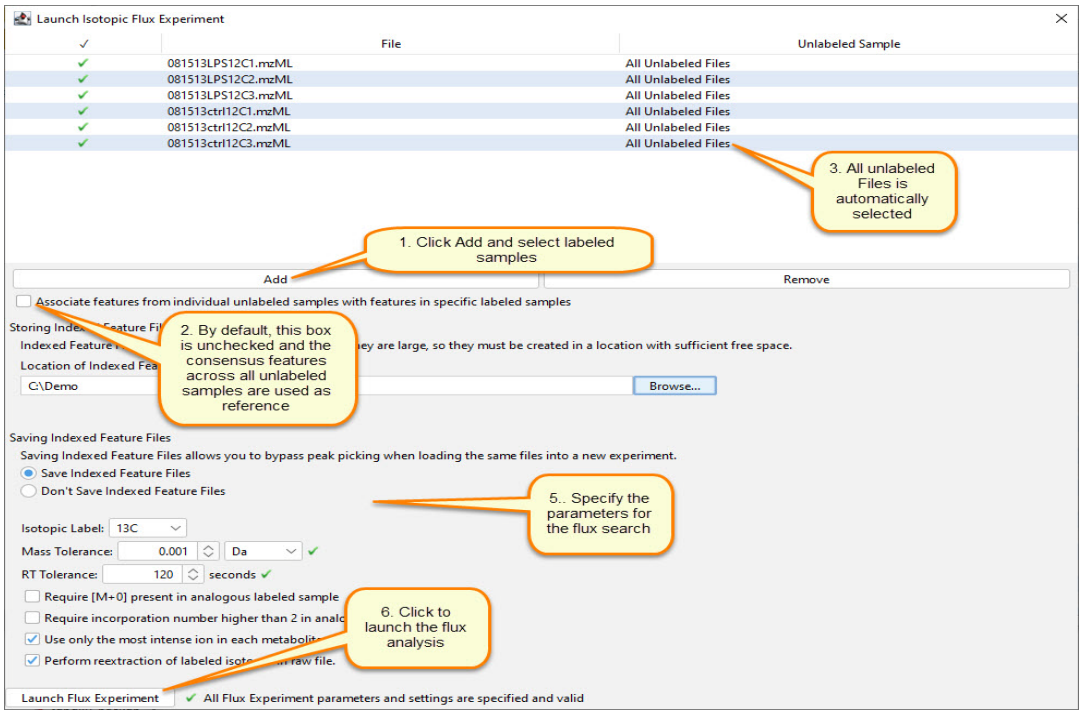
The first step is to click the Add button, which opens a file browser through which the user may select a labeled sample. Any number of labeled samples may be added in this way. The Remove button allows the user to correct any mistakes in labeled sample file selection.

When the labeled samples have been selected, their names will appear in the File column at the top of the dialog.

There are two options in matching labeled samples to their reference samples:

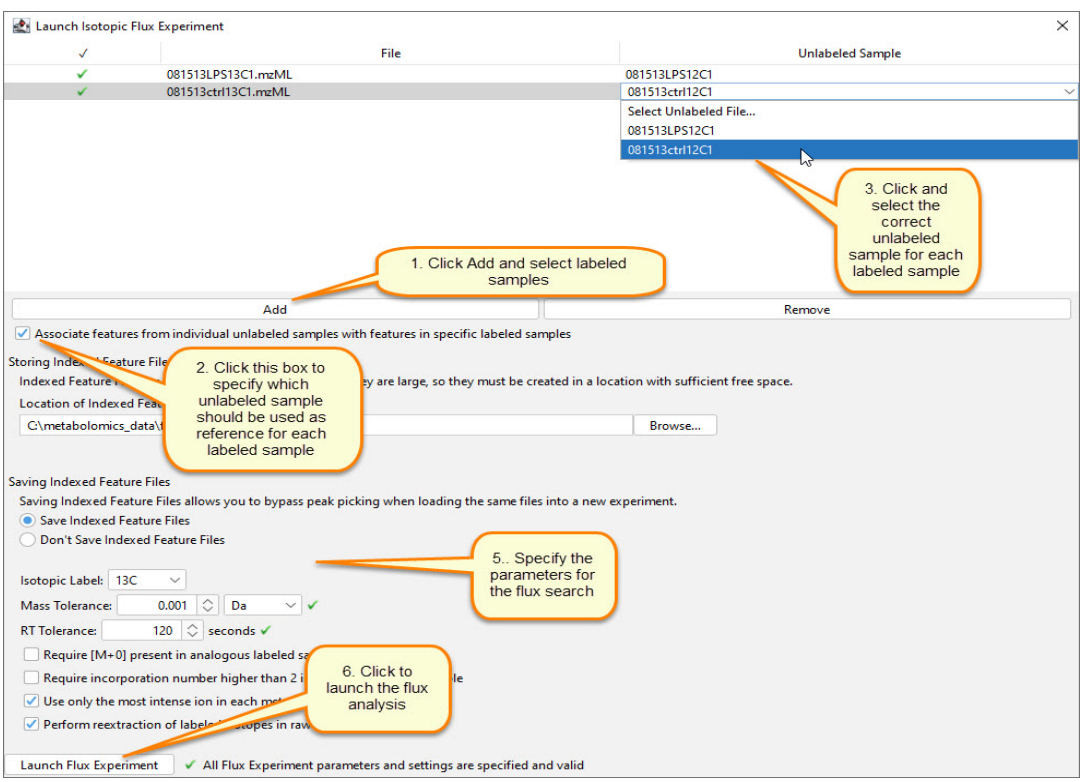
- By default, the consensus features derived from all unlabeled samples in the experiment are used as the reference for all labeled samples.

Figure 12-1: The Isotopic Flux Dialog - Use all unlabeled samples as reference



- Checking a box allows association of labeled samples with specific unlabeled samples.

Figure 12-2: The Isotopic Flux Dialog - Associate specific unlabeled samples as references



If the unlabeled samples are expected to contain different sets of metabolites, checking the box labeled “Associate features from individual unlabeled samples with features in specific labeled samples” allows the user to specify which unlabeled sample should be matched with each labeled sample.

A yellow triangle appears in the Unlabeled Sample column beside each file name. This is a warning to the user that it is necessary to select an unlabeled sample from the dropdown to provide the metabolite information for the labeled sample on the left. It is permissible to select the same unlabeled sample for more than one labeled sample.

The lower portion of the dialog contains parameters which much be specifiedt.

Storing Indexed Feature Files

As in a traditional search, Scaffold Elements must reformat files and perform feature extraction, which requires creation of some potentially large intermediate files. The user should select a location for these files using the provided file browser.

Saving Indexed Feature Files

Radio buttons allow the user to specify whether or not the intermediate files created from the labeled samples should be retained for possible future analysis. If the user chooses “Don’t Save Indexed Feature Files” they will be deleted when the analysis has completed.

Isotopic Label

Specifies the isotopic label used in the experiment. Choices are 13C, 15N and 18O or D (deuterium).

Mass Tolerance

Specifies the mass tolerance and units (Da or ppm) to be used in matching the metabolites in the labeled and unlabeled samples.

RT Tolerance

Specifies the retention time tolerance in seconds to be used in matching the metabolites in the labeled and unlabeled samples.

Options

The following options control which metabolites are reported in the export. They may be set by checking or unchecking the boxes:

Require [M+0] present in analogous labeled sample - filters out any metabolites for which the [M+0] peak is not detected.

Require incorporation number higher than 2 in analogous labeled sample - filters out any metabolites for which no higher levels of label incorporation are observed.

Use only the most intense ion in each metabolite

Launch Flux Experiment

This button begins the analysis of the labeled samples and the creation of the Flux Report. When the export is complete, a message is displayed.

Flux Analysis in Scaffold Elements

Scaffold Elements obtains the location of each analyte from an unlabeled sample and then examines the same region in the corresponding labeled sample. The possible locations of all isotopic peaks which might result from incorporation of the label into the analyte. The program looks for isotopic peaks in the predicted locations, and calculates the intensity values for all such peaks which are found.

The Flux Report

The results of flux analysis are reported in a CSV file, which may be examined and further analyzed in Excel.

Figure 12-3: A Metabolite in the Flux Report

67	Metabolite Information					Sample Information			Intensity Information									
68	Metabolite Name	Accession Number	Molecular For	Retention Time (r	Precursor	Label:	Max Ir	Sample Name	Total Intensity	Ion Intensities								
1505	Raw Intensity Information								Total	[M+0]	[M+1]	[M+2]	[M+3]	[M+4]	[M+5]	[M+6]	[M+7]	[M+8]	[M+9]
1506	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+			081513LPS12C1	3186760.8	2540280	515275.3	131205.7							
1507	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+			081513LPS12C2	3651870.5	2907647	585563.6	158660.5							
1508	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+			081513LPS12C3	5899628.5	4705642	944862.8	249123.3							
1509	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+			081513ctrl12C1	3599075.2	2874827	570870.6	153377.8							
1510	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+			081513ctrl12C2	2650184.5	2106443	432518.1	111223.2							
1511	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+			081513ctrl12C3	3210960.2	2548239	525338.9	137382.2							
1512	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+	13C	15	081513LPS13C1.mzML	4490747.5	3422700	695568.1					313299.4	59179.79		
1513	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+	13C	15	081513LPS13C2.mzML	4663745.5	3453675	705008.8					308960.1	58168.12	137934.2	
1514	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+	13C	15	081513LPS13C3.mzML	4116171	3127090	638179					294750.3	56151.52		
1515	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+	13C	15	081513ctrl13C1.mzML	2244854.2	2165883					41658.67		37312.45		
1516	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+	13C	15	081513ctrl13C2.mzML	3057094.2	2224663	456697.9				42831.54	209280	34737.45	88884.37	
1517	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+	13C	15	081513ctrl13C3.mzML	3230833.8	2345992	482087.5				40354.29	221884.1	38498.5	102017.2	
1518	[M+0] Normalized Intensity Information									[M+0]	[M+1]	[M+2]	[M+3]	[M+4]	[M+5]	[M+6]	[M+7]	[M+8]	[M+9]
1519	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+			Theoretical Distribution		1	0.1667	0.0295							
1520	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+			081513LPS12C1		1	0.2028	0.0517							
1521	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+			081513LPS12C2		1	0.2014	0.0546							
1522	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+			081513LPS12C3		1	0.2008	0.0529							
1523	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+			081513ctrl12C1		1	0.1986	0.0534							
1524	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+			081513ctrl12C2		1	0.2053	0.0528							
1525	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+			081513ctrl12C3		1	0.2062	0.0539							
1526	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+	13C	15	081513LPS13C1.mzML		1	0.2032					0.0915	0.0173		
1527	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+	13C	15	081513LPS13C2.mzML		1	0.2041					0.0895	0.0168	0.0399	
1528	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+	13C	15	081513LPS13C3.mzML		1	0.2041					0.0943	0.018		
1529	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+	13C	15	081513ctrl13C1.mzML		1					0.0192		0.0172		
1530	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+	13C	15	081513ctrl13C2.mzML		1	0.2053				0.0193	0.0941	0.0156	0.04	
1531	Dihydromyricetin	CASNO:27200-12-0	C15H12O8	39.66834878	[606.07]+	13C	15	081513ctrl13C3.mzML		1	0.2055				0.0172	0.0946	0.0164	0.0435	

-

Chapter 13
Quantitative Methods and Tests

Scaffold Elements supports quantitative methods based on MS1 precursor intensity measurements. Values are summarized to the level specified in the summarization hierarchy, and a variety of quantitative statistical tests including the T-Test, ANOVA and Permutation Test are available to identify differential expression across factor levels or treatments.

Details about quantitation are provided in:

- "Quantitative Method" on page 151
- "Quantitative tests" on page 156

Quantitative Method

Elements performs relative quantitation by measuring the signal intensity of the features that represent an analyte within each MS sample, and comparing these intensities across MS samples. Scaffold Elements offers the option to normalize precursor intensity values across MS samples by two different methods. It also calculates fold changes at different summarization levels to support complex experimental designs.

Computing Precursor Intensities

When a Liquid Chromatography - Mass Spectrometry experiment is run, analytes elute from the LC column at different Retention Times (RTs). A specific compound, or analyte, elutes within a certain RT range. The compound is then ionized, forming one or more ion forms. As a result, the ion forms derived from a single analyte manifest as features (peaks) having approximately the same RT but different m/z values.

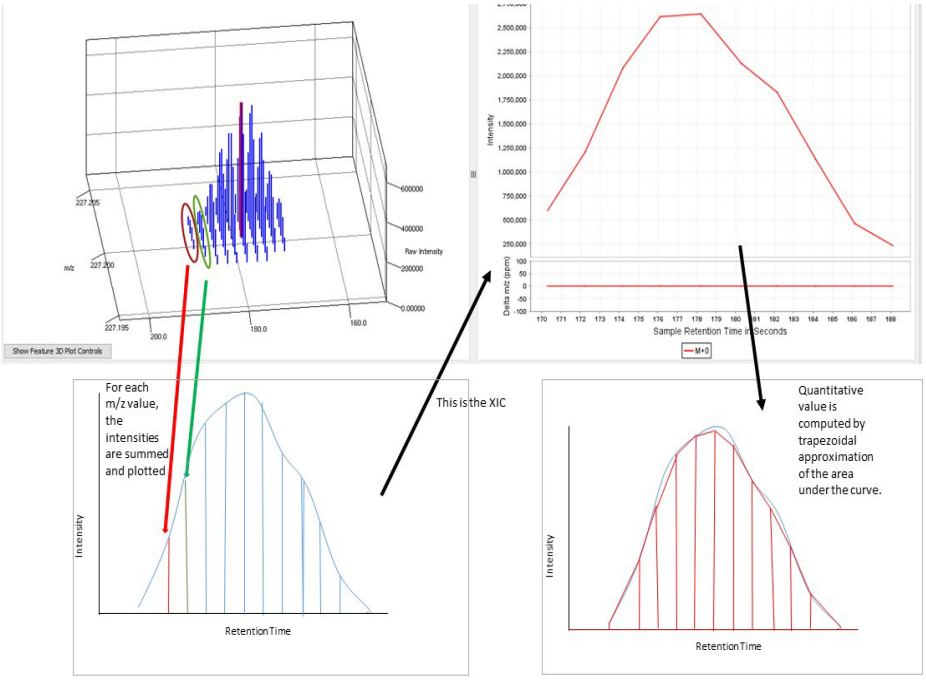
Each ion feature may be viewed as a 3-dimensional plot of (m/z, RT, intensity) points where the m/z bounds are determined by the mass accuracy of the instrument and the RT bounds by the elution time from the column.

In order to estimate the total intensity represented by the feature, a 2-dimensional extracted ion chromatogram (XIC) may be formed by summing the intensities of all raw data points with the same RT, and plotting the summed intensities versus RT. The trapezoidal approximation of the area under this XIC curve is the precursor intensity of the feature, see Figure 13-1

Precursor Intensity Quantitation assumes that for a feature detected at m/z and RT, the area under the corresponding XIC provides a measure of its relative abundance within the sample that eluted at RT. If additional features at the same RT are identified within an MS sample, their precursor intensities are summed to give a measure of the quantity of the analyte that was contained in the sample.

-

Figure 13-1: Calculation of the Precursor Intensity of a Feature



Normalization methods

Once features have been identified and quantified, Elements provides the option to apply normalization to the precursor intensity values.

Normalization is performed at the technical replicate level, which is the MS Sample level unless the experiment contains fractions. If one or more levels of fractions are specified, the values from the MS-Samples within each technical replicate are summed to provide the quantitative values at the technical replicate level.

Elements offers two types of normalizations:

- Sample Intensity Distribution Normalization
- Internal Standard Normalization

Sample Intensity Distribution Normalization

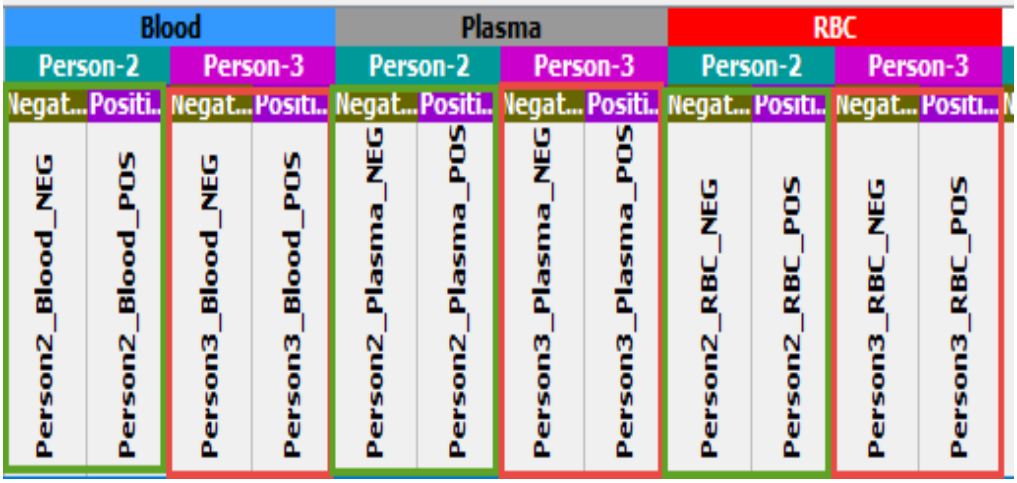
Scaffold Elements normalizes the quantitative values of all of the features in the loaded experiment using the following scheme:

1) If a level above the MS-Sample level is designated as the technical replicate level (i.e. the MS Samples are fractions), for each analyte, the intensity for each technical replicate is calculated by summing the precursor intensities in all MS-Samples within that technical replicate group.

-

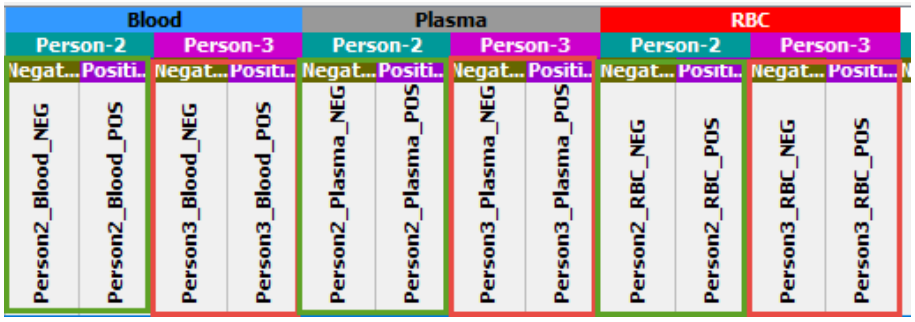
For example, suppose that each biological sample is divided and run in positive and in negative mode, giving two MS-Samples for each technical replicate, as shown in Figure 13-2.

Figure 13-2: Technical Replicates contain fractions



The set of analytes contained in a sample consists of a combination of those that ionize in positive mode and those that ionize in negative mode. Furthermore, for those analytes that ionize in both modes, the quantity of the substance in the sample is best estimated by the total precursor intensity of both positive and negative ions, so Elements sums the intensities of the two MS-Samples that represent positive and negative runs of a single sample, as shown in Figure 2.

2) *MS-Samples combined to form Technical Replicate*Scaffold Elements builds



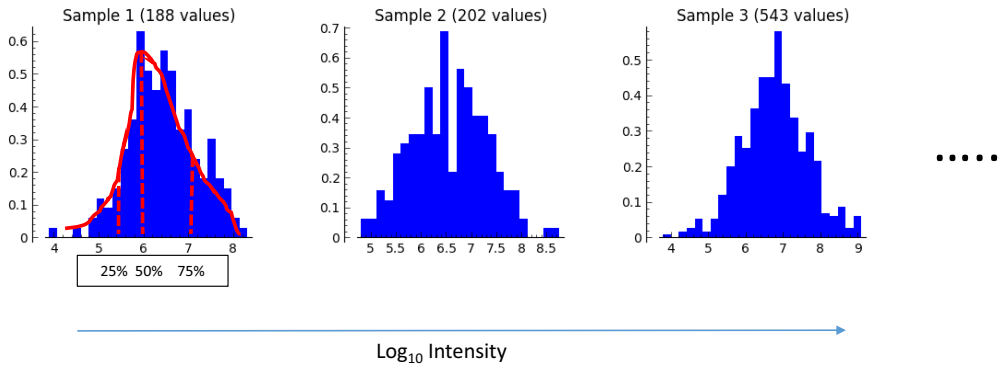
a distribution of the $\log_{10}$ of the Precursor Intensities of all features in each Technical Replicate and applies the procedure described in Normalization procedure.

Normalization procedure

For each histogram, the three quartiles, 25%, 50% and 75% are calculated. The median value of each quartile is computed across all Technical Replicates in the experiment. Then an overall distribution is defined using these median quartiles.

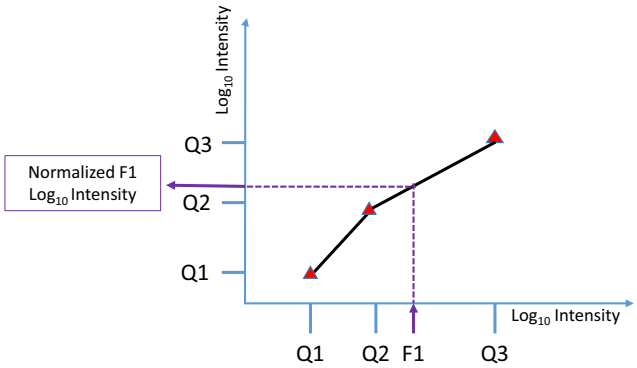
-

Figure 13-3: Log10Intensity distributions for each MS sample



For each Technical Replicate, the sample quartiles are plotted against the overall distribution quartiles, an interpolation is calculated among the plotted points and the Log_{10} Intensities are adjusted accordingly, see Figure 13-4.

Figure 13-4: Normalization plot

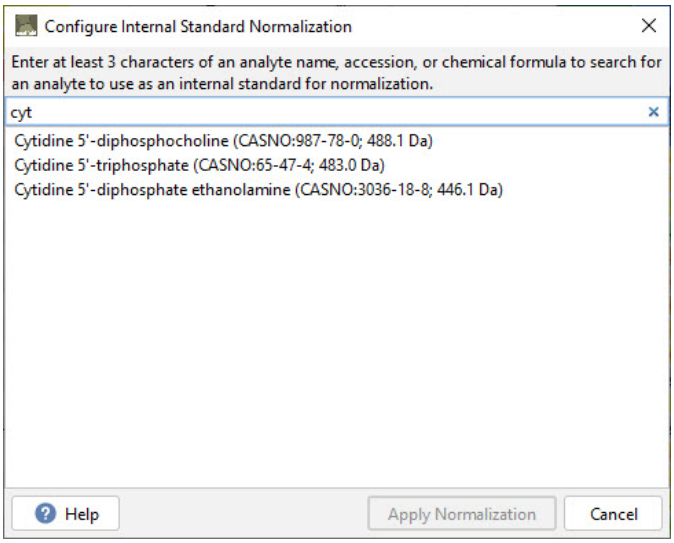


Internal Standard Normalization

In some experiments, a known quantity of a specific analyte is included in each sample and may be used as a standard to normalize quantities across the samples. Normalization to an Internal Standard allows the user to designate a particular analyte as such a standard. The selected analyte must appear in the Analytes List and must be identified in all samples.

The internal analyte is selected through the **Configure Internal Standard Normalization** dialog by selecting the menu option **Experiment > Normalize to Internal Standard**.

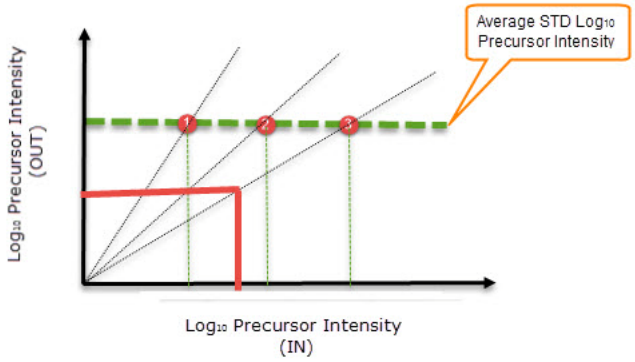
Figure 13-5: Configure Internal Standard Normalization dialog



The dialog contains a search tool that helps the user select the analyte to be used as the standard. Note that if the selected analyte does not appear in all samples or level of summarization, an Error dialog appears warning the user that it is not a suitable internal standard.

In this normalization technique, the normalized values are computed by setting the quantitative value of the standard analyte for each Technical Replicate to the average of the Log_{10} Precursor Intensity in all samples. The normalized values for all other analytes are then calculated using a single linear fit between zero and the average of the standard for each Technical Replicate. See Figure 13-6 for an example showing an experiment with three samples.

Figure 13-6: Normalization by internal standard



-

Quantitative tests

Scaffold Elements provides several statistical tests to identify analytes that exhibit different abundances at the comparison level established in the summarization hierarchy. The experimental design and the number of replicates dictates the most appropriate test to use.

Tests are selected through the "Configure Quantitative Analysis dialog" on page 156 .

The tests are based upon the data that is being displayed in the Samples Table. Adjusting thresholds and filtering the data changes the number of analytes shown in the table and may affect the analytes which will be indicated as having significant abundance level changes.

Configure Quantitative Analysis dialog

Selecting the menu option ***Experiment > Quantitative Analysis*** opens the **Configure Quantitative Analysis** dialog. This dialog consists of two tabs: the **Sample Hierarchy Tab** which allows the user to specify the experimental design (For details, see "Specifying the Design of an Experiment" on page 87)and the Statistical Analysis Tab (see "Statistical Analysis Tab" on page 94)from which the user may choose a quantitative test to apply to the data appearing in the Samples Tableand to define the significance level for the selected test.

The user must first specify the type of experiment and assign categories their appropriate roles in the analysis through the Sample Hierarchy Tab, and then may select and configure a statistical test through the Statistical Analysis Tab.

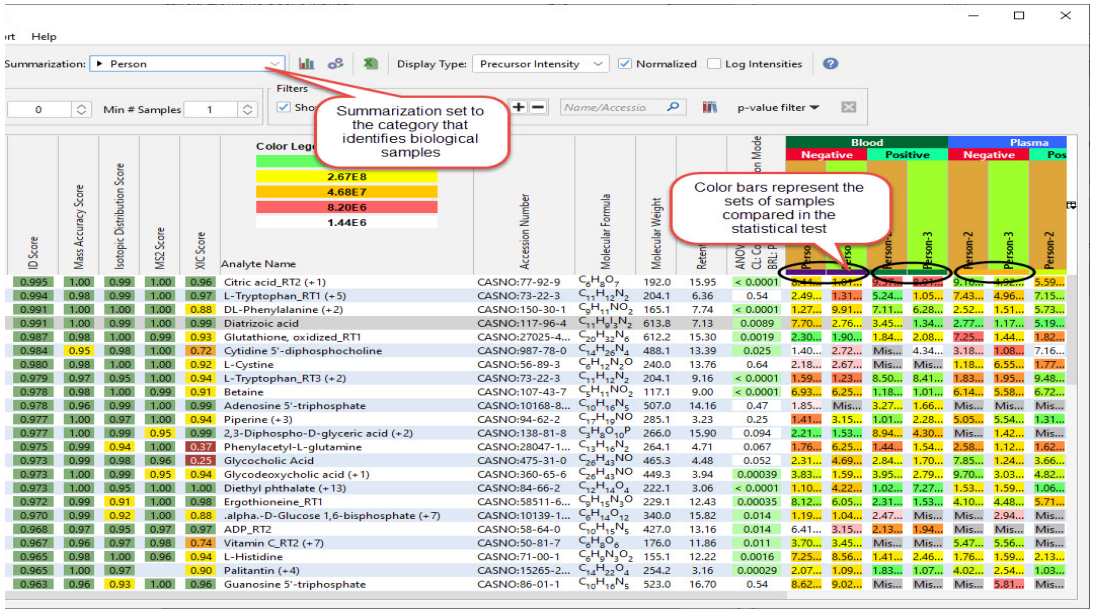
It is not necessary to use all factor levels or treatments in a specific test; only factor levels with a selected check box in the Design Matrix are used in computing the test. This can be useful if the user wants to exclude one or more treatments from the quantitative analysis. Sometimes this may be necessary in order to satisfy the constraints of the test. For example, the Two-Way and Repeated Measures tests require that the experiment be balanced. If one group has fewer samples than the others, it may be necessary to exclude that group from the analysis.

The user may also choose to apply a Multiple Test Correction, and must specify the desired significance level.

Once the experimental design has been specified, the quantitative test chosen, any multiple comparison correction has been selected and the significance threshold has been specified, clicking **Apply** starts the calculation. When the calculation is complete, a new column appears in the Samples Table showing the results of the selected quantitative test.

The heading of the added column lists the type of test applied and the comparison levels utilized. The p-values or q-values shown in the added column are highlighted in green if they are significant even with any multiple test correction, or yellow if they are significant only without the error correction applied. Values which are not significant under either condition are not highlighted.

Figure 13-7: Display of Statistical Test Results



When the selected summarization level corresponds to the chosen biological replicate level, (or blocking level), biological samples belonging to the same treatment group are tagged with a colored band. This helps the user recognize the groups of experimental units blocked together in the current test

Tests available for analyzing Basic Design experiments

•

ANOVA/t-test

ANOVA (Analysis of Variance) is an analysis method for testing equality of means across treatment groups or categories. It tells if there are differences among categories. The result of the test is a p-value which ,when low, indicates a large probability of variation among the different categories considered for the test. It does not, however, indicate which of the categories are different from each other.

Scaffold Elements supports a two-tailed version of ANOVA. When only two treatments are selected from the combinations of factor levels available, the two tailed ANOVA is equivalent to a T-Test.

Permutation Test

This test, depending on the selected treatments used to perform the test, establishes if there are statistically meaningful differences among multiple groups (equivalent to ANOVA) or differences between just two groups (equivalent to a T-test). Rather than depending on assumptions about the distributions of the values, however, it performs the experiment of randomly assigning the observed values to the various categories and assessing how rarely

-

the degree of difference between categories in the experiment is observed.

It is based on an F-statistic calculated on the original set of data points; the data points are then randomly permuted and a new F-statistics is calculated using the randomized values. This randomization and computation of an F-statistic is repeated 10000 times. Finally, a p-value is calculated by counting the number of times the randomized F-statistics were at least as large as the original F-statistics and dividing by 10000.

(See e.g. h*ttp://mathworld.wolfram.com/FishersExactTest.html*)

Mann-Whitney U Test

The Mann Whitney test is a non-parametric statistical hypothesis test for assessing whether one of two samples of independent observations tends to have larger values than the other. It can also be defined as a distribution-free test of whether two medians are equal. The test uses the ranks of the data in the two samples. Although the Mann Whitney test compares well with a t-test, it is independent of the way the data is distributed. Because the Mann Whitney test is the non-parametric version of the t-test, it requires exactly two quantitative sample categories to be selected for testing.

Kruskal-Wallis Test

The Kruskal-Wallis one-way analysis of variance by ranks is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing more three or more samples that are independent, or not related. (The parametric equivalence of the Kruskal-Wallis test is the one-way analysis of variance (ANOVA)). The factual null hypothesis is that the populations from which the samples originate have the same median. When the Kruskal-Wallis test leads to significant results, then at least one of the samples is different from the other samples. The test does not identify where the differences occur or how many differences actually occur. Because it is a non-parametric method, the Kruskal-Wallis test does not assume a normal distribution (unlike the analogous one-way analysis of variance); however, the test does assume an identically-shaped and scaled distribution for each group, except for any difference in medians.

Tests available for analyzing Repeated Measures experiments

Repeated Measures ANOVA/Paired t-test

The Repeated Measures Analysis of Variance test (rANOVA) is a parametric statistical hypothesis test for assessing whether the population means of repeated measurements differ when the measurements have been taken under at least three conditions (e.g., time points). The Paired t-test is similar, but compares exactly two samples. These tests may be used when the data being analyzed is normally distributed and has equal variances across the categories.

Wilcoxon Signed-rank Test

The Wilcoxon Signed-rank test is a nonparametric statistical hypothesis test for assessing whether the population mean ranks of repeated measurements differ when the measurements

-

have been taken under exactly two conditions (e.g., time points). It is a nonparametric alternative to the Paired T-Test, and may be used when the data being analyzed is not normally distributed. The Wilcoxon Signed-rank test does assume that the distributions in the two categories are independent and identically distributed.

Friedman Test

The Friedman test is a nonparametric statistical hypothesis test for assessing whether the population mean ranks of repeated measurements differ when the measurements have been taken under at least three conditions (e.g., time points). It is a nonparametric alternative to the Repeated Measures Analysis of Variance (rANOVA), and may be used when the data being analyzed is not normally distributed. The Friedman test does assume that the distributions in the categories are independent and identically distributed.

Tests available for analyzing Two-Way experiments

Two-way ANOVA

The Two-way ANOVA assesses the effect of two independent variables on analyte levels. It measures the main effect of each of the independent variables, as well as whether there is significant interaction between the variables.

Randomized Block ANOVA

This test is applicable to experiments with a Randomized Block Design. It assesses the effect of the Primary Analysis Category while controlling for the effect of the Secondary Analysis Category, which in this case is the blocking factor. It does not, however, assess the effect of the Secondary Analysis Category itself. An option is available, however, to assess the blocking effect, which assesses whether there is significant interaction between the Primary and Secondary Analysis Categories.

Friedman Test

The Friedman test is a nonparametric statistical hypothesis test for assessing whether the population mean ranks of repeated measurements differ when the measurements have been taken under at least three conditions (e.g., time points). It is a nonparametric alternative to the Repeated Measures Analysis of Variance (rANOVA), and may be used when the data being analyzed is not normally distributed. The Friedman test does assume that the distributions in the categories are independent and identically distributed.

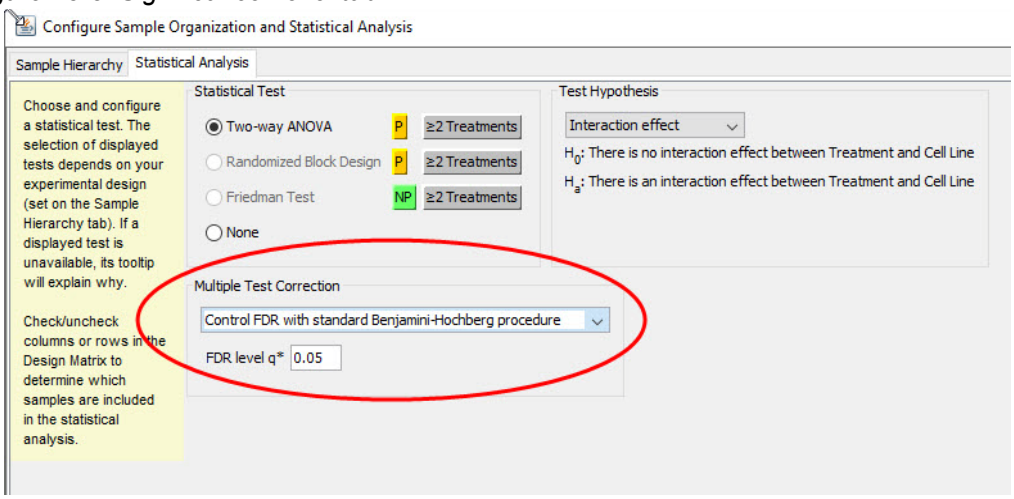
- *In choosing the statistical test to apply, it may be helpful to remember that log intensity values are more likely to be normally distributed while the intensities themselves are not. Parametric tests, therefore, are more suitable when analyzing log values, while it may be preferable to select a non-parametric test for intensities. Nonparametric tests are also better when the data contains outliers which may skew the results.*

-

Significance Level

Controls in the Statistical Analysis tab allow the user to set the significance level required for the selected inference test and to choose methods to control the family-wise error rate through a pull down list.

Figure 13-8: Significance Level tab



Multiple Test Correction

When considering a set of statistical inferences simultaneously and doing multiple comparisons, the risk of making one or more false discoveries (Type I error) grows quite quickly. In these cases it is common to adjust p-values for the number of hypothesis tests performed. A common method is to control the family-wise error rate, which is defined as the probability of making Type I errors. One of the initial and still quite common methods used to control this error is the Bonferroni correction where the significance level for an individual test is found by dividing the family-wise error rate (usually 0.05) by the number of performed tests. Thus when doing 100 statistical tests, the level for an individual test would be 0.05/100=0.0005, and only individual tests with P<0.0005 would be considered significant.

The Bonferroni approach is a fairly conservative one and for a very large number of independent comparisons it may lead to a high rate of false negatives.

To address this issue Scaffold Elements provides two different types of corrections:

- Control FWER with Hochberg's step-up and Holm's step-down
- Control FDR with standard Benjamini-Hochberg procedure

Control FWER with Hochberg's step-up and Holm's step-down

There are various methods described in the literature that control the Family-wise error rate (FWER) using less conservative corrections that are still based on the Bonferroni inequality. These methods are usually quite appropriate to control the FWER in experiments in which a

-

limited number of comparisons are of interest and where the use of the False Discovery Rate is inappropriate. In these cases, such corrections guard against false positives being reported.

Scaffold Elements offers an option to use the following methods to calculate the corrected significance level:

- Holm's step-down method
- Hochberg's step-up method

For more information about these methods, see Controlling the Family-wise Error Rate in the appendix.

When this option is selected the significance level is expressed in terms of o and the related text box appears underneath the pull down list. This text box allows the user to set the significance level to the desired value. The default value is 0.05.

Control FDR with standard Benjamini-Hochberg procedure

This method of controlling the error rate is particularly useful in high-dimensional experiments where a more common goal is to identify as many true positive findings as possible, while incurring a relatively low number of false positives. The false discovery rate (FDR) is designed to quantify this type of trade-off, making it particularly useful for performing many hypothesis tests on high-dimensional data sets.

Scaffold Elements computes the FDR using the Benjamini-Hochberg procedure as developed in the original paper[1].

When this option is selected the significance level is expressed in terms of the FDR level q* and the related text box appears underneath the pull down list. This text box allows the user to set the FDR level q* to the desired value. The default value is 0.05.

Test Hypothesis

This section displays a statement of the null hypothesis, H_0 and the alternative hypothesis, H_a that will be tested by the selected test.

In the case of a two-factor analysis, several different measures are computed. This section then contains a drop-down menu for selecting which test measure should be displayed. Options are:

- Interaction Effect - measures whether there is a statistically significant interaction between the two variables.
- Primary Factor Effect - measures whether the variable designated as the primary analysis category demonstrates a statistically significant effect on analyte level when controlling for the secondary variable.

1. Benjamini Y and Hochberg Y (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," Journal of the Royal Statistical Society, Series B (Methodological), Vol.57, No. 1: 289-300.

-

- Secondary Factor Effect - measures whether the variable designated as the secondary analysis category demonstrates a statistically significant effect on analyte level when controlling for the primary variable.

Design Matrix

The Design Matrix at the bottom of the Statistical Analysis tab is provided to help the user verify the test design or adjust it by adjusting which cells of the matrix should be included in the test. At the beginning of each row and column is a checkbox. If the box is checked, the cells in that row or column will be included in the test. If it is unchecked, they will not. This feature allows the user to, e.g. remove all but two columns to allow performance of a T-test and creation of a Volcano Plot or remove a group that contains an inconsistent number of samples to create a balanced experimental design for a two-way ANOVA.

Multiple Test Corrections

When considering a set of statistical inferences simultaneously and doing multiple comparisons the risk of making one or more false discoveries or a Type I error grows quite quickly. In these cases it is common to adjust p-values for the number of hypothesis tests performed. There are many different methods that provide a way to perform this adjustment. A common one is to control the family-wise error rate, which is defined as the probability of making Type I errors. One of the initial and still quite common methods used to control this error is provided by the Bonferroni correction where the significance level for an individual test is found by dividing the family-wise error rate (usually 0.05) by the number of performed tests. Thus when doing 100 statistical tests, the level for an individual test would be 0.05/100=0.0005, and only individual tests with P<0.0005 would be considered significant.

The Bonferroni approach is a fairly conservative one and for a very large number of independent comparisons it may lead to a high rate of false negatives.

To address this issue Scaffold Elements provides two different types of corrections:

- Control FWER with Hochberg’s step-up and Holm’s step-down
- Control FDR with standard Benjamini-Hochberg procedure

Control FWER with Hochberg’s step-up and Holm’s step-down

There are various methods described in the literature that control the Family-wise error rate (FWER) using less conservative corrections than the Bonferroni one but are still based on the Bonferroni inequality. These methods are usually quite appropriate to control the FWER in control trial experiments in which a limited number of comparisons are of interest and where the use of the False Discovery Rate is inappropriate. In these cases these type of corrections guard against any false positive occurring

When this option is selected Scaffold Elements uses the following methods to calculate the corrected significance level:

- Holm’s step-down method
- Hochberg’s step-up method

-

For more information on how the two methods are developed in Scaffold Elements see the appendix Techniques to Control the Family-wise Error Rate.

When this option is selected the significance level is expressed in terms of o and the related text box appears underneath the pull down list. This text box allows the User to set the significance level to the desired value. The default value is 0.05.

Control FDR with standard Benjamini-Hochberg procedure

This method of controlling the error rate in multiple experiments is particularly useful in high-dimensional type of experiments where a more common goal is to identify as many true positive findings as possible, while incurring a relatively low number of false positives. The false discovery rate (FDR) is designed to quantify this type of trade-off, making it particularly useful for performing many hypothesis tests on high-dimensional data sets.

Scaffold Elements computes the FDR using the Benjamini-Hochberg procedure as developed in the original paper[2].

When this option is selected the significance level is expressed in terms of the FDR level q* and the related text box appears underneath the pull down list. This text box allows the User to set the FDR level q* to the desired value. The default value is 0.05.

2. Benjamini Y and Hochberg Y (1995). “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” Journal of the Royal Statistical Society, Series B (Methodological), Vol.57, No. 1: 289-300.

-

-

Chapter 14
Reports

A variety of reports are available in Scaffold Elements to assist the user in interpreting and working with quantitative analysis data. The Current View may be exported using the right click context menu option Export> Export to Excel. It creates a CSV file corresponding to the information displayed in the table or graph from which it is selected. A number of programmed reports are also available through the Export option on the Scaffold Elements main menu. Each report is saved in a CSV format and may be opened in Excel.

The user cannot change the report format, but may select a different location in which to save the report. When the user saves an Excel report, a default name in the format <Report Name><Scaffold Elements File name> is provided for the report, but the name and location may be changed in the file browser. Finally, the user can open and view any Scaffold Elements report in Excel or another spreadsheet application, or using a text editor.

The following reports are available in Scaffold Elements:

- Export Attributes File...
- Export Samples Report to Excel...
- Export Features Report to Excel...
- Export Heatmap Report to Excel...
- Run SQL Query for Export...
- Export for MetaboAnalyst...
- Export Current View to Excel

-

Export Current View to Excel

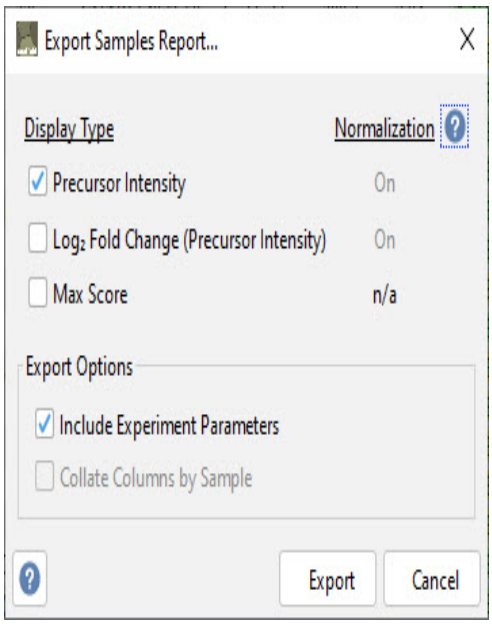
This is accomplished by right-clicking in any pane of any View and choosing Export>Export to Excel. Exports the information contained in the current view to a comma-delimited text file that can be opened and viewed in Excel.

Export Attributes File...

Generates an attributes file that captures the sample organization of the current experiment.

Export Samples Report to Excel...

Opens a dialog that allows the user to produce a customized analyte-level report.This option generates a comma-delimited Samples table similar to the one appearing in the Samples View, but allows the user to select whether or not to display each of the Display Types, whether to include a header specifying the experimental analysis parameters, and if more than one Display Type is selected, how to group the quantitative values. If “Collate Columns by Sample” is selected, the columns containing all quantitative values for a sample will be adjacent to each other, if it is not, all columns of a single Display Type for all samples will be adjacent, followed by columns for the next Display Type, etc.



Export Features Report to Excel...

Generates a comma-delimited table of all Features for all analytes appearing in the Samples View.

-

Export Heatmap Report to Excel...

Generates a comma-delimited file containing a header section detailing the parameter settings and filtering applied when the Heatmap was generated, along with the information contained in the Heatmap.

Export Publish Report to Excel...

Generates a comma-delimited report of the information contained in the Publish View Experiment Information table.

Run SQL Query for Export...

Opens the SQL Query tab of the Publish View, see

Export for MetaboAnalyst...

Generates a comma-delimited file formatted for import into the MetaboAnalyst online anlaysis tool.

-

Appendix

- Appendix A.Creating A Personal Spectral Library
- Appendix B.Creating a Custom Spectral Library using a tab-delimited text file
- Appendix C.Elements Scoring Algorithms
- Appendix D.Rolling up Values
- Appendix E.Agglomerative Point Clustering Feature Finding Algorithm
- Appendix F.Isotopic Clustering
- Appendix G.Forming Consensus MS1 peak groups
- Appendix H.Exporting a Transition List to Skyline
- Appendix I.Terminology
- Appendix J.Heat map clustering
- Appendix J.Techniques to Control the Family-wise Error Rate
- Appendix M.Using Principal Component Analysis in Scaffold Elements
- Appendix N.How PCA is Performed in Scaffold Elements
- Appendix O.Description of Mouse Right Click Context Menu Commands

Appendix A.Creating A Personal Spectral Library

Why create a personal spectral library?

A powerful feature of Elements is the ability to transform experimental results into a personal spectral library. A personal spectral library offers distinct advantages over commonly used libraries like NIST and METLIN, such as the ability to:

- **Remove instrument-specific effects.** Instrumentation can have an effect on the MS2 spectra. By comparing MS2 spectra captured with the same instrument, instrument-dependent variation is factored out.
- **Search by retention time and exact mass rather than by exact mass alone.** It is often challenging to compare RT information across instrumental setups, but if the chromatographic system does not change between runs, you can confidently use RT in your spectral library matching.
- **Capture unknown analytes.** Even if an analyte identification cannot be determined for a species of interest, it is now possible to search for the same species again. Unidentified species might be confidently associated with known species based on similar variations in precursor intensity across samples, or have MS2 spectra that could be analyzed by hand.
- **Narrow the set of hypotheses.** Searching analytes against a large spectral library can produce an intimidating number of analyte identification hypotheses. Using smaller, higher-confidence personal libraries produces a more manageable number of analyte identification hypotheses.

Getting Started

Three steps are required:

1) Load the data required to create a personal spectral library. For example, you may load a standards data set containing known compounds you wish to identify and quantify in subsequent experiments and perform a search against a general database containing these compounds.

2) Use Elements' spectral library creation tool to generate a spectral library from the results.

3) Search experimental data against the new library.

Following is a tutorial demonstrating the creation and use of personal spectral libraries in Elements.

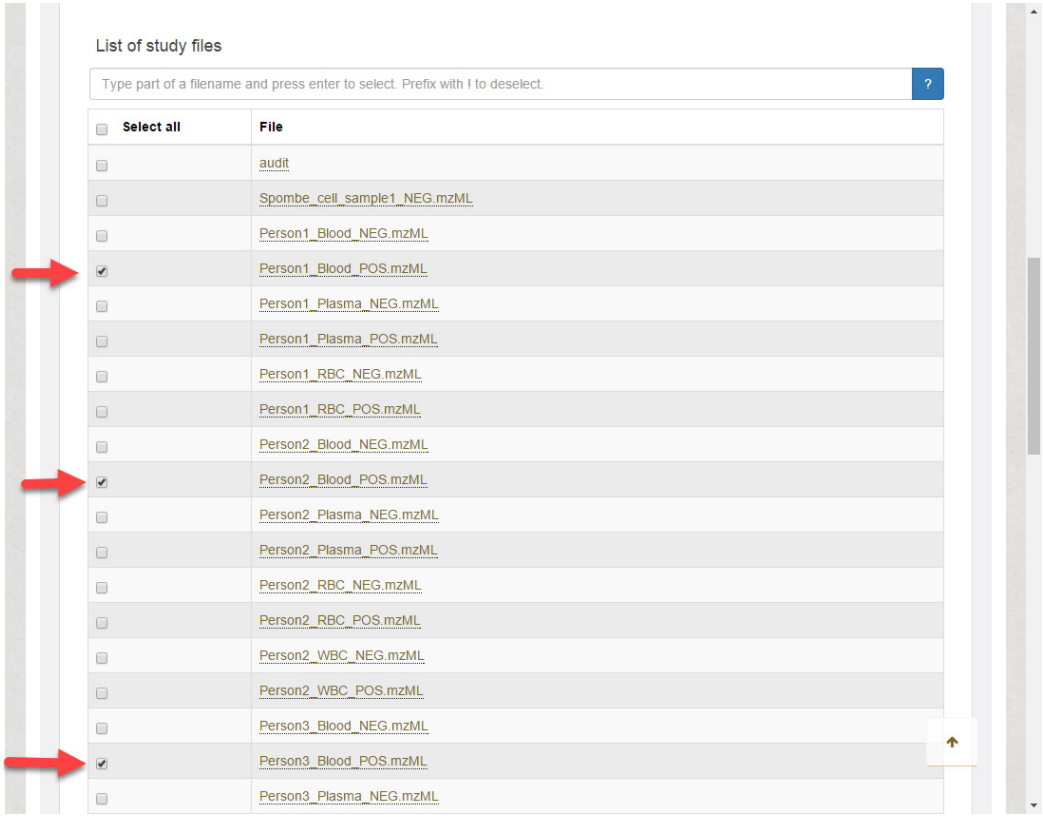
Personal Spectral Library Tutorial

In this short tutorial, we will build a personal spectral library from some publicly available metabolomics data from the MetaboLights webserver. You may want to use this tutorial as a guide while working with your own data, or if you wish to follow along, please first download the data available on this web page:

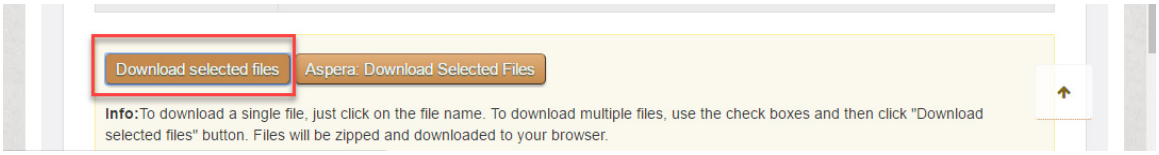
http://www.ebi.ac.uk/metabolights/MTBLS87

Instructions for Downloading Tutorial Data

On the web page, scroll down to the "List of Study Files". In the checkbox, select the files "Person1_Blood_POS.mzML", "Person2_Blood_POS.mzML", and "Person3_Blood_POS.mzML".



Scroll down to the bottom of the page and click the "Download selected files" button.



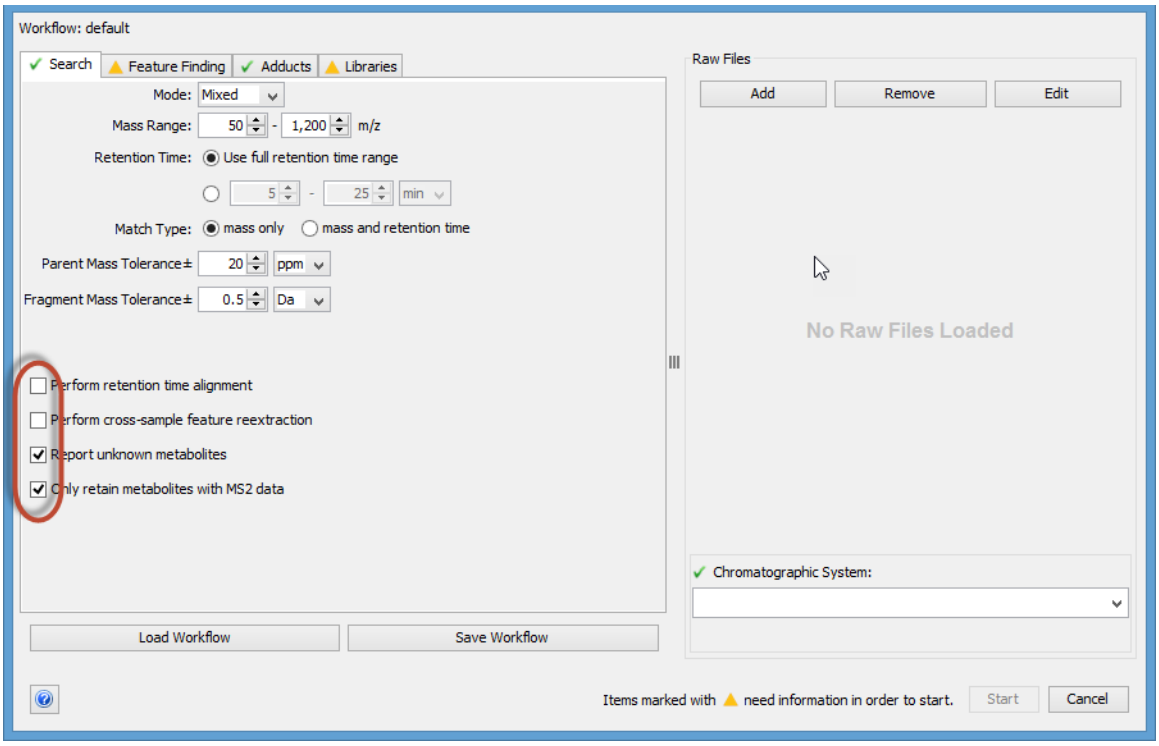
This will start a download of the files, into a zipped directory. Unzip the files into a directory on your computer.

Step 1: Load the data required to create a spectral library.

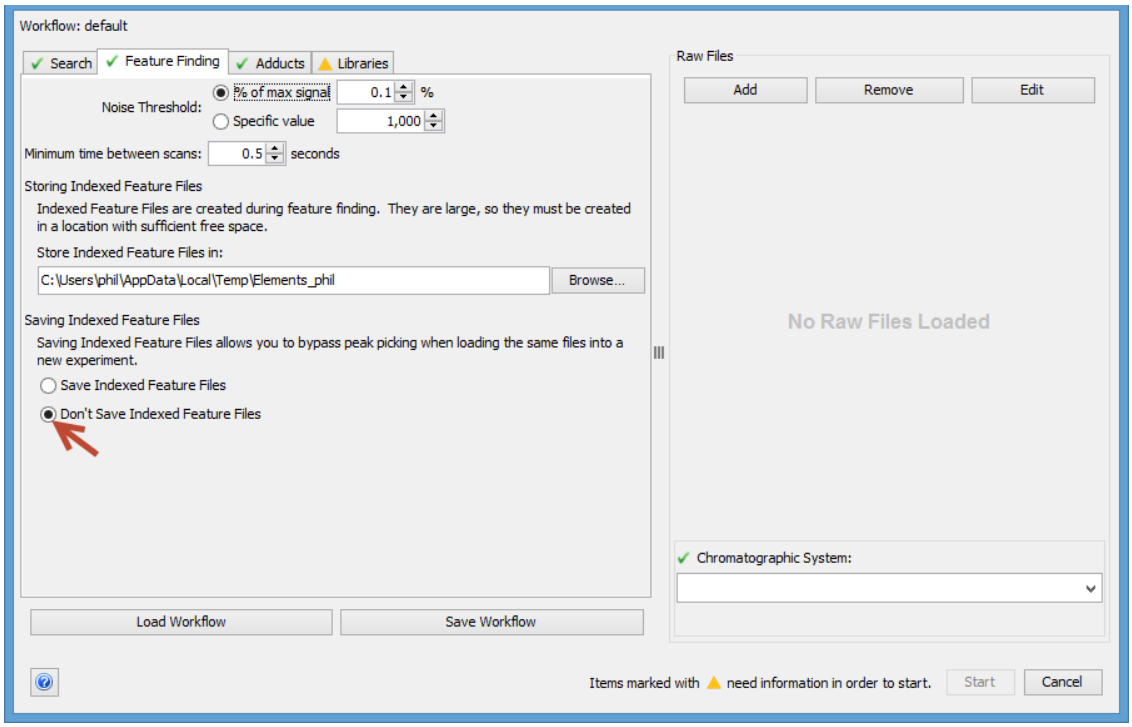
Searching:

Certain search settings are recommended in order to create a high quality spectral library. To apply these settings, we create a new experiment and load our files with the following settings:
First, in the search tab, we will uncheck “perform retention time alignment” so that the analytes we identify and add to our personal spectral library will contain unaligned retention times. We will match to these unaligned retention time values in future searches. We will also uncheck “perform cross-sample feature re-extraction” because would like to generate analyte identifications based exclusively on high-quality features, and re-extracted features may be lower in quality. We will check “Report unknown analytes”, so that unidentified ions can be added to the library. We will also check “Only retain analytes with MS2 data”. If you have run your instrument in DDA mode, some features will not have an associated MS2, but without an MS2, it can be very hard to make a confident identification, as there may be many structural isomers with identical precursor mass and very different fragmentation patterns. We choose, therefore, to include only analytes with MS2 spectra in our library.

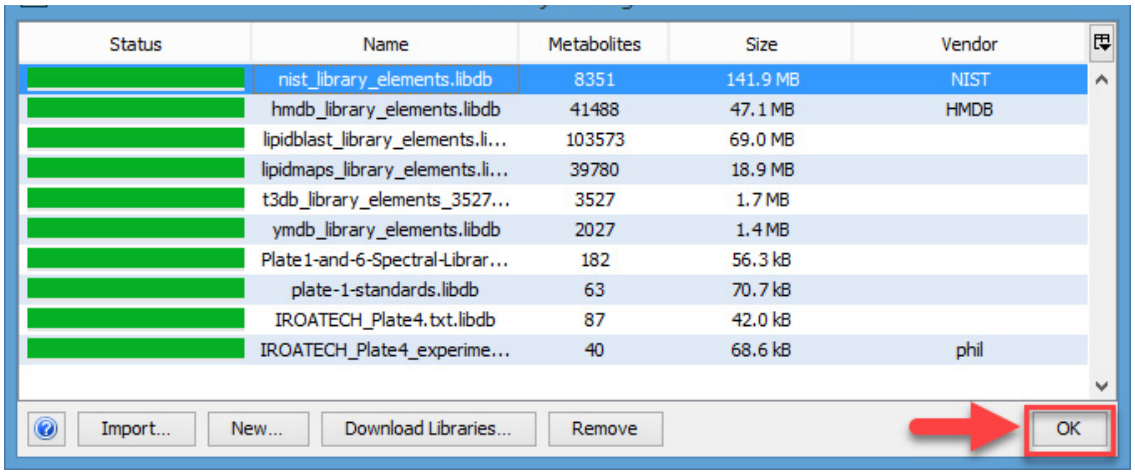
The dialog will look like this:



Let's move on to the Feature Finding tab. Under the "Saving Indexed Feature Files" section, click the "Don't Save Indexed Feature Files" radio button. Leave all other defaults.



In this tutorial, we can skip the "Adducts" tab, and move to the "Libraries" tab. Click the "Add Library…" button, and select the NIST library from the library manager dialog that appears. Once you have selected the appropriate library, click the "OK" button in the lower right-hand corner.



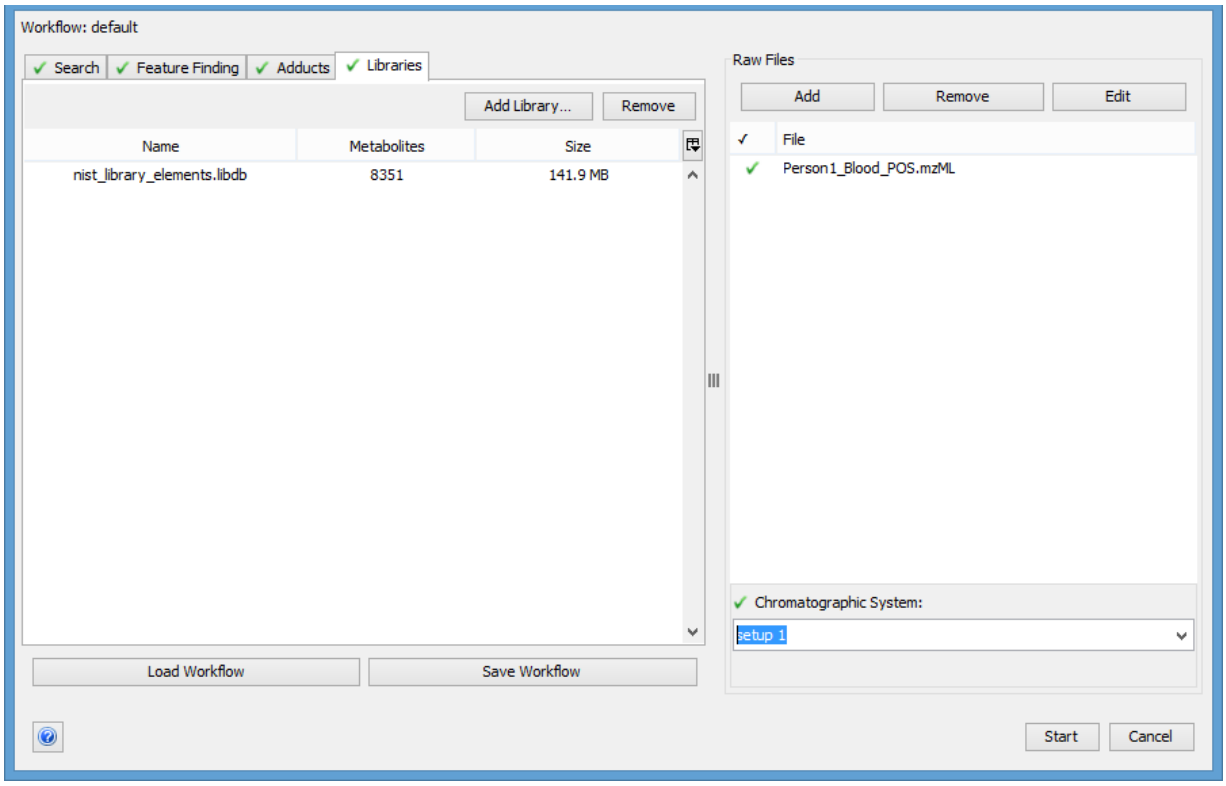
The NIST library will appear in the Library Tab.
Now we will add files to the experiment. In the "Raw Files" panel on the right

side of the loading dialog, click the “Add” button, and navigate to the folder where you have put the raw study files for this tutorial. Select only “Person1_Blood_POS.mzML”.

At this stage, all components are enabled, and it is possible to click the “Start” button. However, there is one more important field to fill out: The Chromatographic System, located directly beneath the raw files panel. It is essential that we enter this information because we will need it when we use our library in future experiments.

The Chromatographic System is an identifier to allow the user to track all of the information needed to describe the chromatography associated with the files loaded. This may include the mobile phase, stationary phase, filter size, column length, and pore size, just to name a few important factors. It is up to the user to keep track of all of the pertinent details associated with their chromatography, and assign a single string representation of their chromatographic system.

For this example, let us suppose that the details of the chromatography have been documented elsewhere, and we know that these files were run with “setup 1”. Type “setup 1” in the “Chromatographic System” panel.

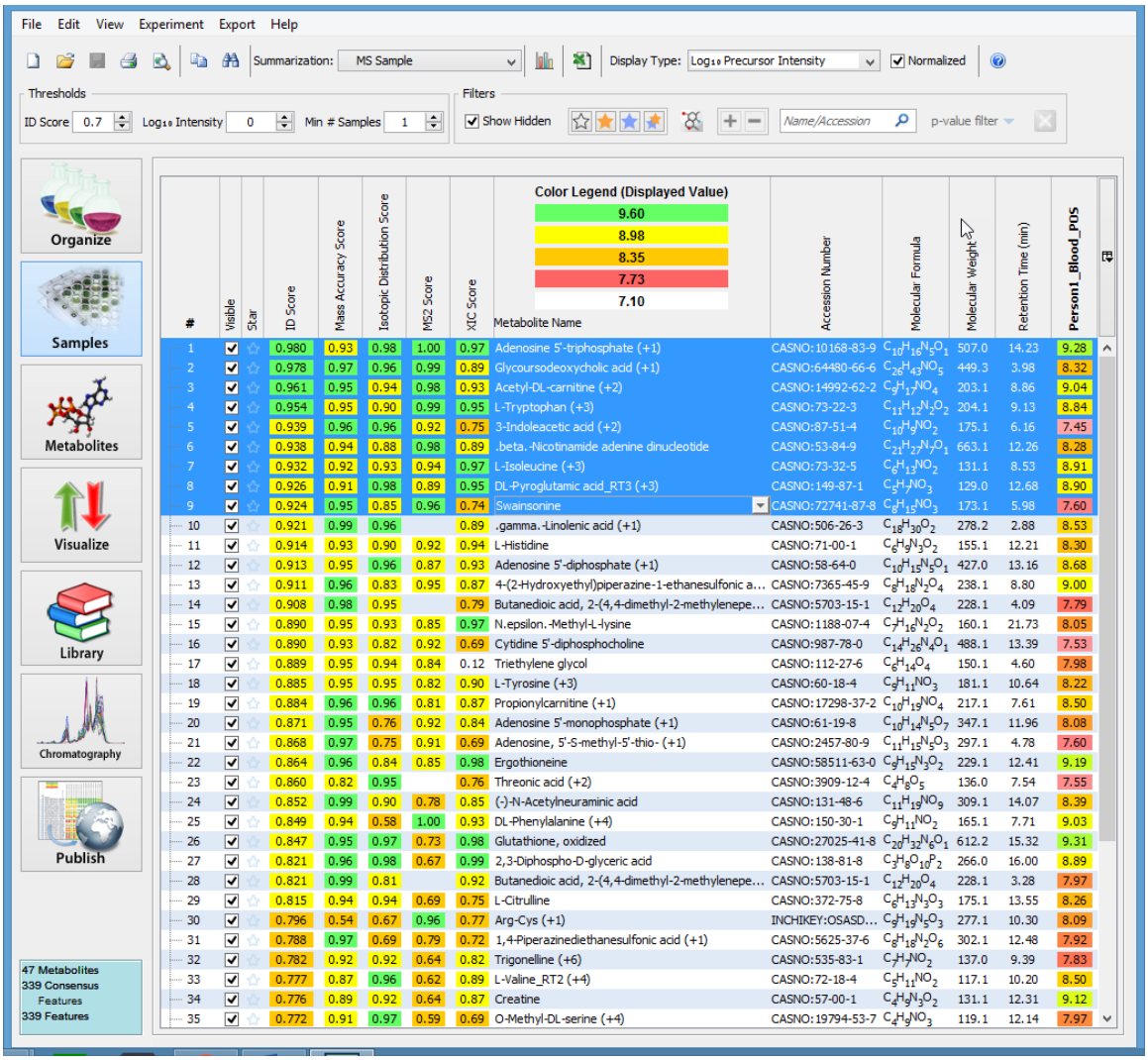


Now we are ready to load our data. Click the "Start" button in the lower left-hand corner of the Loading Dialog.

After a few minutes, the file will load, and the Elements Samples View will populate with analyte identifications. Save the file, and title it "MTBLS87-Person1-NIST".

Preparing to add a group of analytes to a personal spectral library

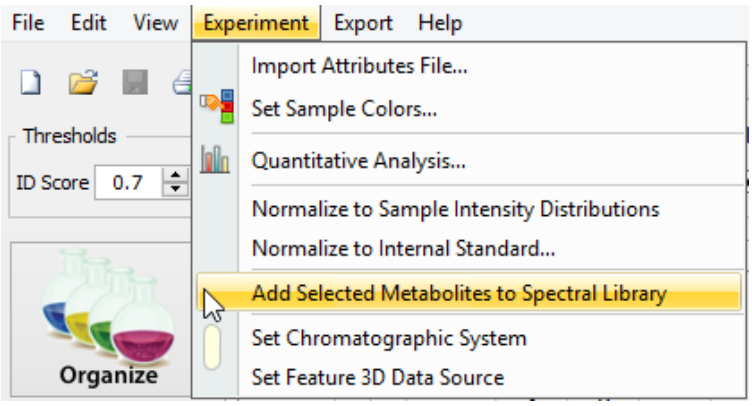
The first row in the table, "Adenosine 5'-triphosphate" is selected. Hold down the shift key and click the 9th row in the table, "Swainsonine".



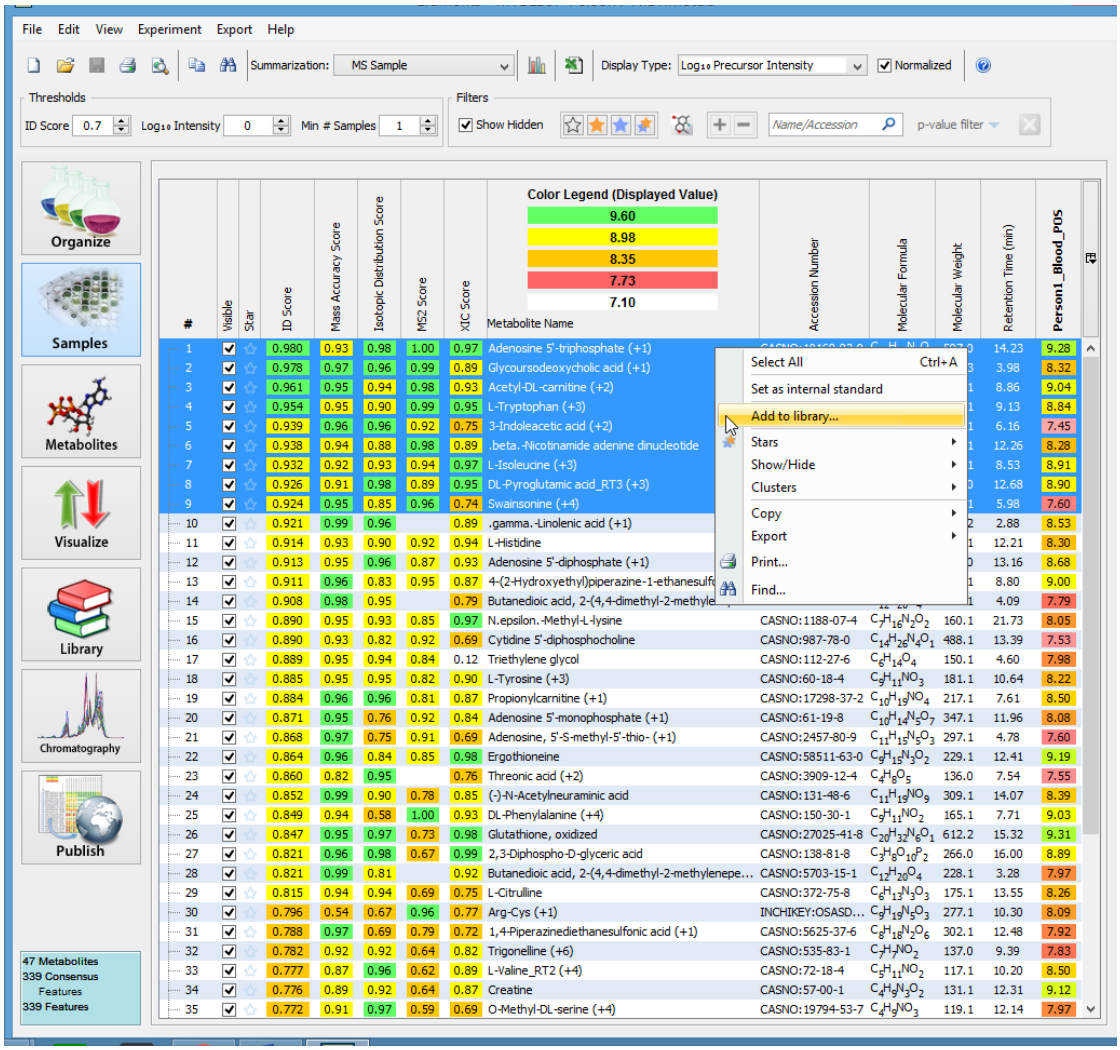
#	Visible	Star	ID Score	Mass Accuracy Score	Isotopic Distribution Score	MS2 Score	XIC Score	Metabolite Name	Accession Number	Molecular Formula	Molecular Weight	Retention Time (min)	Person1_Blood_POS
1	✓		0.980	0.93	0.98	1.00	0.97	Adenosine 5'-triphosphate (+1)	CASNO:10168-83-9	$C_{10}H_{16}N_5O_1$	507.0	14.23	9.28
2	✓		0.978	0.97	0.96	0.99	0.89	Glycoursodeoxycholic acid (+1)	CASNO:64480-66-6	$C_{26}H_{43}NO_5$	449.3	3.98	8.32
3	✓		0.961	0.95	0.94	0.98	0.93	Acetyl-DL-carnitine (+2)	CASNO:14992-62-2	$C_9H_{17}NO_4$	203.1	8.86	9.04
4	✓		0.954	0.95	0.90	0.99	0.95	L-Tryptophan (+3)	CASNO:73-22-3	$C_{11}H_{12}N_2O_2$	204.1	9.13	8.84
5	✓		0.939	0.96	0.96	0.92	0.75	3-Indoleacetic acid (+2)	CASNO:87-51-4	$C_{10}H_9NO_2$	175.1	6.16	7.45
6	✓		0.938	0.94	0.88	0.98	0.89	.beta.-Nicotinamide adenine dinucleotide	CASNO:53-84-9	$C_{21}H_{27}N_7O_1$	663.1	12.26	8.28
7	✓		0.932	0.92	0.93	0.94	0.97	L-Isoleucine (+3)	CASNO:73-32-5	$C_6H_{13}NO_2$	131.1	8.53	8.91
8	✓		0.926	0.91	0.98	0.89	0.95	DL-Pyroglutamic acid_RT3 (+3)	CASNO:149-87-1	$C_5H_7NO_3$	129.0	12.68	8.90
9	✓		0.924	0.95	0.85	0.96	0.74	Swainsonine	CASNO:72741-87-8	$C_8H_{15}NO_3$	173.1	5.98	7.60
10	✓		0.921	0.99	0.96		0.89	.gamma.-Linolenic acid (+1)	CASNO:506-26-3	$C_{18}H_{30}O_2$	278.2	2.88	8.53
11	✓		0.914	0.93	0.90	0.92	0.94	L-Histidine	CASNO:71-00-1	$C_6H_9N_3O_2$	155.1	12.21	8.30
12	✓		0.913	0.95	0.96	0.87	0.93	Adenosine 5'-diphosphate (+1)	CASNO:58-64-0	$C_{10}H_{15}N_5O_1$	427.0	13.16	8.68
13	✓		0.911	0.96	0.83	0.95	0.87	4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic a...	CASNO:7365-45-9	$C_8H_{18}N_2O_4$	238.1	8.80	9.00
14	✓		0.908	0.98	0.95		0.79	Butanedioic acid, 2-(4,4-dimethyl-2-methylenepe...	CASNO:5703-15-1	$C_{12}H_{20}O_4$	228.1	4.09	7.79
15	✓		0.890	0.95	0.93	0.85	0.97	N.epsilon.-Methyl-L-lysine	CASNO:1188-07-4	$C_7H_{16}N_2O_2$	160.1	21.73	8.05
16	✓		0.890	0.93	0.82	0.92	0.69	Cytidine 5'-diphosphocholine	CASNO:987-78-0	$C_{14}H_{26}N_4O_1$	488.1	13.39	7.53
17	✓		0.889	0.95	0.94	0.84	0.12	Triethylene glycol	CASNO:112-27-6	$C_6H_{14}O_4$	150.1	4.60	7.98
18	✓		0.885	0.95	0.95	0.82	0.90	L-Tyrosine (+3)	CASNO:60-18-4	$C_9H_{11}NO_3$	181.1	10.64	8.22
19	✓		0.884	0.96	0.96	0.81	0.87	Propionylcarnitine (+1)	CASNO:17298-37-2	$C_{10}H_{19}NO_4$	217.1	7.61	8.50
20	✓		0.871	0.95	0.76	0.92	0.84	Adenosine 5'-monophosphate (+1)	CASNO:61-19-8	$C_{10}H_{14}N_5O_7$	347.1	11.96	8.08
21	✓		0.868	0.97	0.75	0.91	0.69	Adenosine, 5'-S-methyl-5'-thio- (+1)	CASNO:2457-80-9	$C_{11}H_{15}N_5O_3$	297.1	4.78	7.60
22	✓		0.864	0.96	0.84	0.85	0.98	Ergothioneine	CASNO:58511-63-0	$C_9H_{15}N_3O_2$	229.1	12.41	9.19
23	✓		0.860	0.82	0.95		0.76	Threonic acid (+2)	CASNO:3909-12-4	$C_4H_8O_5$	136.0	7.54	7.55
24	✓		0.852	0.99	0.90	0.78	0.85	(-)-N-Acetylneuraminic acid	CASNO:131-48-6	$C_{11}H_{19}NO_9$	309.1	14.07	8.39
25	✓		0.849	0.94	0.58	1.00	0.93	DL-Phenylalanine (+4)	CASNO:150-30-1	$C_9H_{11}NO_2$	165.1	7.71	9.03
26	✓		0.847	0.95	0.97	0.73	0.98	Glutathione, oxidized	CASNO:27025-41-8	$C_{20}H_{32}N_6O_1$	612.2	15.32	9.31
27	✓		0.821	0.96	0.98	0.67	0.99	2,3-Diphospho-D-glyceric acid	CASNO:138-81-8	$C_3H_8O_{10}P_2$	266.0	16.00	8.89
28	✓		0.821	0.99	0.81		0.92	Butanedioic acid, 2-(4,4-dimethyl-2-methylenepe...	CASNO:5703-15-1	$C_{12}H_{20}O_4$	228.1	3.28	7.97
29	✓		0.815	0.94	0.94	0.69	0.75	L-Citrulline	CASNO:372-75-8	$C_6H_{13}N_3O_3$	175.1	13.55	8.26
30	✓		0.796	0.54	0.67	0.96	0.77	Arg-Cys (+1)	INCHIKEY:OSASD...	$C_9H_{19}N_5O_3$	277.1	10.30	8.09
31	✓		0.788	0.97	0.69	0.79	0.72	1,4-Piperazinediethanesulfonic acid (+1)	CASNO:5625-37-6	$C_8H_{18}N_2O_6$	302.1	12.48	7.92
32	✓		0.782	0.92	0.92	0.64	0.82	Trigonelline (+6)	CASNO:535-83-1	$C_7H_7NO_2$	137.0	9.39	7.83
33	✓		0.777	0.87	0.96	0.62	0.89	L-Valine_RT2 (+4)	CASNO:72-18-4	$C_5H_{11}NO_2$	117.1	10.20	8.50
34	✓		0.776	0.89	0.92	0.64	0.87	Creatine	CASNO:57-00-1	$C_4H_9N_3O_2$	131.1	12.31	9.12
35	✓		0.772	0.91	0.97	0.59	0.69	O-Methyl-DL-serine (+4)	CASNO:19794-53-7	$C_4H_9NO_3$	119.1	12.14	7.97

Now, we will turn the information associated with these first nine analytes into a personal spectral library.

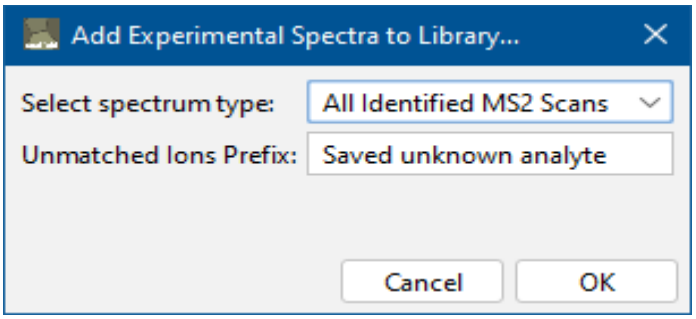
Keeping these nine analytes selected, navigate to the Experiment Menu, and select Experiment -> Add Selected Analytes to Spectral Library



Alternatively, you can also right click on the samples view while the 9 analytes are selected, and choose “Add to library…”



After this, the following dialog will appear:



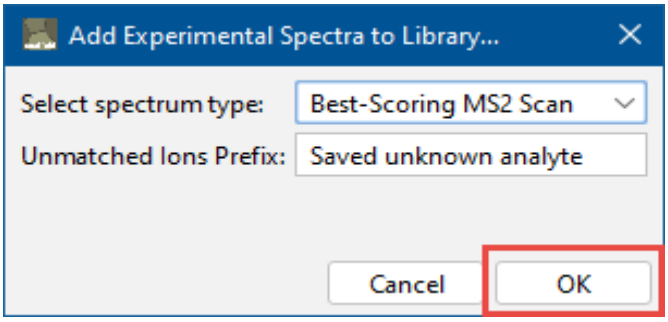
This dialog deserves some explanation:

The first option, “Select spectrum type:”, allows the user to specify which MS2 information gathered from the experiment should be written into a spectral library of interest. There are two possible choices for spectrum type:

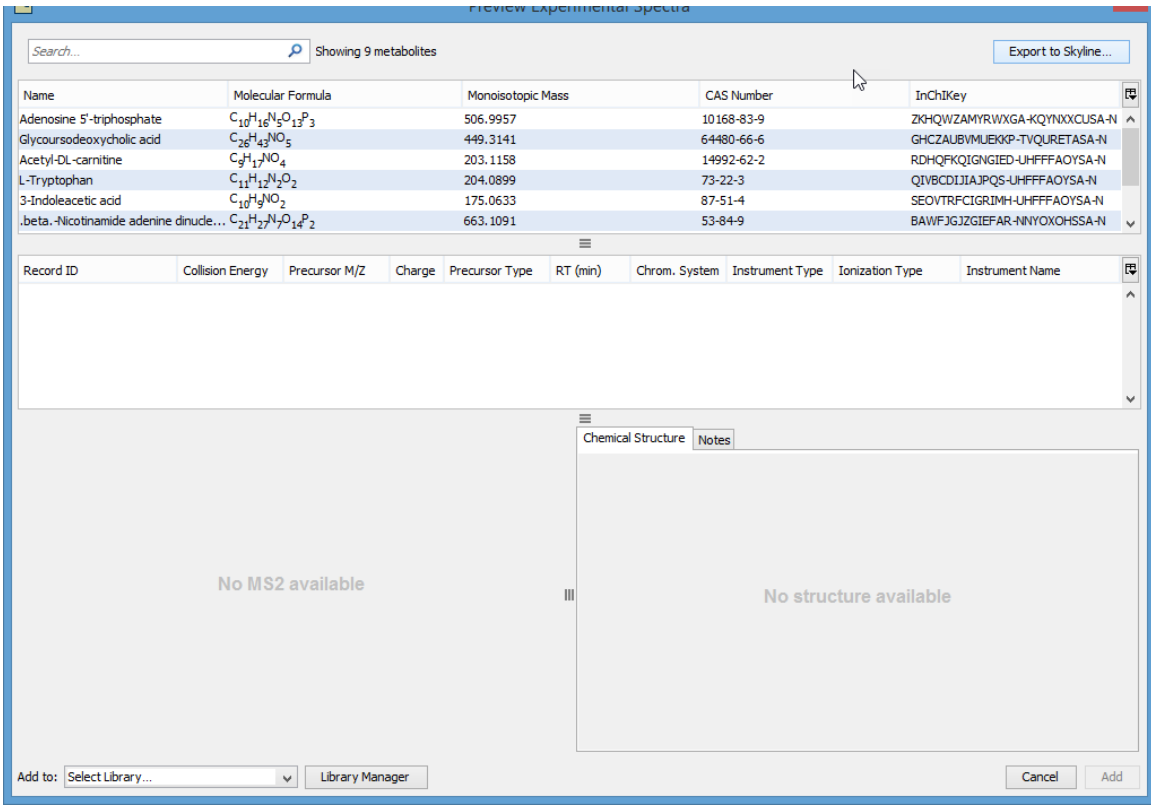
- “All Identified MS2 Scans”, which retains MS2 scans for all ion forms and features associated with each analyte. The highest scoring MS2 scan is retained for each feature.
- “Best-Scoring MS2 Scan”, which retains only the highest-scoring MS2 per analyte (across all ion forms and features discovered).

The second option, “Unmatched Ions Prefix”, is relevant when dealing with unknown analytes. We will return to this option later.

From the “Select spectrum type:” drop-down menu, select the second option, “Best-Scoring MS2 Scan.” Then, click the OK button in the lower right-hand corner.

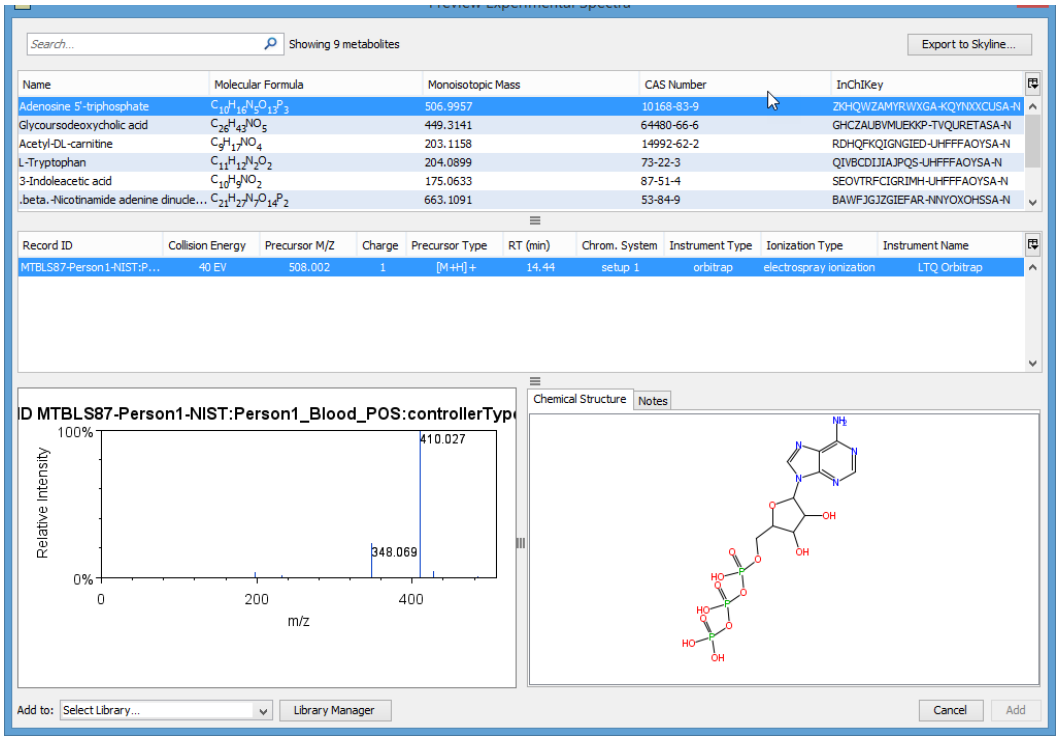


At this point the “Preview Experimental Spectra” dialog will appear.

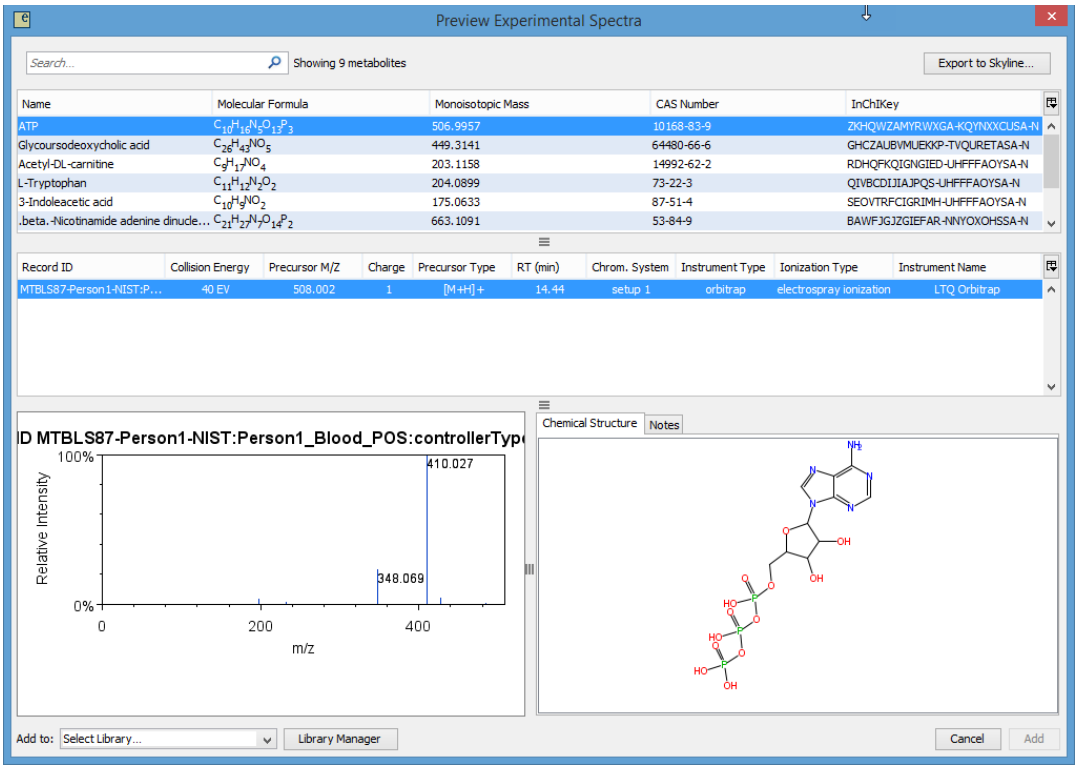


This dialog allows you to make adjustments to the data imported from Elements before it is written to a spectral library.

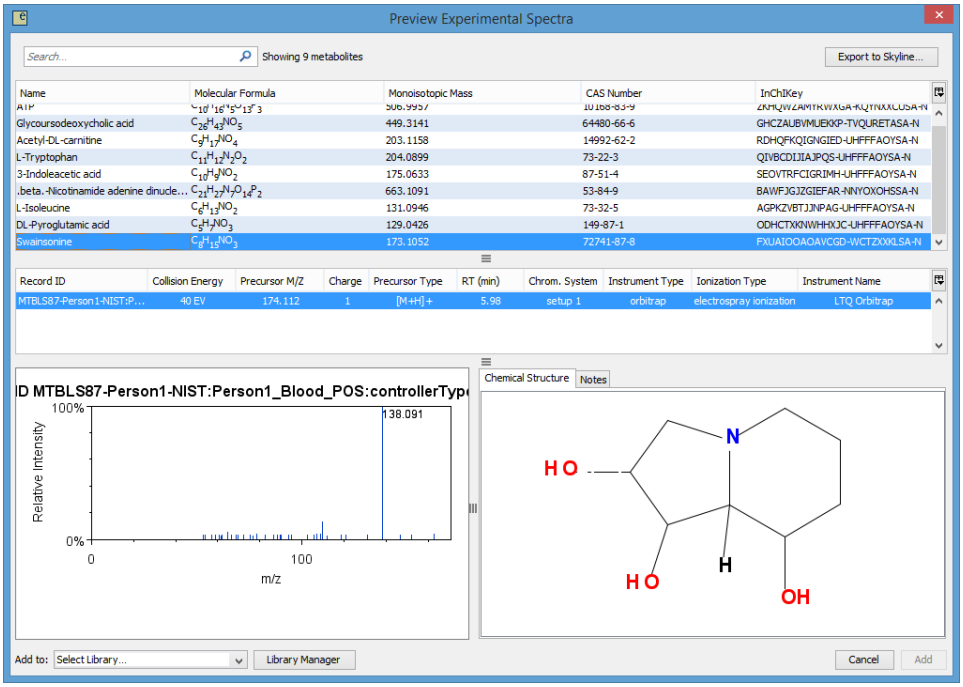
In the dialog, select the first analyte, “Adenosine 5’-triphosphate”. Note that a single record of this analyte is present, along with a single MS2 spectrum and a chemical structure.



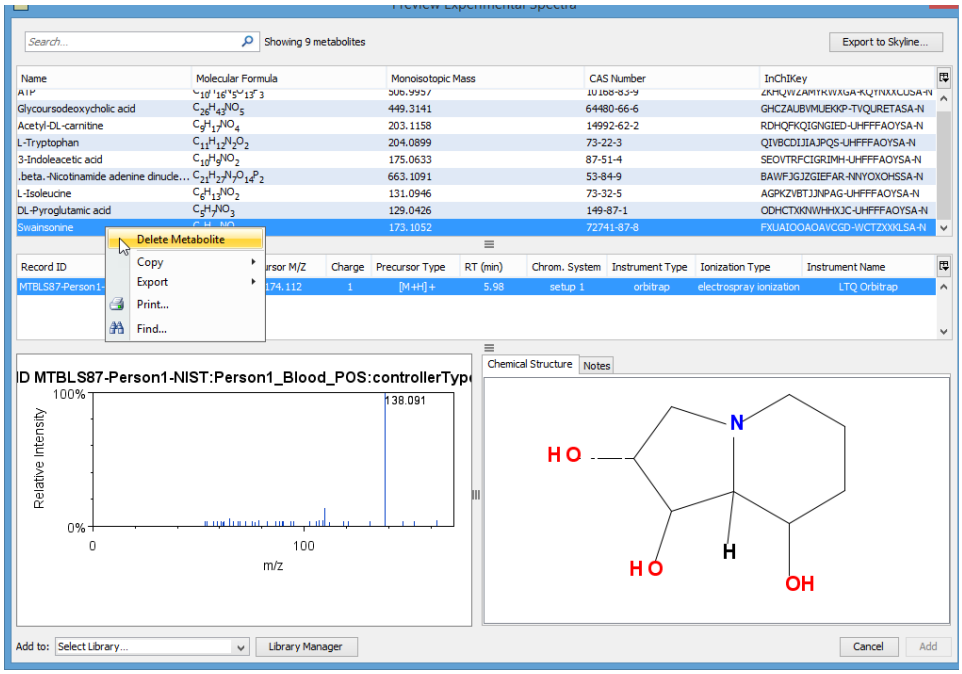
This dialog allows the user to edit many features of the analyte entries before they are added to the library. For example, double-click on the cell containing “Adenosine 5’-triphosphate”, and replace this text with “ATP”. Click the “enter” key to update the analyte name.



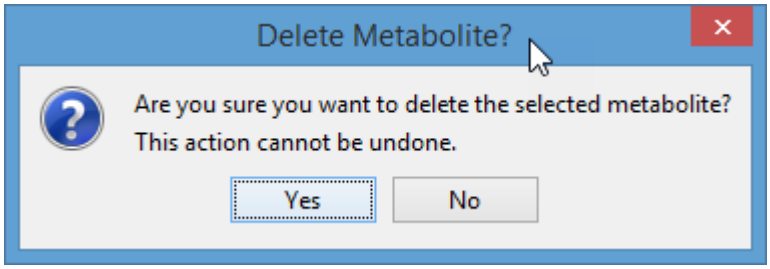
Now, select the last analyte in the list, Swainsonine.



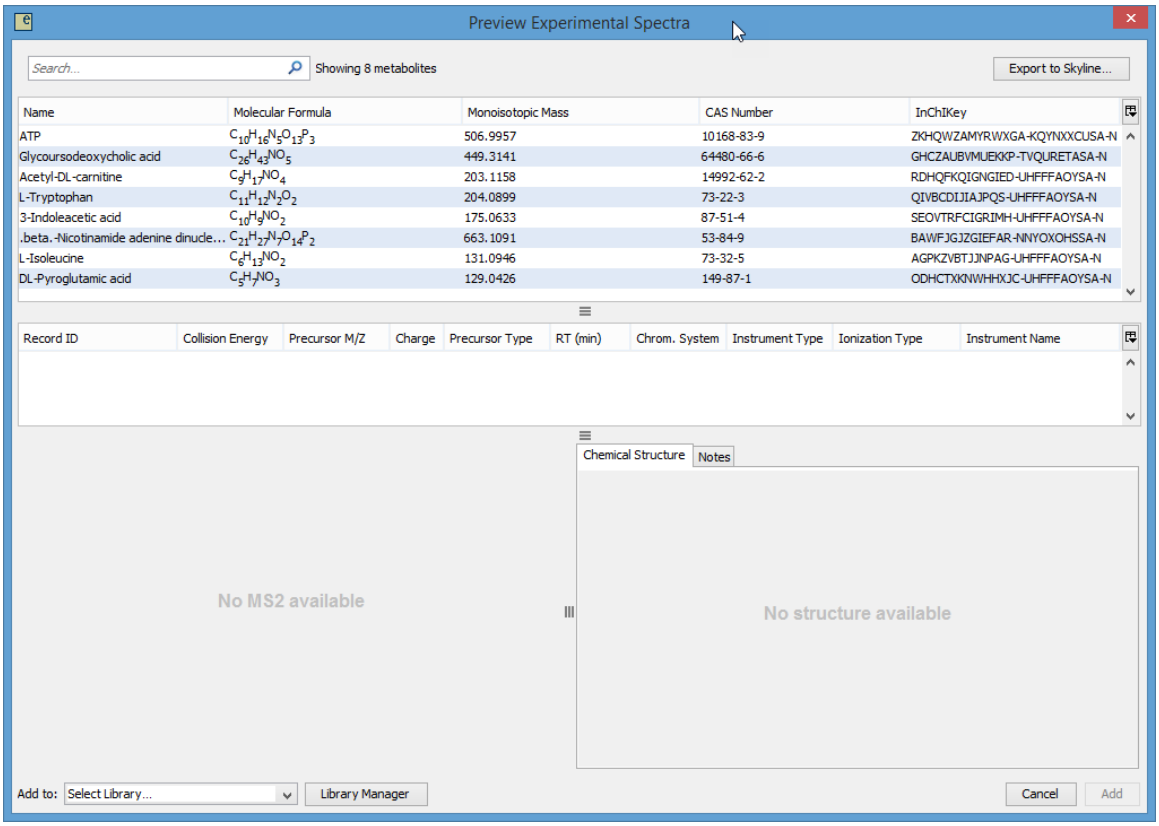
Let's suppose that we have decided that we do not want to retain this analyte after all. Right click on the row in the analytes table (the upper table), and select "Delete Analyte".



The following dialog will appear, asking you if you really want to delete Swainsonine.



Click "Yes". Notice that now, the Preview Dialog only contains 8 analytes (instead of 9).

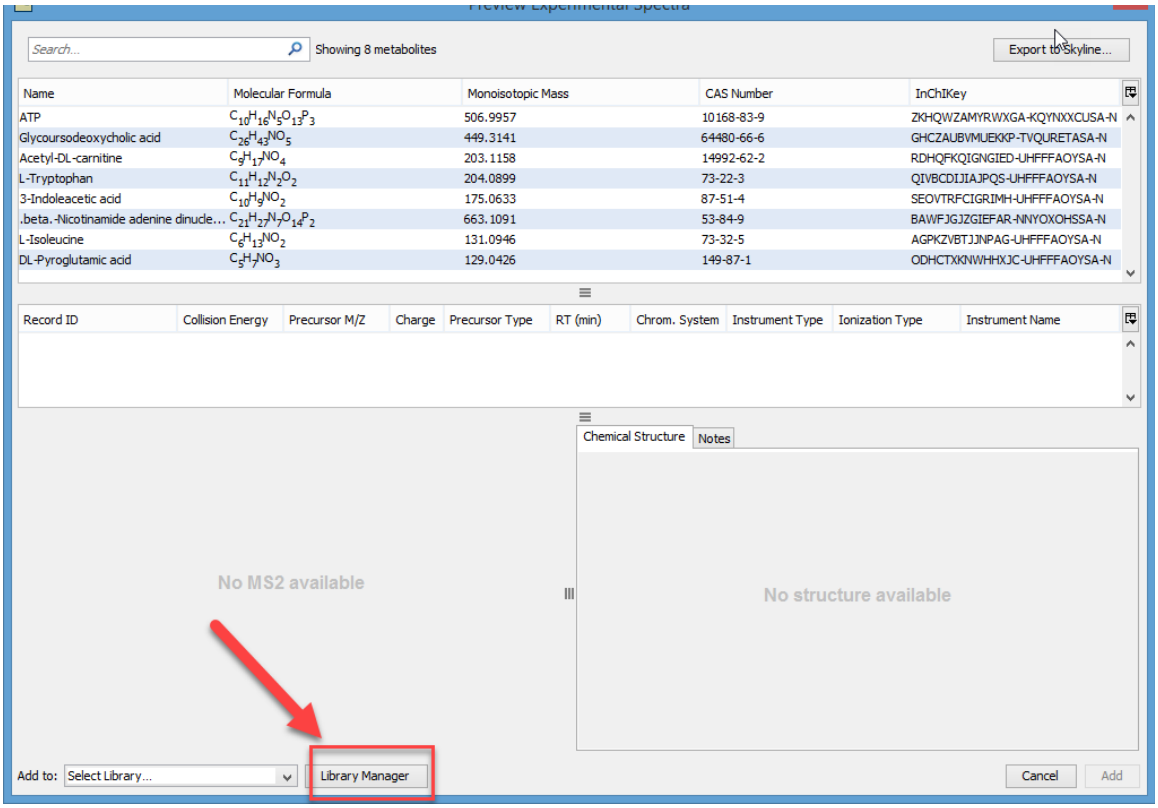


At this stage, we are ready to write this information into a new spectral library.

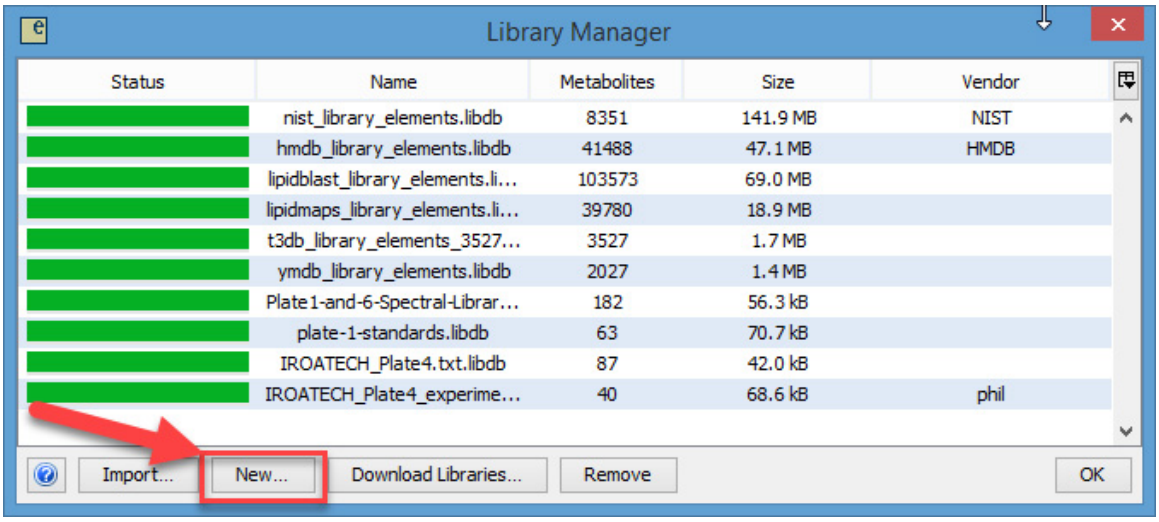
At the bottom of the Preview Dialog, click the “Library Manager” button, which will launch the Library Manager.

Step 2: Creating a new personal spectral library

We first need to create a new personal spectral library. Only personal spectral libraries created from within Elements are editable. To maintain the integrity of existing spectral libraries, Elements does not allow the user to add to or modify libraries imported from outside sources.



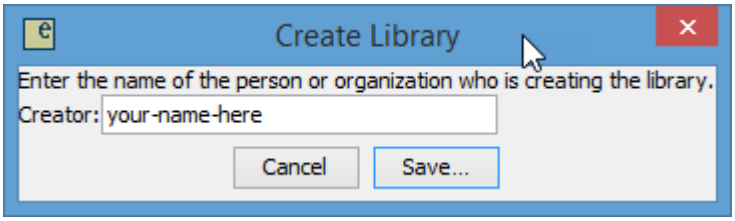
In the Library Manager dialog that appears, click the “New Library” button.



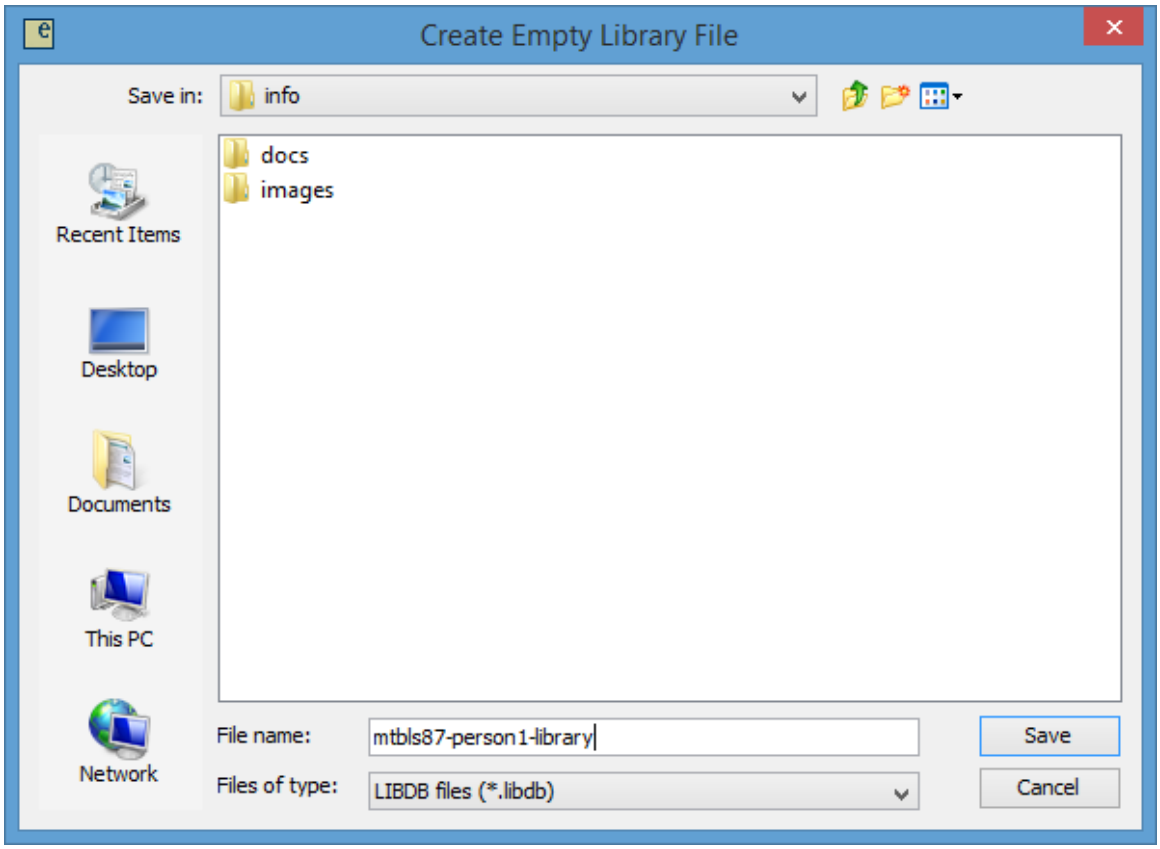
Clicking the "New…" button creates a new, empty Elements library file (.libdb file). Clicking the "Import…" button will invite the user to select a file, which will then be converted into an Elements library file. Acceptable file types for import include structural database (.SDF) files, .MSP files such as the Mass Bank of North America libraries, (available for download at http://mona.fiehnlab.ucdavis.edu/downloads) and tab-delimited text files (.txt).

When a libdb file is created with the "New…" button, it is editable – entries can be added, deleted, or modified. When a file is created with the "Import…" button, the resulting libdb is an exact reflection of the imported file, and is not editable.

Once you click the "New…" button, the following dialog will appear, inviting you to enter the name of the creator or organization creating the library.

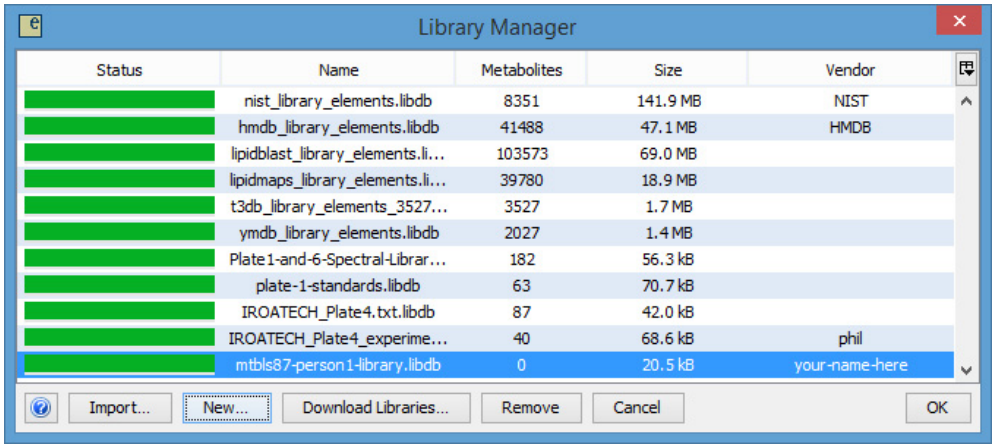


Enter your name, and click the "Save…" button. Navigate to an appropriate location on your file system, and name the library "mtbls87-person1-library".

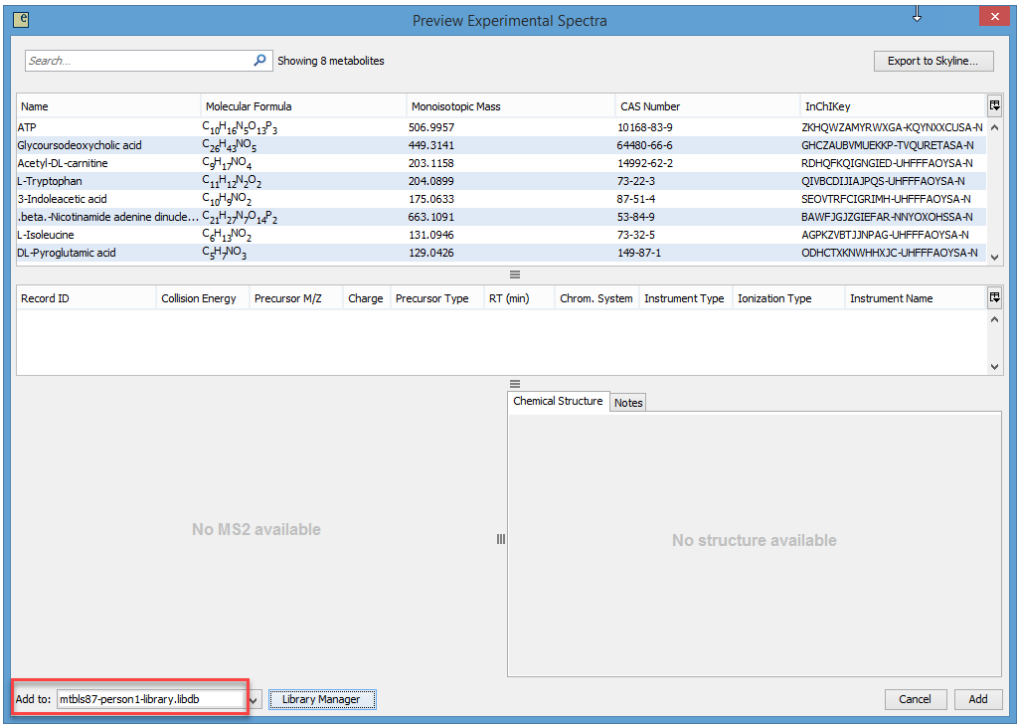


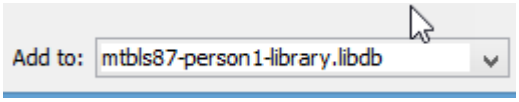
Click the “Save” button.

The library will now appear in the Library Manager, with 0 Analytes, created by “your-name-here”.



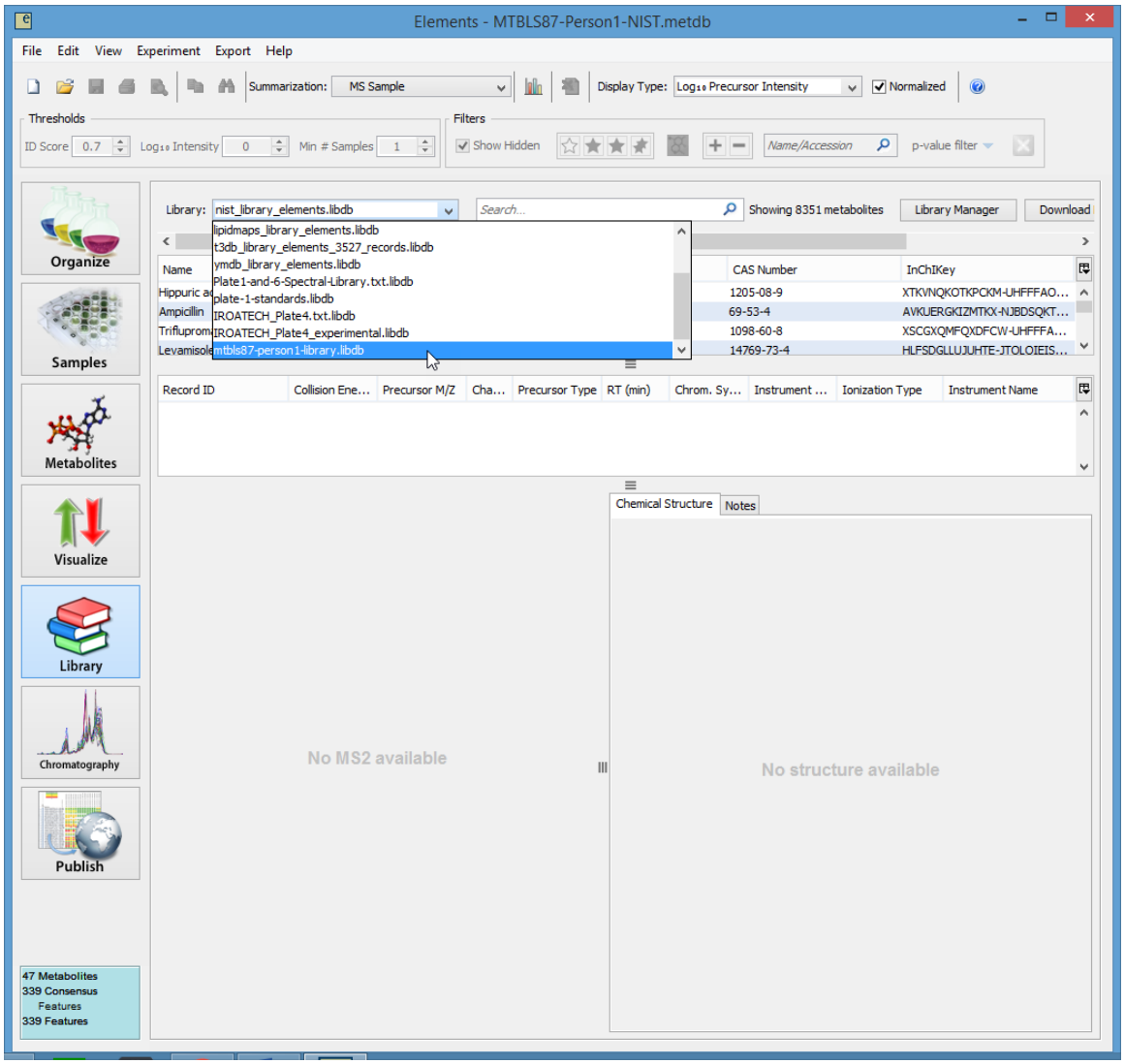
With this newly created library selected, click the “OK” button in the bottom right-hand corner of the Library Manager to return to the Preview Experimental Spectra dialog. Note that the newly created library is now selected in the drop-down list of available libraries.





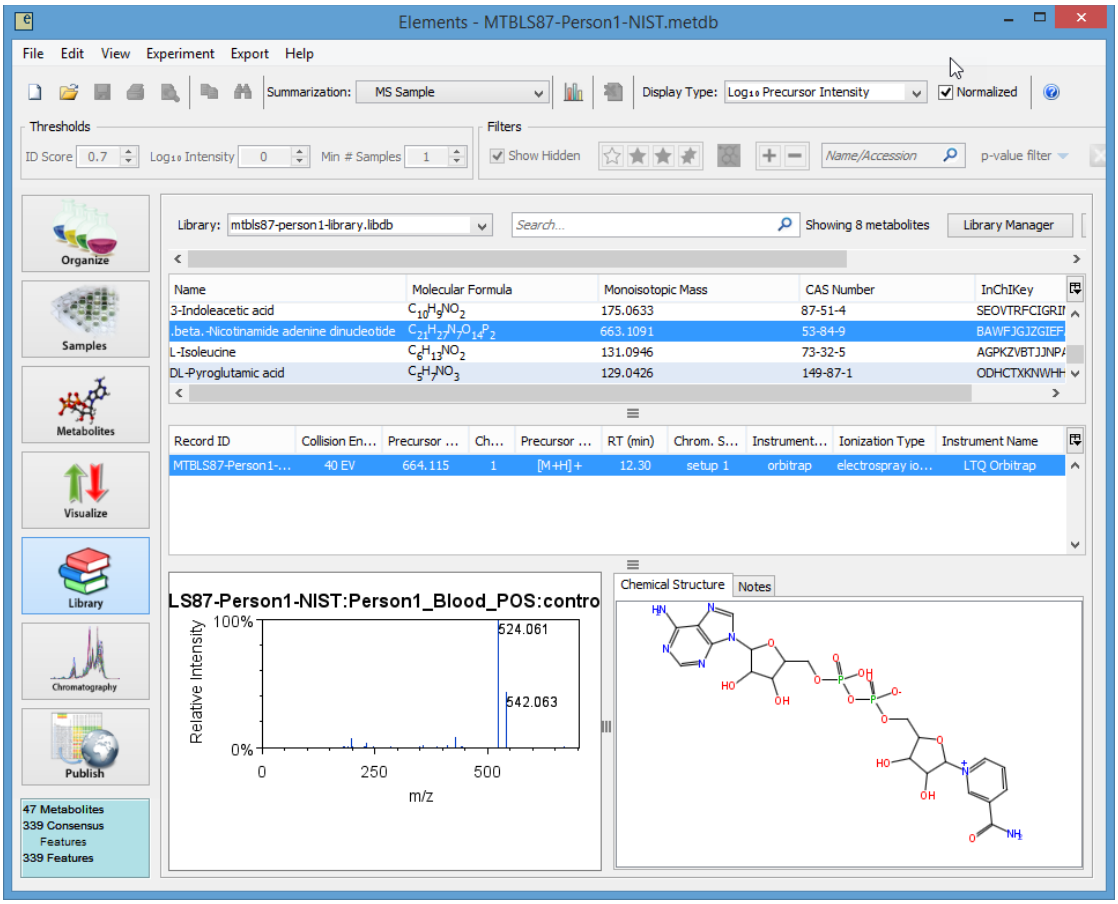
Now we are ready to write the analyte information shown in the preview dialog into our newly created, empty spectral library “mtbls87-person1-library”. Click the “Add” button in the lower left-hand corner of the Preview Experimental Spectra dialog. A dialog will appear as the data is transferred, and then disappear.

Returning to Elements, navigate to the Library View, and select the library “mtbls87-person1-library” from the drop-down menu in the Library drop-down menu located in the upper left-hand area of the Library View.

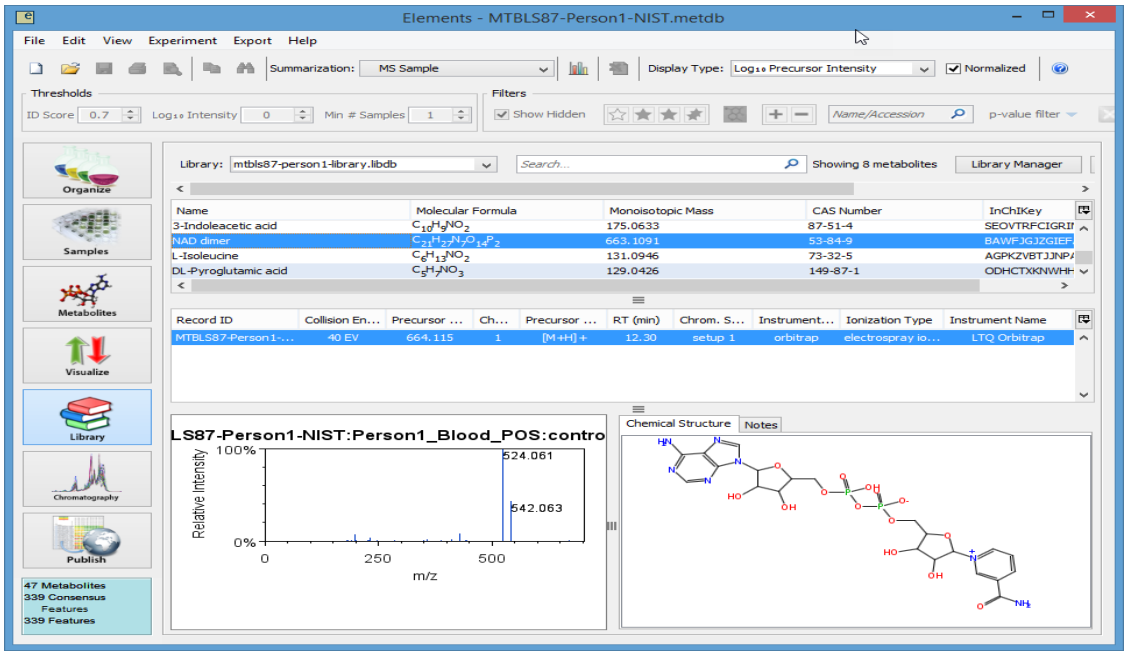


Because this library was created from your experimental data, it is editable. Not only is it editable in the Experimental Spectra Preview dialog, but also here in the Library View.

Select the analyte named ".beta.-Nicotinamide adenine dinucleotide".



Double click on this name, and replace ".beta.-Nicotinamide adenine dinucleotide" with "NAD dimer".

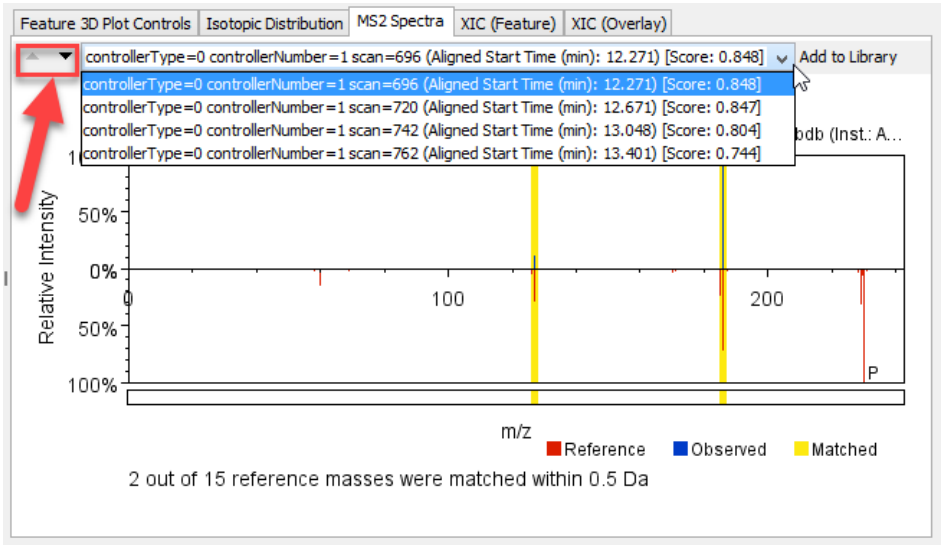


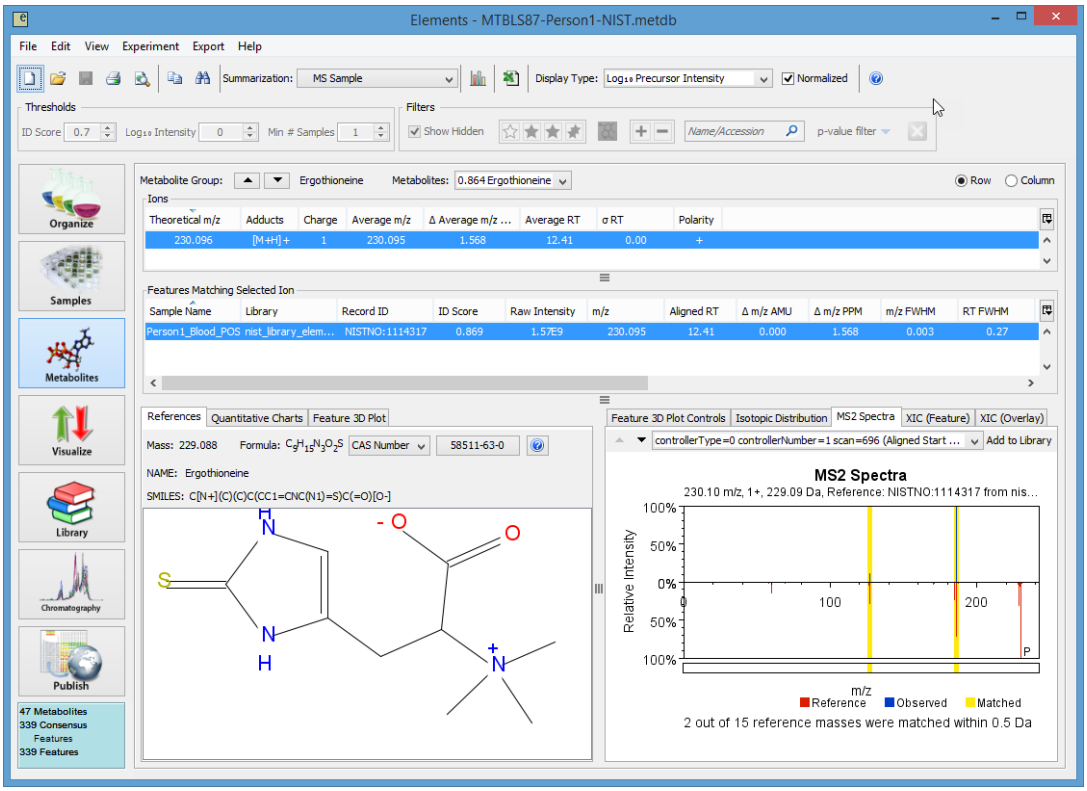
Adding individual entries to a personal spectral library

In addition to the ability to add spectra to a personal spectral library through the Samples View, it is also possible to add individual spectra to the library.

At this point, return to the Samples View. Double click on the analyte named “Ergothioneine” (analyte # 22). This will open the analytes view.

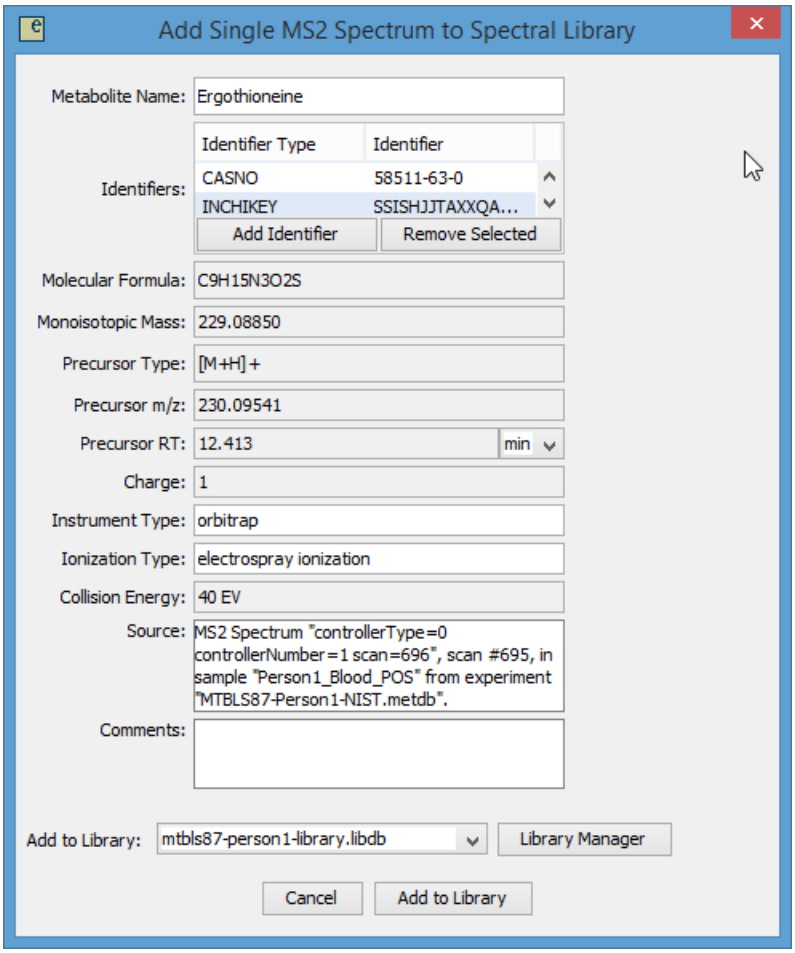
In the Analytes View that appears, the MS2 Spectra tab indicates that there were several MS2 spectra collected for this analyte. This is apparent in the drop-down list of four experimental spectra, as well as the enabled up and down arrow buttons to the left of the drop-down menu.


Feature 3D Plot Controls
Isotopic Distribution
MS2 Spectra
XIC (Feature)
XIC (Overlay)
controllerType=0 controllerNumber=1 scan=696 (Aligned Start Time (min): 12.271) [Score: 0.848]
Add to Library
controllerType=0 controllerNumber=1 scan=696 (Aligned Start Time (min): 12.271) [Score: 0.848]
controllerType=0 controllerNumber=1 scan=720 (Aligned Start Time (min): 12.671) [Score: 0.847]
controllerType=0 controllerNumber=1 scan=742 (Aligned Start Time (min): 13.048) [Score: 0.804]
controllerType=0 controllerNumber=1 scan=762 (Aligned Start Time (min): 13.401) [Score: 0.744]
Relative Intensity
50%
0%
50%
100%
100
200
m/z
Reference
Observed
Matched
2 out of 15 reference masses were matched within 0.5 Da


Elements - MTBLS87-Person1-NIST.metdb
File
Edit
View
Experiment
Export
Help
Summarization:
MS Sample
Display Type:
Normalized
Thresholds
Filters
Organize
Samples
Metabolites
Visualize
Library
Chromatography
Publish
Metabolite Group:
Ergothioneine
Metabolites:
0.864 Ergothioneine
Row
Column
Ions
Theoretical m/z
Adducts
Charge
Average m/z
Average RT
Polarity
230.096
[M+H]+
230.095
1.568
12.41
0.00
Features Matching Selected Ion
Sample Name
Library
Record ID
ID Score
Raw Intensity
m/z
Aligned RT
Person1_Blood_POS
nist_library_elem...
NISTNO:1114317
0.869
1.57E9
230.095
12.41
0.000
1.568
0.003
0.27
References
Quantitative Charts
Feature 3D Plot
Mass: 229.088
Formula:
CAS Number
58511-63-0
NAME: Ergothioneine
SMILES: C[N+](C)(C)C(CC1=CNC(N1)=S)C(=O)[O-]
MS2 Spectra
230.10 m/z, 1+, 229.09 Da, Reference: NISTNO:1114317 from nis...
Relative Intensity
100%
50%
0%
50%
100%
100
200
m/z
Reference
Observed
Matched
2 out of 15 reference masses were matched within 0.5 Da
47 Metabolites
339 Consensus Features
339 Features

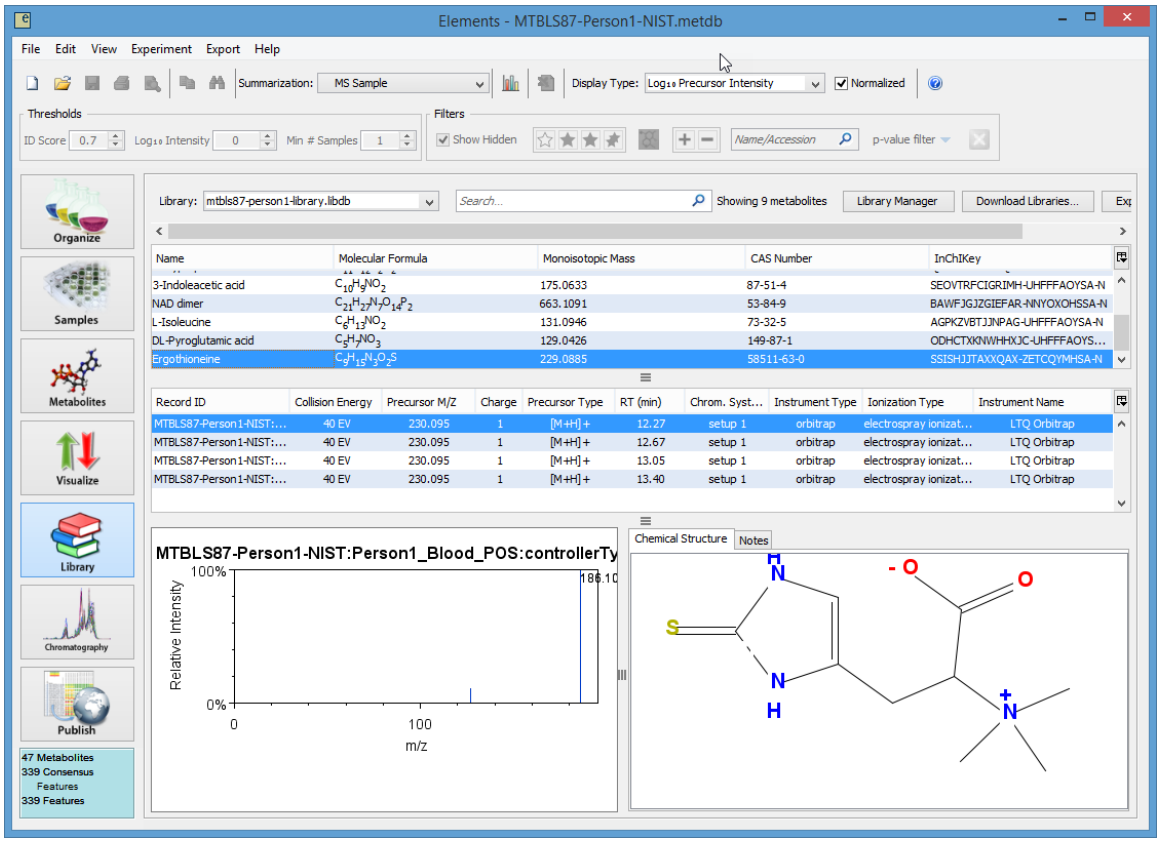
In order to add this spectrum to the personal spectral library, either click the "Add to Library" button or right-click on the chart and select "Add to library".

A dialog will appear, inviting the user to edit certain fields. In the "Add to Library" field at the bottom, select "mtbls87-person1-library", and click "add to library".



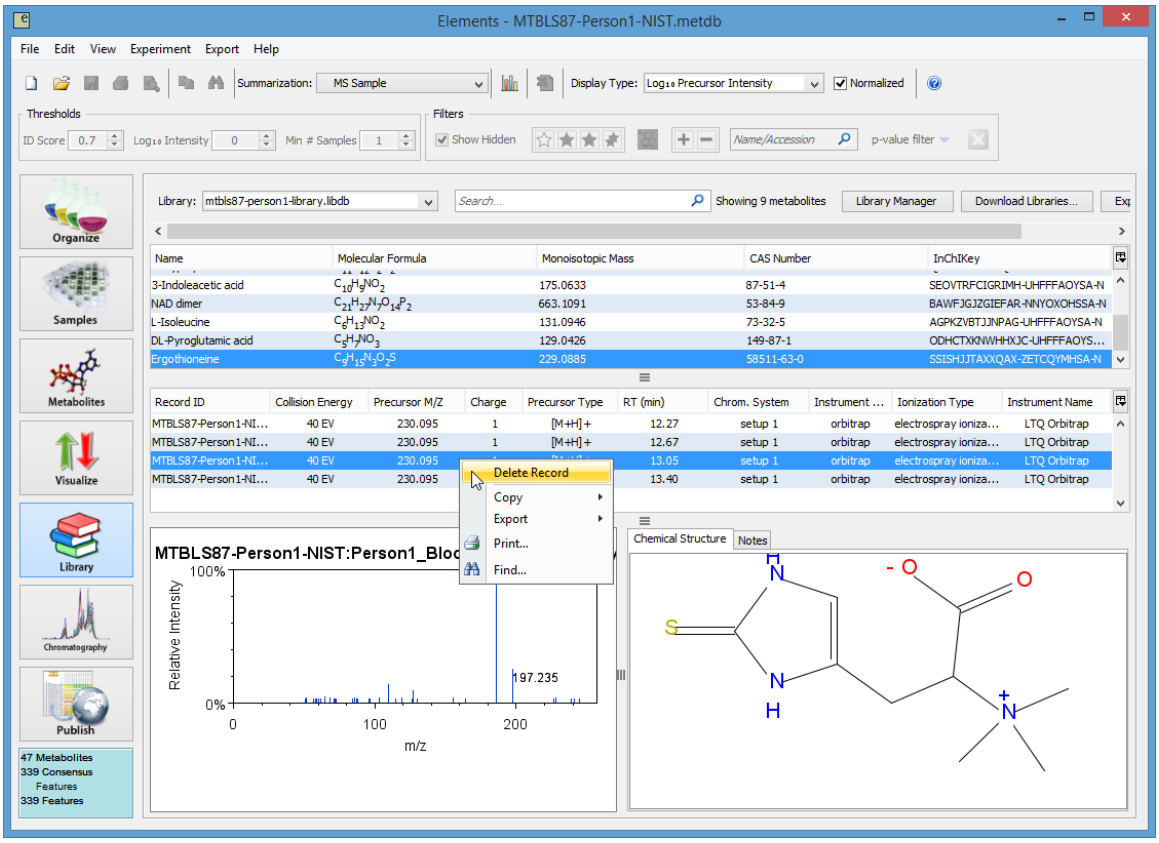
Repeat this process for all four Ergothioneine MS2 spectra, by selecting each in turn from the MS2 Spectra drop-down menu or clicking through the up and down arrow buttons to the left of the drop-down menu.

Switch to the Library View and select Ergothioneine. There are now four different records for Ergothioneine, each with a different MS2 spectrum.

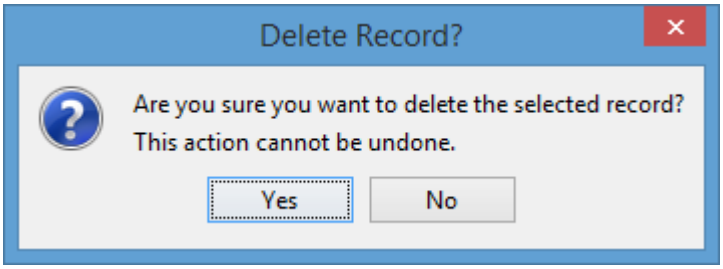


Investigate these four records by selecting each row in turn in the records table (you can also use the up and down arrow keys). Note that the third and fourth records, collected at RT = 13.05 min and 13.40 min, have much messier MS2s than the other two. Therefore, we will remove these records.

To remove a record from the table, simply right click on the selected row in the records table and choose “Delete Record”.



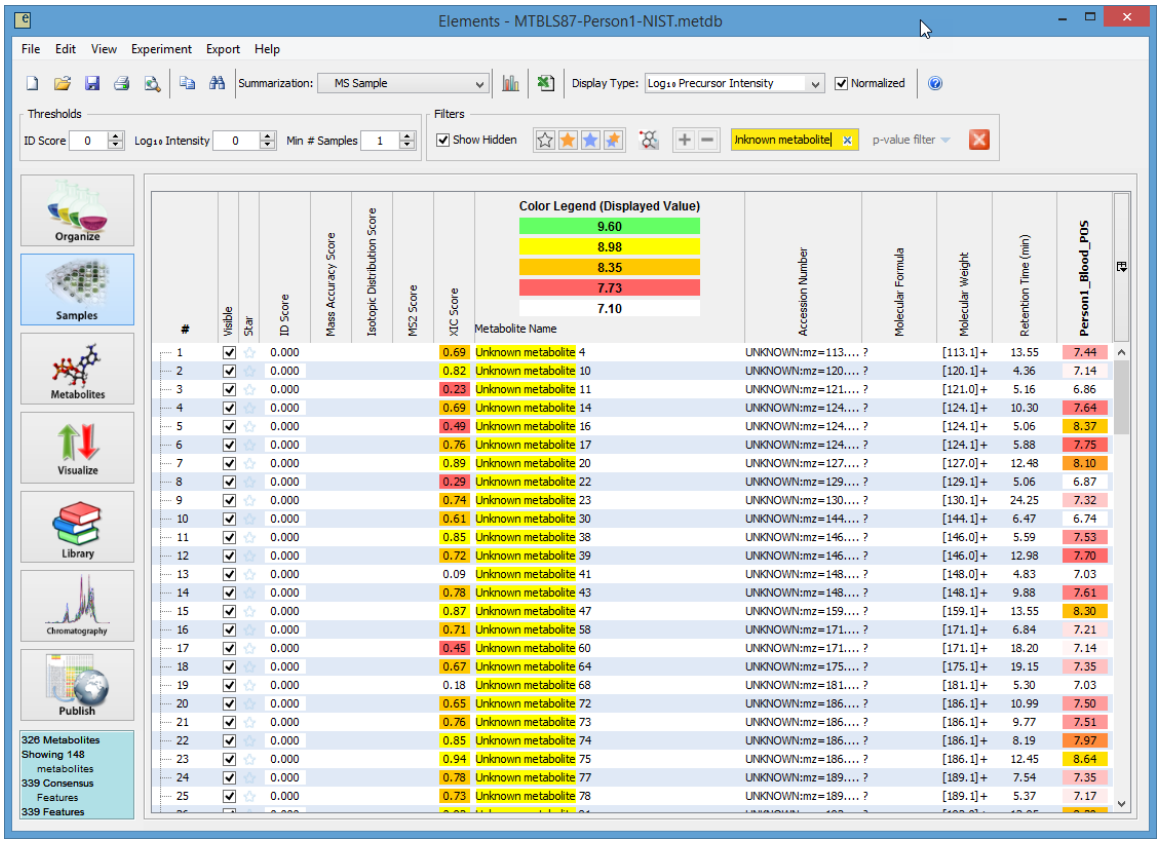
Note that it is not possible to delete records from an analyte if the analyte has only one record. In that case, delete the entire analyte instead by right-clicking on the analyte's name in the (upper) analytes table.



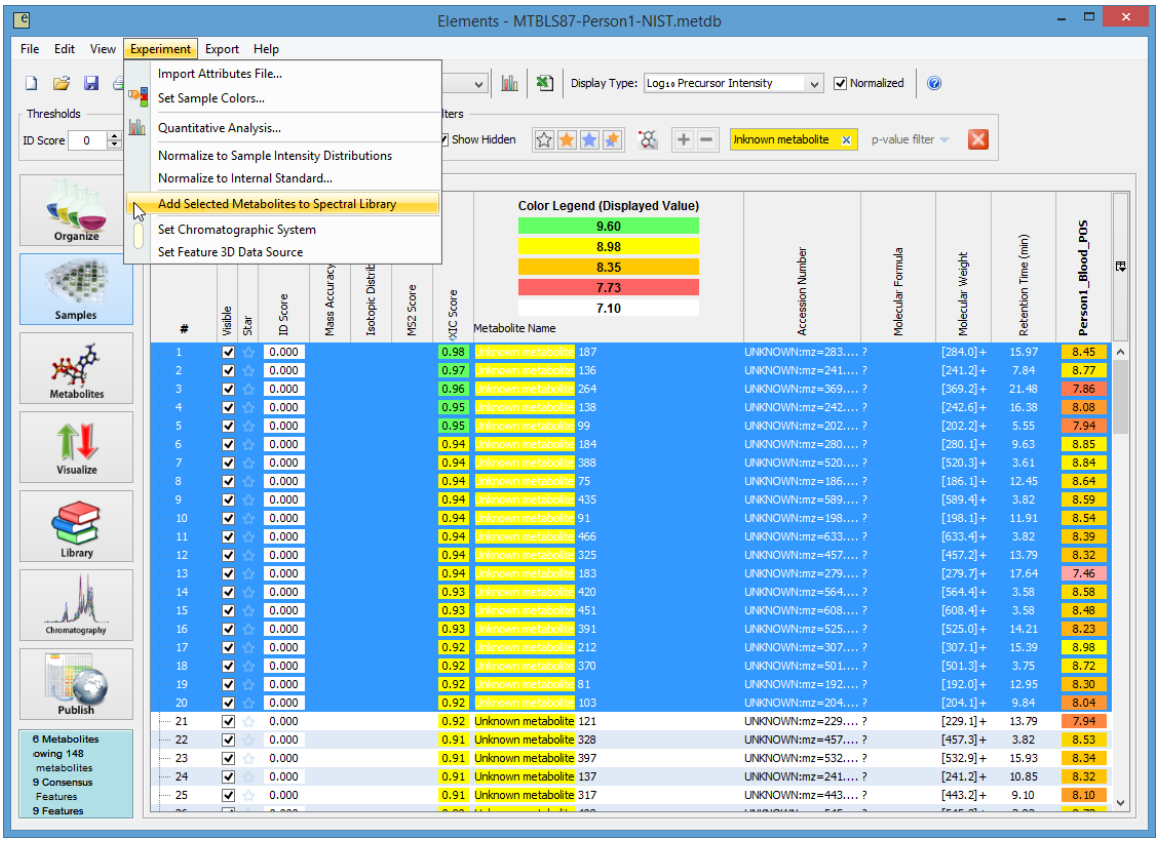
Adding Unknown Analytes to a personal spectral library

At this point, we will add analytes that could not be identified with NIST to our personal spectral library.

Return to the samples view. Set the ID score to 0.0, and type the text “Unknown analyte” in the Name/Accession filter. This will show all of the unidentified analytes in the experiment.



Click the column named “XIC Score” twice to sort the analytes in order of descending XIC score. Select the first 20 analytes, and select “Add Selected Analytes to Spectral Library” from the Experiment menu.

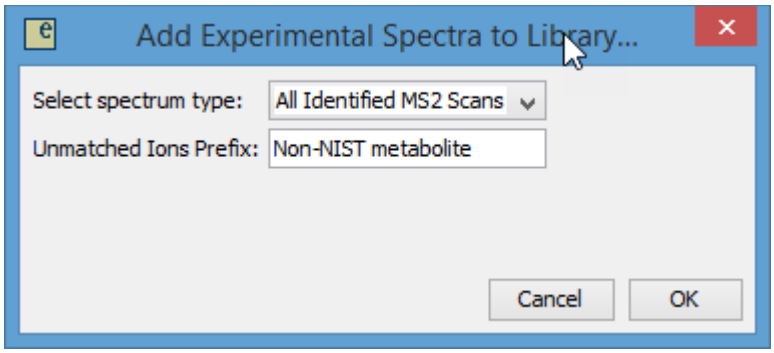


The “Add Experimental Spectra to Library…” dialog will appear.

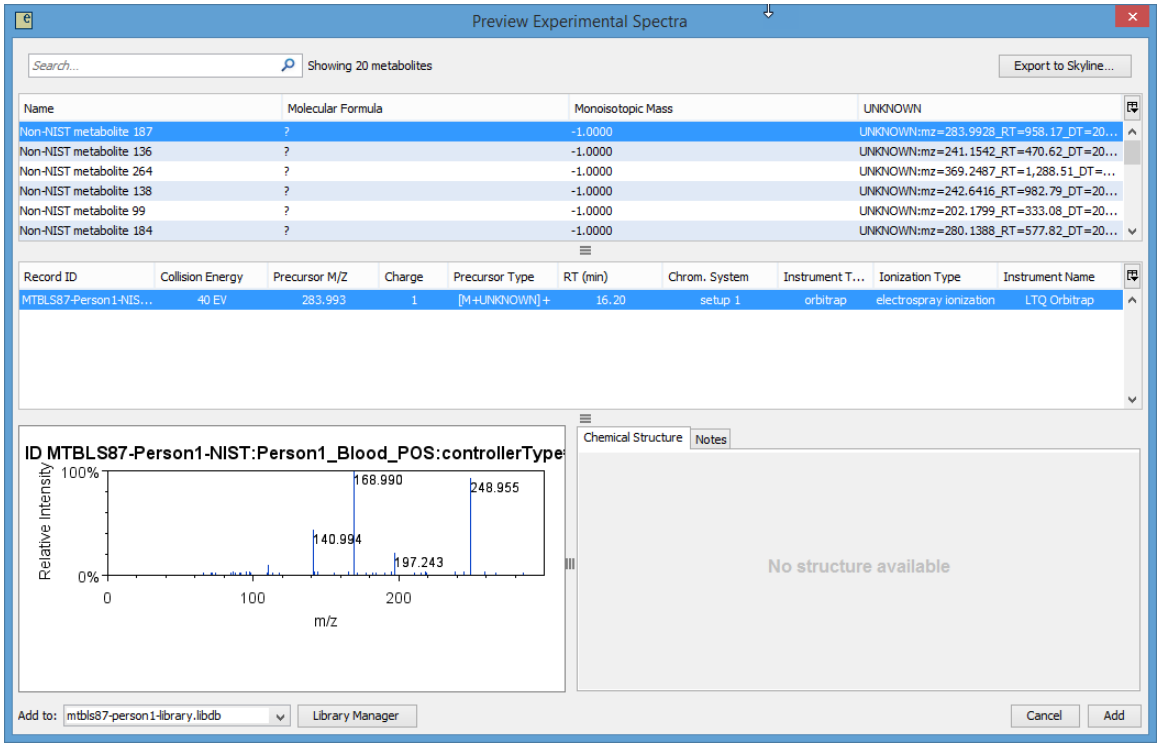
As described previously, the first option, “Select spectrum type:”, allows the user to specify how they would like the MS2 information gathered from the experiment to be written into a spectral library of interest.

The second option, “Unmatched Ions Prefix”, is relevant when dealing with unknown analytes.

Let us change this prefix from “Saved unknown analyte” to “Non-NIST analyte”



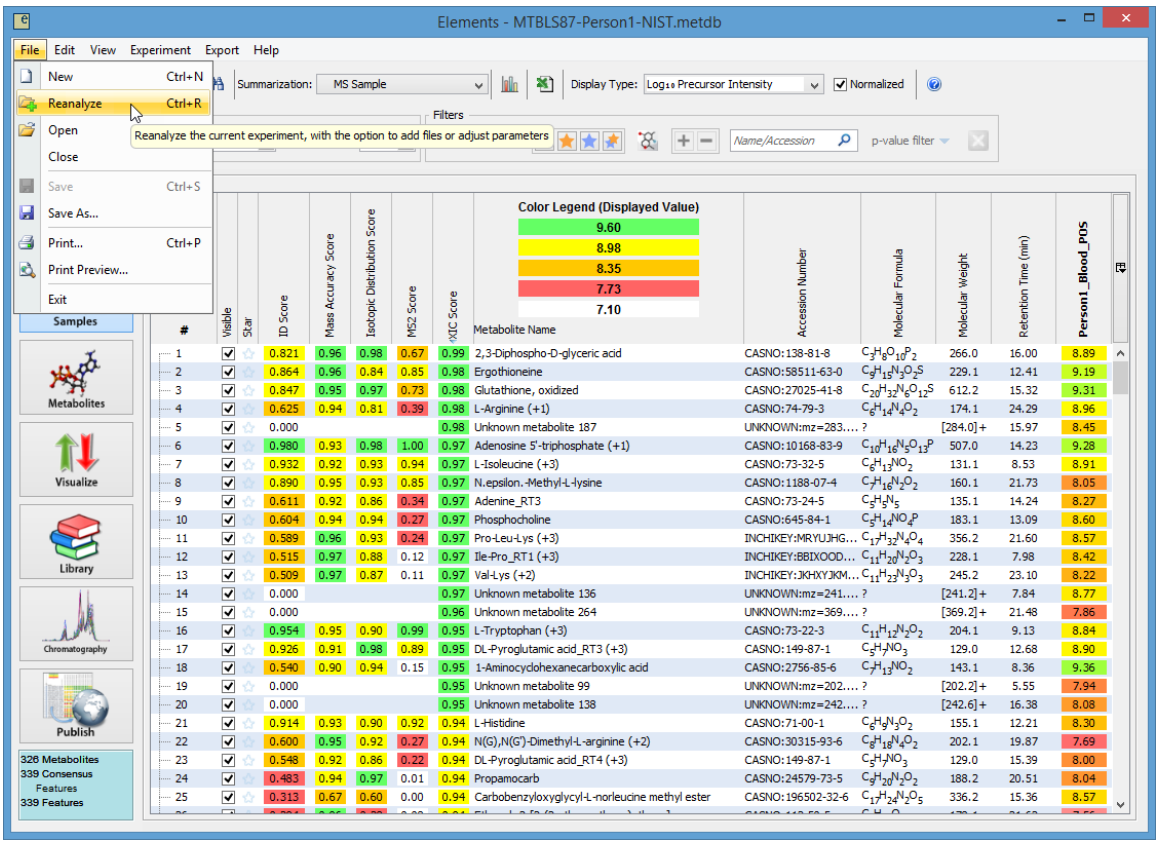
Click the “OK” button to launch the “Preview Experimental Spectra” dialog. Note that in the dialog that appears, all analytes are named “Non-NIST analyte” followed by a serial number. In the “Add to:” drop-down menu in the lower left-hand corner, select “mtbls87-person1-library”.



Click the “Add” button in the lower left-hand corner to add these analytes to the spectral library “mtbls87-person-1-library.”

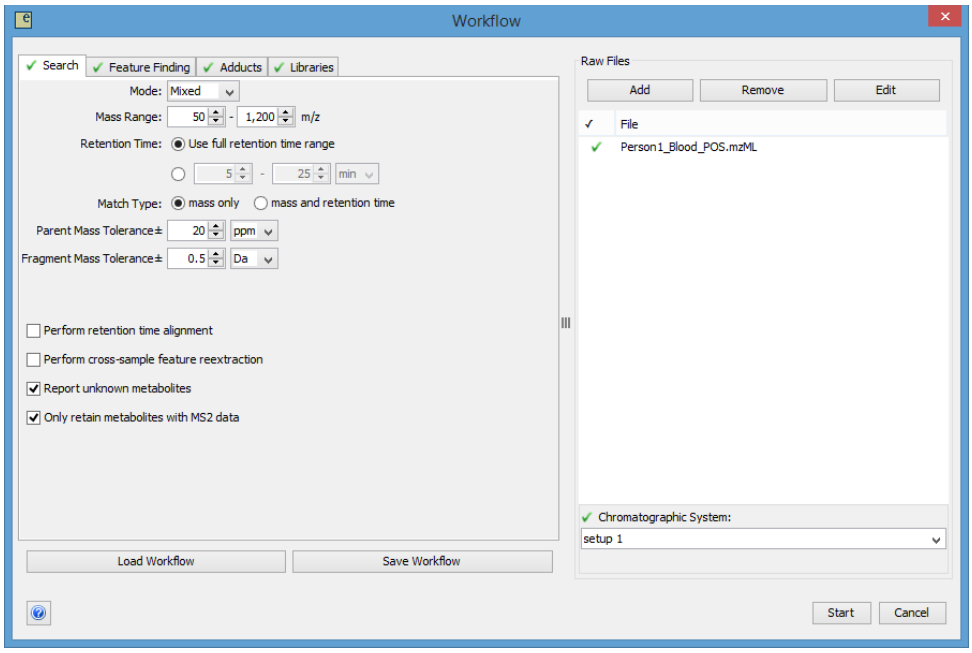
Step 3: Searching a personal spectral library

At this stage, we are ready to use our new spectral library. Return to the samples view, remove the “Unknown analyte” text filter, and save the file. Now, select “Reanalyze” from the File menu.



It is not necessary to use Reanalyze – it is also possible to create a new experiment, but Reanalyze provides the advantage that all previous experimental parameters and files are already filled in. We would like to reuse most of these parameters, so this option makes sense.

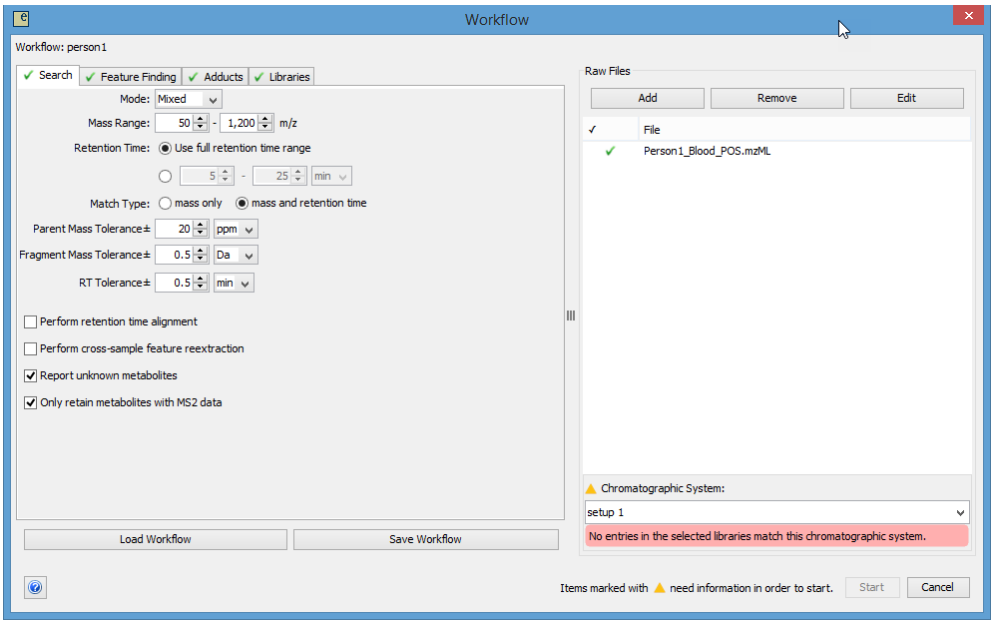
The loading dialog will appear, showing all of the parameters and files associated with the current experiment.



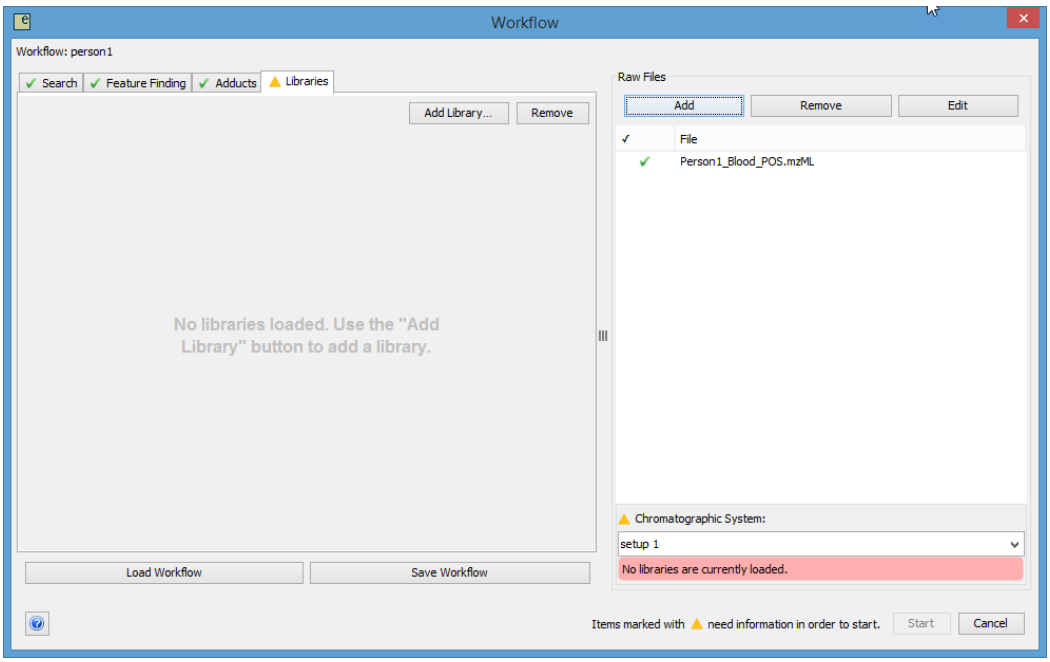
Note that all parameters and settings are valid, and we could click the Start button now. This is exactly what we would expect, as we launched this dialog from a valid experiment. If we did click the Start button, we would produce an exact copy of the original file. At this point, for good measure, let us save this workflow by clicking the “Save Workflow” button. Call the workflow “person1”.

Now, we will make some key modifications. In the search tab, change the “Match Type:” radio button from “mass only” to “mass and retention time”. Note that the option to search by both mass and retention time is a new feature in Elements 1.3.

Once this radio button has been changed, the loading dialog will become invalid, displaying the error message “No entries in the selected libraries match this chromatographic system.”

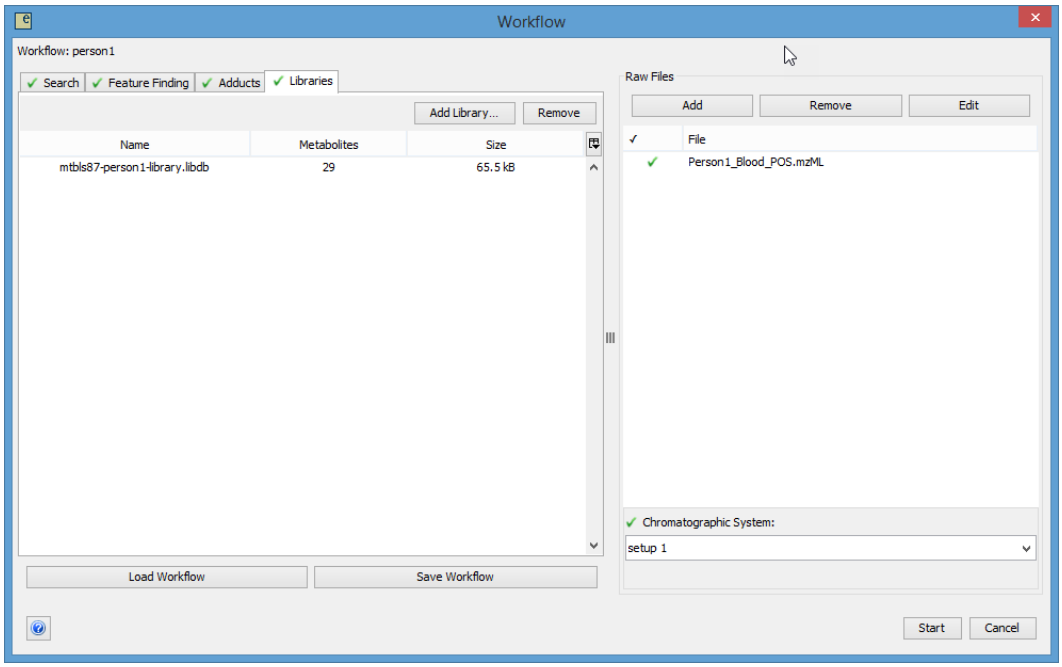


Switch to the Libraries tab, select the NIST library, and remove it by clicking the "Remove" button in the Libraries tab. The error message in the Chromatographic System loading dialog will now change to "No libraries are currently loaded".



Now click the "Add Library…" button, and navigate to our personal spectral

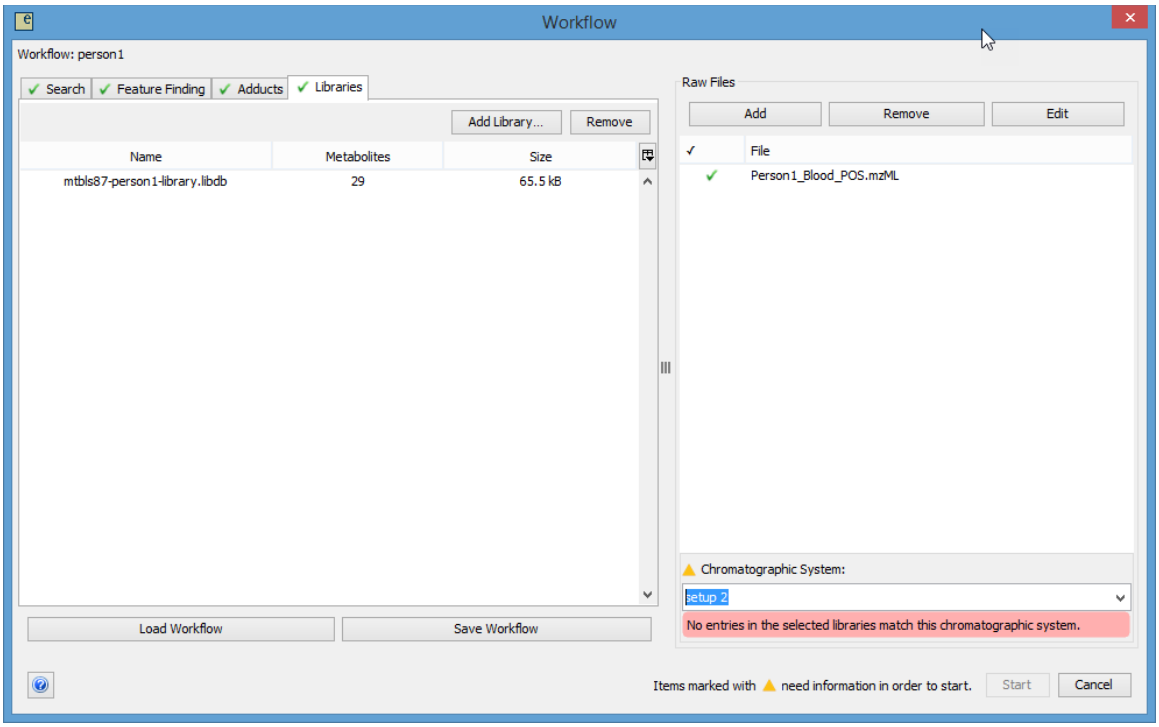
library, “mtbls87-person1-library” in the library manager. Once this library is selected, click the “OK” button in the lower right-hand corner of the Library Manager. This will cause the error message associated with the Chromatographic System panel to disappear.



Why did the validation logic related to the Chromatographic System panel change? Initially, we searched data against NIST, which does not contain retention time information. We are now switching our mode to search by mass and retention time, so we need to search against at least one or more library that has retention time information. Our personal spectral library, mtbls87-person1-library, has retention time information.

However, it is not enough to have any retention time information in the library, we need to have retention time information that is comparable. To illustrate this case, change the text in the Chromatographic System panel from “setup 1” to “setup 2”.

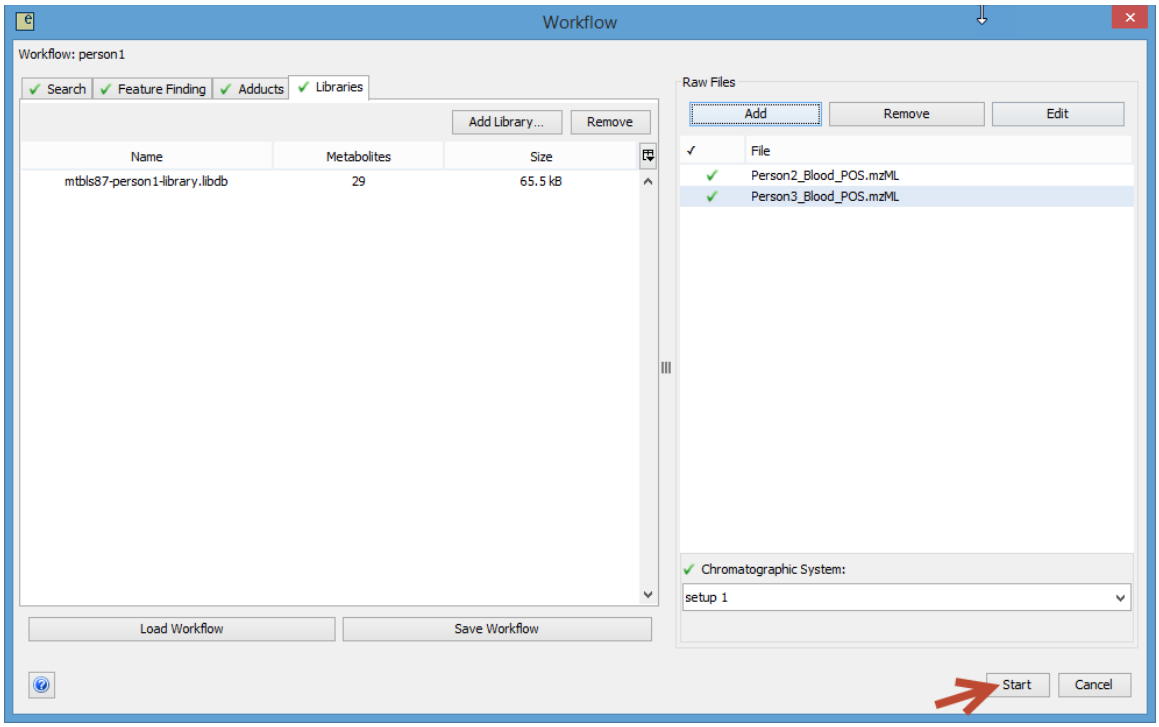
A new error message appears: “No entries in the selected libraries match this chromatographic system.”



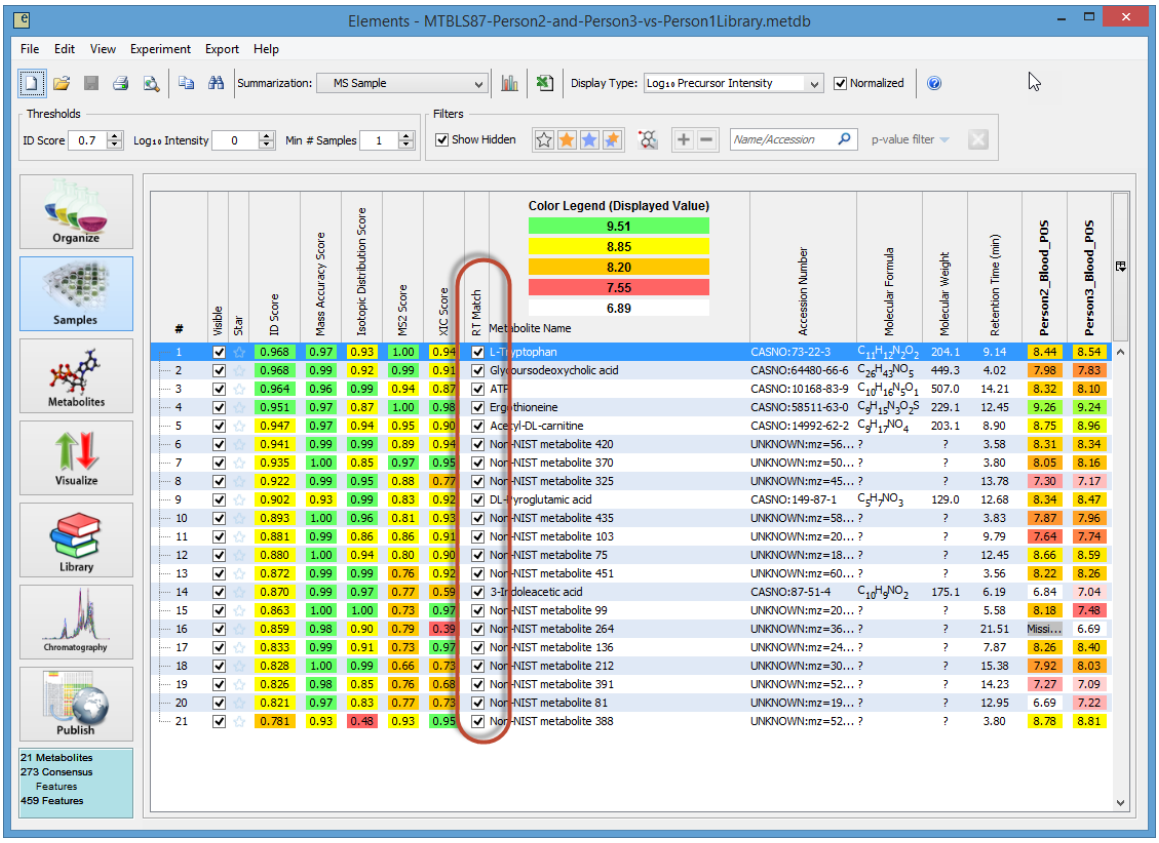
All of the information in mtbls87-person1-library was extracted from Person1, which has the chromatographic system type of “setup 1.” Entries with a chromatographic system of “setup 1” cannot be compared to “setup 2”, so the program will not let you proceed until you have specified comparable setups (In other words, the chromatographic system must be the same).

Change the chromatographic system back to “setup 1”. Now the RT information can be compared, so the dialog is valid again.

Finally, we need to adjust the Raw Files that we loaded. As our personal spectral library is build from Person1_Blood_POS.mzML data, we should remove this file, and add the other two raw data files (Person2_Blood_POS.mzML and Person3_Blood_POS.mzML). This corresponds to specifying experimental data to be searched against the records from known analytes used in creation of the spectral library.



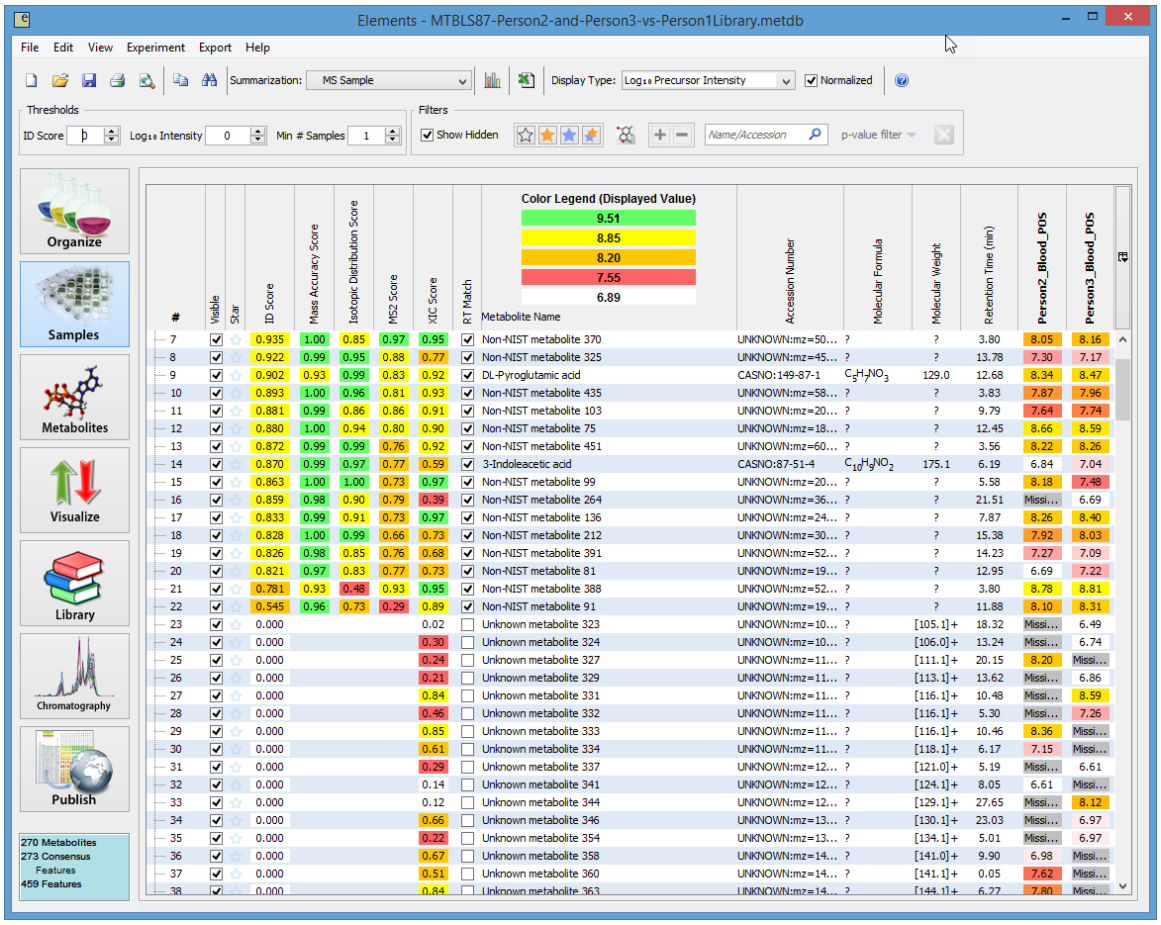
Click the start button to start the analysis. After a few minutes, the results will appear in the samples view. Save the file, calling it "MTBLS87-Person2-and-Person3-vs-Person1Library"



Note that we see a new column, “RT Match” with a checkbox for every entry. This indicates that every analyte identification agrees in RT as well as mass with the searched spectral library.

Note the presence of the “Non-NIST analytes”. Though we don’t know what these analytes are (in terms of a molecular formula), it is clear that they exist in Person2_Blood_POS and Person3_Blood_POS. We are even able to score the quality of the matches based on m/z, RT, observed MS2 spectrum, and observed isotopic distribution (based on the observed intensities of previously identified [M+1] and [M+2] peaks).

Set the ID score down to 0 by changing the “ID Score” threshold in the upper left-hand corner.

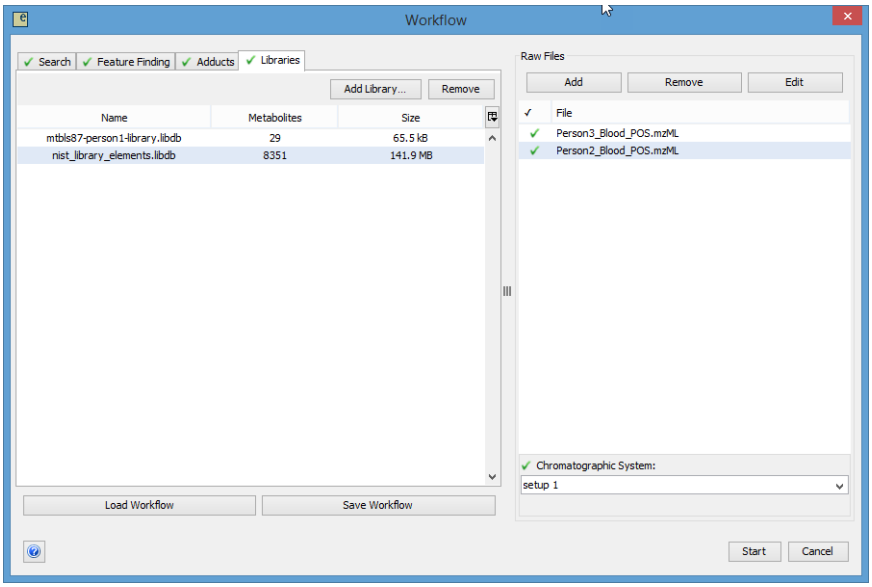


#	Visible	Star	ID Score	Mass Accuracy Score	Isotopic Distribution Score	MS2 Score	XIC Score	RT Match	Metabolite Name	Accession Number	Molecular Formula	Molecular Weight	Retention Time (min)	Person2_Blood_POS	Person3_Blood_POS
7	✓		0.935	1.00	0.85	0.97	0.95	✓	Non-NIST metabolite 370	UNKNOWN:mz=50...	?	?	3.80	8.05	8.16
8	✓		0.922	0.99	0.95	0.88	0.77	✓	Non-NIST metabolite 325	UNKNOWN:mz=45...	?	?	13.78	7.30	7.17
9	✓		0.902	0.93	0.99	0.83	0.92	✓	DL-Pyroglutamic acid	CASNO:149-87-1	$C_5H_7NO_3$	129.0	12.68	8.34	8.47
10	✓		0.893	1.00	0.96	0.81	0.93	✓	Non-NIST metabolite 435	UNKNOWN:mz=58...	?	?	3.83	7.87	7.96
11	✓		0.881	0.99	0.86	0.86	0.91	✓	Non-NIST metabolite 103	UNKNOWN:mz=20...	?	?	9.79	7.64	7.74
12	✓		0.880	1.00	0.94	0.80	0.90	✓	Non-NIST metabolite 75	UNKNOWN:mz=18...	?	?	12.45	8.66	8.59
13	✓		0.872	0.99	0.99	0.76	0.92	✓	Non-NIST metabolite 451	UNKNOWN:mz=60...	?	?	3.56	8.22	8.26
14	✓		0.870	0.99	0.97	0.77	0.59	✓	3-Indoleacetic acid	CASNO:87-51-4	$C_{10}H_9NO_2$	175.1	6.19	6.84	7.04
15	✓		0.863	1.00	1.00	0.73	0.97	✓	Non-NIST metabolite 99	UNKNOWN:mz=20...	?	?	5.58	8.18	7.48
16	✓		0.859	0.98	0.90	0.79	0.39	✓	Non-NIST metabolite 264	UNKNOWN:mz=36...	?	?	21.51	Missi...	6.69
17	✓		0.833	0.99	0.91	0.73	0.97	✓	Non-NIST metabolite 136	UNKNOWN:mz=24...	?	?	7.87	8.26	8.40
18	✓		0.828	1.00	0.99	0.66	0.73	✓	Non-NIST metabolite 212	UNKNOWN:mz=30...	?	?	15.38	7.92	8.03
19	✓		0.826	0.98	0.85	0.76	0.68	✓	Non-NIST metabolite 391	UNKNOWN:mz=52...	?	?	14.23	7.27	7.09
20	✓		0.821	0.97	0.83	0.77	0.73	✓	Non-NIST metabolite 81	UNKNOWN:mz=19...	?	?	12.95	6.69	7.22
21	✓		0.781	0.93	0.48	0.93	0.95	✓	Non-NIST metabolite 388	UNKNOWN:mz=52...	?	?	3.80	8.78	8.81
22	✓		0.545	0.96	0.73	0.29	0.89	✓	Non-NIST metabolite 91	UNKNOWN:mz=19...	?	?	11.88	8.10	8.31
23	✓		0.000				0.02		Unknown metabolite 323	UNKNOWN:mz=10...	?	[105.1]+	18.32	Missi...	6.49
24	✓		0.000				0.30		Unknown metabolite 324	UNKNOWN:mz=10...	?	[106.0]+	13.24	Missi...	6.74
25	✓		0.000				0.24		Unknown metabolite 327	UNKNOWN:mz=11...	?	[111.1]+	20.15	8.20	Missi...
26	✓		0.000				0.21		Unknown metabolite 329	UNKNOWN:mz=11...	?	[113.1]+	13.62	Missi...	6.86
27	✓		0.000				0.84		Unknown metabolite 331	UNKNOWN:mz=11...	?	[116.1]+	10.48	Missi...	8.59
28	✓		0.000				0.46		Unknown metabolite 332	UNKNOWN:mz=11...	?	[116.1]+	5.30	Missi...	7.26
29	✓		0.000				0.85		Unknown metabolite 333	UNKNOWN:mz=11...	?	[116.1]+	10.46	8.36	Missi...
30	✓		0.000				0.61		Unknown metabolite 334	UNKNOWN:mz=11...	?	[118.1]+	6.17	7.15	Missi...
31	✓		0.000				0.29		Unknown metabolite 337	UNKNOWN:mz=12...	?	[121.0]+	5.19	Missi...	6.61
32	✓		0.000				0.14		Unknown metabolite 341	UNKNOWN:mz=12...	?	[124.1]+	8.05	6.61	Missi...
33	✓		0.000				0.12		Unknown metabolite 344	UNKNOWN:mz=12...	?	[129.1]+	27.65	Missi...	8.12
34	✓		0.000				0.66		Unknown metabolite 346	UNKNOWN:mz=13...	?	[130.1]+	23.03	Missi...	6.97
35	✓		0.000				0.22		Unknown metabolite 354	UNKNOWN:mz=13...	?	[134.1]+	5.01	Missi...	6.97
36	✓		0.000				0.67		Unknown metabolite 358	UNKNOWN:mz=14...	?	[141.0]+	9.90	6.98	Missi...
37	✓		0.000				0.51		Unknown metabolite 360	UNKNOWN:mz=14...	?	[141.1]+	0.05	7.62	Missi...
38	✓		0.000				0.84		Unknown metabolite 363	UNKNOWN:mz=14...	?	[144.1]+	6.27	7.80	Missi...

Note the presence of Unknown analytes. Non-NIST analytes were analytes that could not be identified by NIST, but were present in Person1_Blood_POS, and saved in our personal spectral library metbls87-person1-library. Unknown analytes are analytes that were not found in our personal library, mtbls87-person1-library. It is important to keep in mind that “unknown” is only unknown with respect to the spectral libraries that were searched.

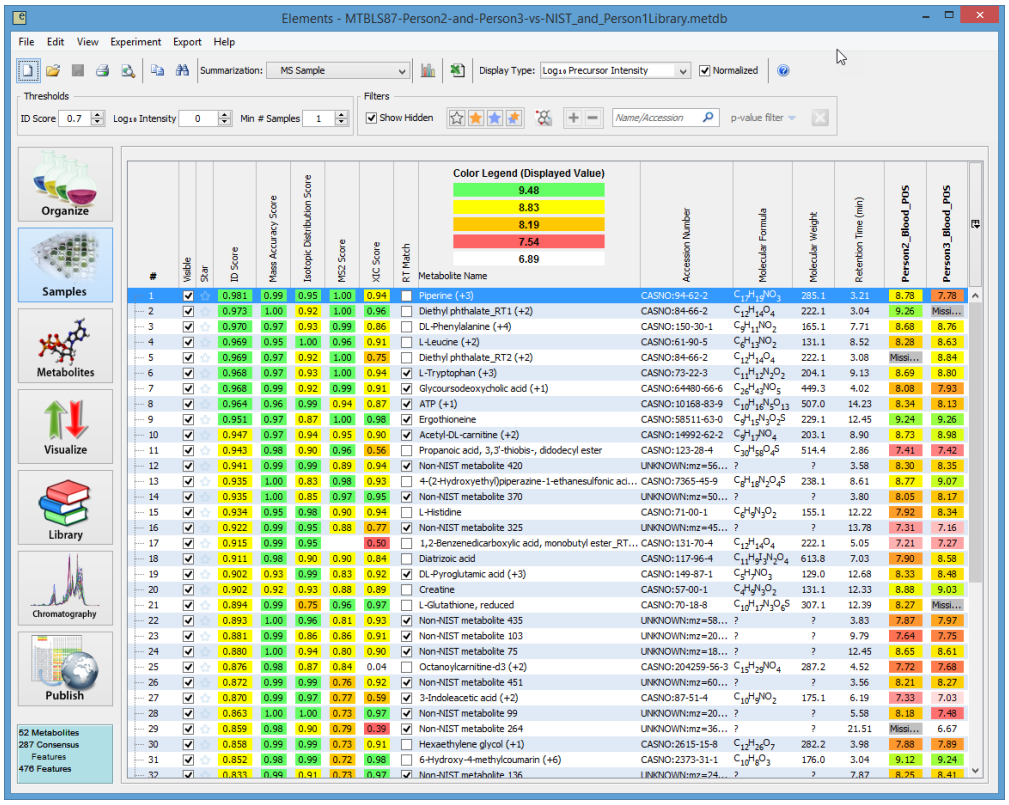
Finally, we will re-analyze one more time, this time combining our results from mtbls87-person1-library and NIST.

Select “Reanalyze” from the File menu. In the loading dialog that appears, switch to the Libraries tab and add the NIST library to the mtbls87-person1-library. No other changes are necessary.



Click the “Start” button in the bottom right-hand corner.

After a few minutes, the results should appear. Save the file, naming it “MTBLS87-Person2-and-Person3-vs-NIST_and_Person1Library.metdb”.



Notice that now, not every entry in the RT Match column is checked – only those entries where RT information was available in the library.

This concludes our tutorial. Personal spectral libraries can enable you to transform experimental results into useful metrics for future experiments. In this tutorial, we also highlighted the utility of “Reanalyze”, which can be useful for fast re-analysis of data with slight parameter adjustments (for example, adding a spectral library and re-searching). In this tutorial, we did not choose to save indexed feature files, but if we had, in many cases we could have reused these results when we reanalyzed.

Appendix B.Creating a Custom Spectral Library using a tab-delimited text file

Elements can load custom spectral libraries if the information is saved in a tab-delimited text file structured as described hereafter:

- Analytes should be specified in a table, one analyte per row. No missing values are allowed in a row.
- Columns are interpreted according to their titles and may appear in any order. Empty columns are not allowed.
- Column titles are case-insensitive.
- Any additional columns will be ignored.
- Tables should be saved as tab-delimited text files with extensions TXT or TSV and loaded via the Library Manager dialog accessible through the menu command **Edit > Library Manager...** or the Workflow dialog.
- Required Columns:
- Optional Columns:
- Examples of acceptable table formats:

Required Columns:

- **Analyte Name** - required, must be unique. Acceptable titles: “analyte name”, “name”, “analyte”, “ion”, “compound”, “species”
- **Formula** - required, designates the chemical formula of the species. Acceptable titles: “formula”, “molecular formula”, “chemical formula”

Optional Columns:

- ***Ionization Mode** - ionization mode used to generate an MS2 spectrum. This parameter is not required, and is currently only used for display in the library view. If it is not specified, MS2 spectra will have a charge value of “0” in the library view, indicating that the charge is unknown.

  Values: + for positive or - for negative mode

  Acceptable titles: “ionization mode”, *“mode”*
- **Precursor** - the mass of the precursor form of the analyte. This column is included for transition libraries.

  Acceptable titles: *“*precursor*”*

- **Product** - the mass of a single fragment of the analyte. This column is included for transition libraries.

  Acceptable titles: “product”

- **COLL_ENERGY_EV** - the collision energy, given in electron volts (EV). It will be used for MS2 spectrum display only.

  Acceptable titles: any string that contains the words “collision energy” and *“ev”*

- **COLL_ENERGY_PCT** - the collision energy, given as a percentage. It will be used for MS2 spectrum display only.

  Acceptable titles: any string that contains the words “collision energy” and “%” or *“pct”*

- **SMILES** - the SMILES representation of the analyte, which is necessary to produce a molecular structure drawing.

  Acceptable titles: any string that contains the word "smiles"

- ***MS2_SPECTRUM** - contains a feature list of a single MS2 spectrum. A specific format is required, as shown in the examples below.

  Acceptable titles: “ms2_spectrum”, *“ms2”*, *“ms2s”*, *“ms2_spectra”*

- **CAS_NUMBER** - the CAS number of the analyte. If the CAS number is valid, this will be passed on and associated with the analyte. CAS numbers, like all identifiers, are used only as metadata.

  Acceptable titles: “cas”, “casno”, “cas number”

  * NOTE- If the MS2_SPECTRUM column is included, the Ionization_Mode column should also be included.

Examples of acceptable table formats:

The tables should be saved as tab-delimited text files.

Figure 1: Example 1

name	formula	cas	smiles	mode	ms2_spectrum
LEUKOTRIENE B4	C20H32O4	71160-24-2	CCCCCC=CCC(C=CC=CC=CC(CCCC(=O)O)O)O	+	{(105.0719,7663.8); (107.0509,16581.0);(151.1136,222091.9);(196.1064,6746.3);}
Prostaglandin D2	C20H32O5	41598-07-6	CCCCCC(C=CC1C(C(CC1=O)O)CC=CCCCC(=O)O)	+	{(121.0667,2730.2); (135.0823,1499.6);(351.2185,9177.2);}
9-HODE	C18H32O3	98524-19-7	CCCCCC=CC=CC(CCCCCCCC(=O)O)O	+	{(123.1193,14822.8); (125.0986,23147.5);(277.1678,13643.2);(295.22911051823.5);}

Figure 2: Example 2

Ionization mode	Precursor	Product	Dwell time (sec)	Metabolite name	collision energy (eV)	Chemical formula
-	87	43	3	pyruvate	-14	C3H4O3
-	89	43.2	3	lactate	-16	C3H6O3
-	115	71	3	fumarate	-13	C4H4O4

Appendix C.Elements Scoring Algorithms

Identifications in Scaffold Elements are made by comparing the set of features in each MS1 peak group to entries in one or more spectral libraries. For each consensus MS1 peak group, all consensus features whose m/z values match library entries are noted. A series of scores is computed to assess the likelihood that the match is correct. These include isotopic distribution, mass accuracy, and MS2 similarity (if both the library and the sample have MS2 information). These scores are assessed on the individual features in each sample, and the highest score of each type is assigned to the consensus feature. An overall analyte ID score is computed for each potential identification of an MS1 peak group by computing a weighted average of these scores and combining this with a score based on the number of ions in the MS1 peak group which are explained by this identification.

Identification of In-source Fragments

If the option Match to In-source fragments in the Adducts tab is selected, all features that have not matched to any user-provided adduct for an analyte are compared to the m/zs observed in all library MS2 spectra. MS2 spectral matches with an intensity greater than or equal to a user-provided threshold intensity are annotated as in-source fragments. If multiple in-source fragment annotations can be made (ie, multiple MS2 spectra contain the same fragment peak), only the highest intensity peak is retained.

Score Calculations

Analyte ID Score

An ID Score is computed by combining the following scores in a weighted average:

- “Mass Accuracy Score” on page 214
- “MS2 Score” on page 216
- “Isotopic Distribution Score” on page 214

A score that is computed but not included in the ID Score calculation is the

- “XIC Score” on page 217

A score based on the number of annotated ions in the MS1 peak group is factored into the ID Score:

- “MS1 Annotation Score” on page 217

A penalty may be assessed for missing information.

Individual Scores

Conceptually, the individual scores represent the following:

- Mass Accuracy Score: How close the observed mass is to the theoretically expected mass

- Isotopic Distribution Score: How well the higher-order isotopic feature intensities match what is theoretically expected
- MS2 Score: How well the observed MS2 feature masses match an experimentally derived library entry
- XIC Score: How smooth the XIC curve is between adjacent RT points.

Mass Accuracy Score

The mass accuracy score is a function of the user's specific Parent Mass Tolerance.

Feature - library matches with no difference in mass receive a mass accuracy score of 1.0, while those matches with a difference in mass exactly at the tolerance limit score a mass accuracy score of 0.5.

Intermediate matches are scored based on the following linear equation:

$$\text{Mass Accuracy Score} = \frac{-0.5 \cdot \text{Delta Mass (AMU)}}{\text{Precursor Mass Tolerance(AMU)}} + 1$$

It is not possible to achieve a mass accuracy score of less than 0.5

Isotopic Distribution Score

The isotopic distribution score is based on a comparison of the relative abundance of observed features within an **isotopic feature cluster** to the theoretically expected abundance, based on the molecular formula of the matched analyte.

An isotopic feature cluster contains a monoisotopic feature [M+0], as well as (possibly), a feature with one extra neutron [M+1] and a feature with two extra neutrons [M+2]. If neither an [M+1] nor [M+2] feature is discovered, the isotopic distribution score is not assessed. If the matched analyte is an unknown, the molecular formula is not known, so it is not possible to generate a theoretical distribution with which to compare, so the isotopic distribution score is not assessed.

To generate the theoretical relative abundances of the [M+0], [M+1], and [M+2] features for a given analyte, we use PNNL's isotopic distribution calculator.

We compute the two sets of observed relative intensities, using two different quantities:

1. Area under XIC curve (trapezoidal approximation)
2. Max intensity of raw (m/z, RT, I) signal points

For each of these sets of relative intensities, the [M+0], [M+1], and [M+2] intensities are all divided by the relative intensity of the [M+0] feature, so the [M+0] feature always has a value of 1.0.

For each set of observed relative intensities (area under XIC curve and max intensity of raw signal), we perform the following normalization of the [M+1] and [M+2] relative intensities, for theoretical Th and observed Ob:

Theoretical normalization:

$$[M+1]_{\text{Th,norm}} = \frac{[M+1]_{\text{Th}}}{\sqrt{[M+1]_{\text{Th}}^{2} + [M+2]_{\text{Th}}^{2}}}$$

$$[M+2]_{\text{Th,norm}} = \frac{[M+2]_{\text{Th}}}{\sqrt{[M+1]_{\text{Th}}^{2} + [M+2]_{\text{Th}}^{2}}}$$

Observed normalization:

$$M+1_{\text{Ob,norm}} = \frac{[M+1]_{\text{Ob}}}{\sqrt{[M+1]_{\text{Ob}}^{2} + [M+2]_{\text{Ob}}^{2}}}$$

$$[M+2]_{\text{Ob,norm}} = \frac{[M+2]_{\text{Ob}}}{\sqrt{[M+1]_{\text{Ob}}^{2} + [M+2]_{\text{Ob}}^{2}}}$$

Where “Th” stands for theoretical, “Ob” for observed, and “norm” for normalized.

The total intensities from the theoretical and observed intensities are also computed:

$$\text{Th_Total} = [M+1]_{\text{Th}} + [M+2]_{\text{Th}}$$

$$\text{Ob_Total} = [M+1]_{\text{Ob}} + [M+2]_{\text{Ob}}$$

For each of the observed sets of relative intensities, we first compute the **Cosine Score**:

$$\text{Cosine_Score} = [M+1]_{\text{Th,norm}} \cdot [M+1]_{\text{Ob,norm}} + [M+2]_{\text{Th,norm}} \cdot [M+2]_{\text{Ob,norm}}$$

Finally, we normalize this cosine score based on the total intensity observed:

$$\text{Isotopic Distribution Score} = \frac{Min(\text{Th_Total, Ob_Total})}{Max(\text{Th_Total, Ob_Total})} \cdot \text{Cosine_Score}$$

The above isotopic distribution score is computed for each of the two sets of relative intensities and the maximum is taken to be the overall isotopic distribution score for the feature-analyte

Note also that both values are displayed in the Analyte View's isotopic distribution plot.

MS2 Score

Within a spectral library, it is possible that a given analyte has multiple **Analyte Records**. An **Analyte Record** is a record of a specific experiment that was performed. When comparing a feature versus an analyte, there may be multiple analyte records, each with disparate MS2s (but only one MS2 per analyte record), and multiple MS2 scans per feature. To determine the overall MS2 score, we compare every analyte record to every MS2 scan and take the highest-scoring combination of MS2 scan-analyte record MS2.

Each MS2 Spectrum consists of an array of (mass, intensity)(*m,I)* values. These are the fragment masses. In some cases, the precursor may be included in this array. If so, it is identified and removed from further processing.

Each fragment intensity I_i from each spectrum is normalized according to the following sum-of-squares normalization formula:

$$Inorm, i = \frac{Ii}{\sqrt{I_1^{\,2} + I_2^{\,2} + \ldots + I_n^{\,2}}}$$

Where there are n(m,I) fragment mass-intensity values per MS2 Scan (applies to both the MS2s from the experimental data as well as MS2s associated with analyte records).

Based on the user-supplied MS2 tolerance threshold, fragment masses from the analyte record MS2 scan are associated with fragment masses from the experimental data MS2 scan, where mass fragments from the two scans are associated together in **pairs** based on the difference in fragment mass. If more than one fragment mass from an experimental data MS2 scan could be associated with an analyte record scan, the fragment association with the smaller mass difference is given preference. Once a fragment mass from one MS2 scan is in a pair, it may not pair with any other fragment mass.

The normalized intensities of the associated fragment masses are summed in a cosine product between the theoretical (analyte record MS2 scan) and observed (experimental data

MS2 scan):

$$\text{MS2 Score} = I_{\text{norm,1,Th}} \cdot I_{\text{norm,1,Ob}} + I_{\text{norm,2,Th}} \cdot I_{\text{norm,2,Ob}} + \ldots + I_{\text{norm,n,Th}} \cdot I_{\text{norm,n,Ob}}$$

Where, in this case, the subscript numbers refer to **pairs** of associated fragment masses (not necessarily the total number of fragments per MS2 scan).

XIC Score

The Mass Accuracy Score, Isotopic Distribution Score, and MS2 Score all involve comparisons of experiment-specific data to spectral library data, while the XIC Score is computed entirely based on experiment-specific data (feature shape).For this reason, the XIC Score is not included with the other scores in the computation of an overall ID Score, though the XIC Score may still be useful in evaluating feature quality.

The **XIC Score** does not actually measure the quality of a match between a feature and an analyte, rather it measures the likelihood that a feature is actually derived from the presence of an ionized analyte and is not the result of machine noise.

Note that because an XIC Score has no bearing on whether or not a given feature is a particular analyte, we exclude the XIC Score from the **ID Score** computation.

In Elements, we use the **Zig Zag Index**, which is described in more detail in this publication:

> Zhang, W., & Zhao, P. X. (2014). Quality evaluation of extracted ion chromatograms and chromatographic features in liquid chromatography/ mass spectrometry-based metabolomics data. BMC Bioinformatics, 15(Suppl 11), S5. http://doi.org/10.1186/1471-2105-15-S11-S5

For our baseline value, we take the minimum intensity value used in the XIC.

The ZigZag Index naturally outputs a value from a nonlinear distribution, so in order to make this value comparable to our other scores, we adjust it in the following way:

$$XICScore = e^{(20\ln(0.4 \cdot ZigZagIndex))}$$

MS1 Annotation Score

First, the number of annotated MS1 ions is noted. This includes both user-specified adducts and In-source fragments of different mz values (for example, ISF #1 at mz=100 and ISF #2 at mz=200 count as two annotated MS1 ions). An MS1 peak group score is computed as follows:
1 annotated ion: MS1 peak group score of 0.70
2 annotated ions: MS1 peak group score of 0.90
3 annotated ions: MS1 peak group score of 0.95

4 annotated ions: MS1 peak group score of 0.97
5 annotated ions: MS1 peak group score of 0.98
6 – 10 annotated ions: MS1 peak group score of 0.99 + 0.002* (numAnnotatedIons – 6)
11+ annotated ions: MS1 peak group score of 1.0

ID Score

The ID Score communicates the likelihood that a given analyte was found in the experiment.

The Mass Accuracy Score, the Isotopic Distribution Score and the MS2 Score are all helpful metrics in assessing the likelihood that a single feature found in a single sample is a given analyte. However, to provide an overall, experiment-wide likelihood of an analyte being present, we must combine:

1. Computed Mass Accuracy Score, Isotopic Distribution Score and MS2 Score values

2. Consideration of the above scores computed from different samples

3. Consideration of the above scores computed from different feature types (i.e. different adducts) - the MS1 Annotation Score.

All individual features associated with an analyte are identified, and the Mass Accuracy Score, Isotopic Distribution Score, MS2 Scores and MS1 Annotation Score are all assessed, if possible.

The highest Mass Accuracy Score, Isotopic Distribution Score and MS2 Scores are identified. All known analytes will have a Mass Accuracy Score, however they may or may not have an Isotopic Distribution Score and/or an MS2 Score, depending on whether or not the experiment contains MS2 data and whether the higher-order isotopic features could be identified for any monoisotopic features.

A Penalty is assessed if no MS2 score is present for any of the ion forms in an MS1 peak group.

Penalty Calculation

A penalty of 0.05 is assessed if there is no MS2 spectral match for any of the features in an MS1 peak group and there are fewer than 3 ion forms in the MS1 peak group. If there are 3 or more ions, the penalty is 0.

A weighted average of the highest Mass Accuracy, Isotopic Distribution and MS2 scores is computed, however there are four cases to be considered, depending on missing information:

- **Isotopic Distribution and MS2 Scores present**

$$\text{ID Score} = 1/3 \cdot \text{MS2 Score} + 2/9 \cdot \text{Isotopic Distribution Score} + 1/9 \cdot \text{Mass Accuracy Score} + 1/3 \cdot \text{MS1 Annotation Score}$$

- **MS2 Score missing, Isotopic Distribution score present**

$$\text{ID Score} = 1/3 \cdot \text{Isotopic Distribution Score} + 1/6 \cdot \text{Mass Accuracy Score} - \text{Penalty} + 1/3 \cdot \text{MS1 Annotation Score}$$

- **MS2 Score present, Isotopic Distribution Score missing**

$$\text{ID Score} = 3/7 \cdot \text{MS2 Score} + 1/7 \cdot \text{Mass Accuracy Score} + 3/7 \cdot \text{MS1 Annotation Score}$$

- **Both MS2 Score and Isotopic Distribution Score missing**

$$= 1/4 \cdot \text{Mass Accuracy Score} - \text{Penalty} + 3/4 \cdot \text{MS1 Annotation Score}$$

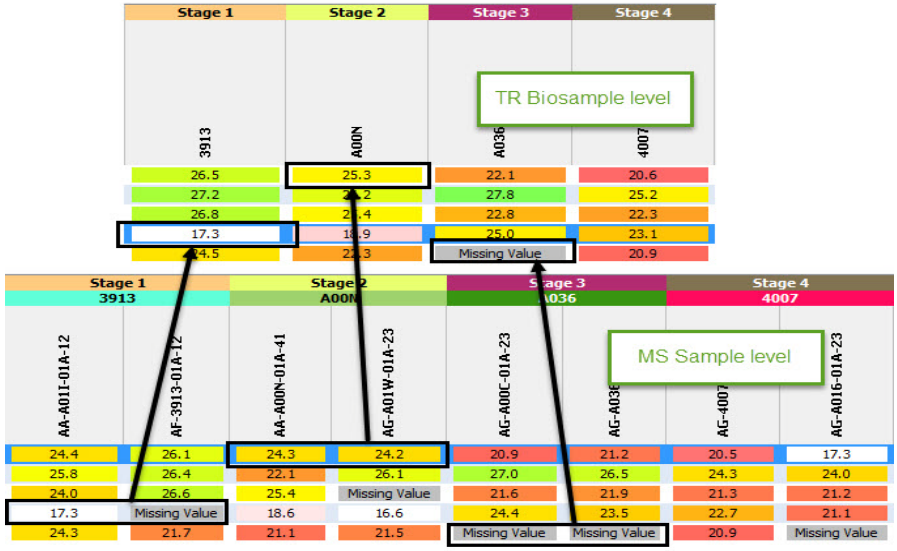
Appendix D.Rolling up Values

As can be seen when Display Type: **Log_{10} Precursor Intensity** is selected and the level of summarization is changed, Scaffold Elements rolls up the values listed in a row, or group of analytes, to a higher level of summarization, in two distinct ways depending on whether the MS Samples are regarded as fractions of a single technical replicate.

At the MS sample level, for each analyte, Scaffold Elements reports the Log_{10} of the sum of the maximum precursor intensity values for each ion feature in the MS sample. When a tecnicak replicate includes more than one MS sample, Scaffold Elements rolls up the values from the MS sample level to the technical replicate level by summing all of the precursor intensity values in the group of MS samples and then calculating the Log_{10} of the result. This is the value reported in the samples table at the technical replicate level. see Figure 3.

When MS samples do not include data for a particular analyte but that analyte is found in another sample, Scaffold Elements labels the corresponding cells in the Samples table with the tag “Missing Value”. When rolling up a group of MS samples that includes some values that are missing and some not to the technical replicate level, the program ignores the missing values and assigns a value which is the log_{10} of the sum of the existing intensities. However, if all the values are missing, “Missing Value” is assigned to the group as shown in Figure 3.

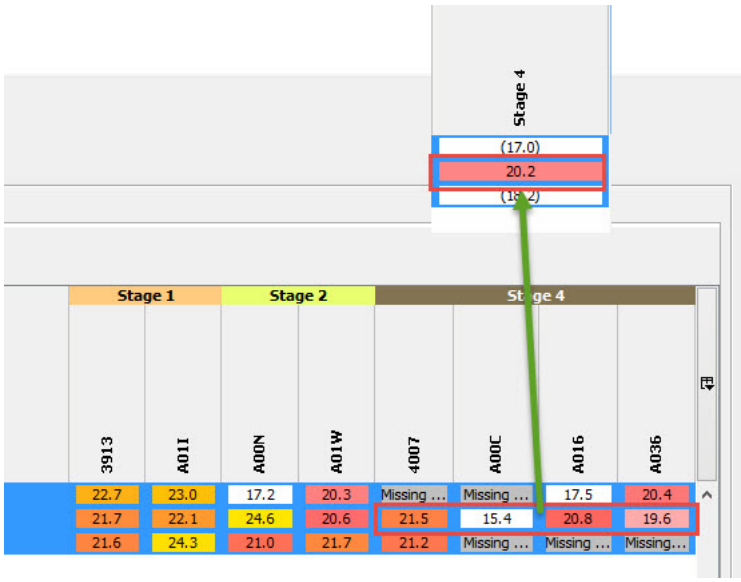
Figure 3: Rolling up values to a TR Level of summarization



When the summarization level is switched to a level higher than the TR Level, Scaffold Elements rolls up the values using the median of the Log_{10} Intensity data included in an attribute group, see Figure 4 The picture shows the log_{10} Precursor Intensity values of four biosamples included in the attribute group Stage 4 and the log_{10} Precursor Intensity rolled up value when the higher summarization level Stage is selected.The value corresponds to the median of the values appearing when the immediately lower level of summarization is

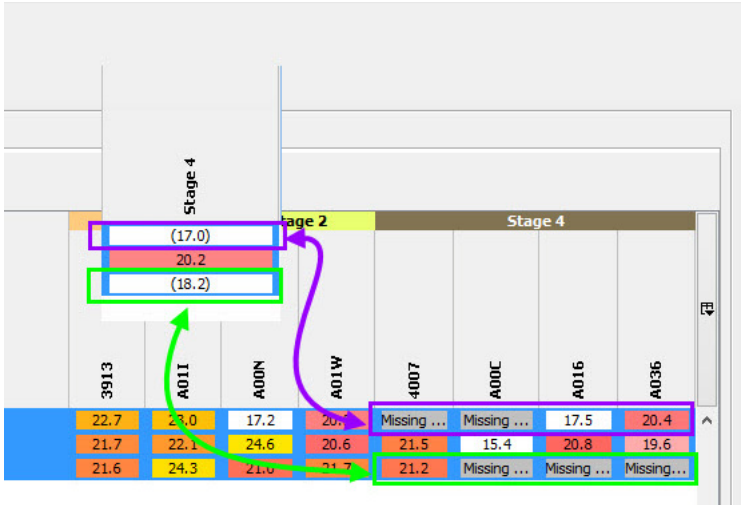
chosen, which, in this particular example, is called Biosample level.

Figure 4: Rolling up values to a higher summarization level



If fifty percent or more of the values belonging to a group are missing, the QRILC method of Missing Value Imputation is used to compute the proper rolled up values, and in the Samples Table the values so calculated are shown in parentheses, see Figure 5.

Figure 5: Rolling up of values to a higher summarization level with missing values



Missing Value Imputation

Missing values affect various computations in Scaffold Elements. In order to be able to roll up values to higher levels of summarization, to perform statistical testing and to perform Principle Component Analysis, it may be desirable to impute values when no measurement has been obtained. On the assumption that missing values in Scaffold Elements are generally

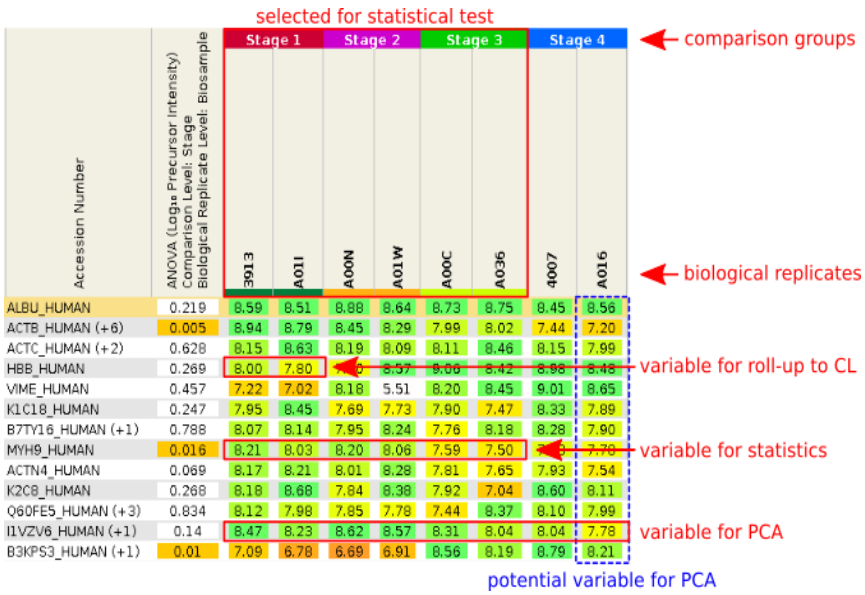
a result of either absence of a compound or presence at very low intensity, Scaffold Elements uses the method of “Quantile Regression for Imputation of Left-Censored data” (QRILC)[1].

QRILC imputes values by drawing from a truncated normal distribution whose parameters are estimated from the observed (non-missing) values and the number of missing and observed values, assuming that missing values are lower than any observed value.

When rolling up values where over 50% are missing, the mean estimated by QRILC is used as the rolled up value (which is treated as partially-missing, indicated by surrounding the value in parentheses).

When computing statistics where any number of values are missing (provided there are at least two observed values), QRILC is used to estimate a distribution for values across all comparison groups, and replaces missing values with values drawn from the estimated distribution, truncated at the proportion of missing values (so that if X% of values are missing, the imputed values fall below the Xth percentile of the estimated distribution).

PCA does the same, but across values for all biological replicates in the file.



Accession Number	ANOVA (Log₂ Precursor Intensity) Comparison Level: Stage Biological Replicate Level: Biosample	3913	A011	A00N	A01W	A00C	A036	4007	A016
ALBU_HUMAN	0.219	8.59	8.51	8.88	8.64	8.73	8.75	8.45	8.56
ACTB_HUMAN (+6)	0.005	8.94	8.79	8.45	8.29	7.99	8.02	7.44	7.20
ACTC_HUMAN (+2)	0.628	8.15	8.63	8.19	8.09	8.11	8.46	8.15	7.99
HBB_HUMAN	0.269	8.00	7.80	[illegible]	8.57	9.06	8.42	8.98	8.48
VIME_HUMAN	0.457	7.22	7.02	8.18	5.51	8.20	8.45	9.01	8.65
K1C18_HUMAN	0.247	7.95	8.45	7.69	7.73	7.90	7.47	8.33	7.89
B7TY16_HUMAN (+1)	0.788	8.07	8.14	7.95	8.24	7.76	8.18	8.28	7.90
MYH9_HUMAN	0.016	8.21	8.03	8.20	8.06	7.59	7.50	[illegible]	7.78
ACTN4_HUMAN	0.069	8.17	8.21	8.01	8.28	7.81	7.65	7.93	7.54
K2C8_HUMAN	0.268	8.18	8.68	7.84	8.38	7.92	7.04	8.60	8.11
Q60FE5_HUMAN (+3)	0.834	8.12	7.98	7.85	7.78	7.44	8.37	8.10	7.99
I1VZV6_HUMAN (+1)	0.14	8.47	8.23	8.62	8.57	8.31	8.04	8.04	7.78
B3KPS3_HUMAN (+1)	0.01	7.09	6.78	6.69	6.91	8.56	8.19	8.79	8.21

Parameter estimation is done by a linear regression between the quantiles of the observed data and quantiles of a truncated normal distribution (if X% of values are missing, the distribution used for fitting is truncated below the Xth percentile, so that the 0th percentile of the observed data is matched to the Xth percentile of the distribution, and so on up to the 95th percentile). This gives an estimated mean and standard deviation for the distribution of values. For more information, see the documentation for the R package "imputeLCMD v2.0" which provides the reference implementation for QRILC.

1. Cosmin Lazar (2015): imputeLCMD v2.0". R package version 2.0

Appendix E.Agglomerative Point Clustering Feature Finding Algorithm

Elements' feature finding algorithm relies on agglomerative clustering of raw (m/z, RT, intensity) data points into collections of points, then performs a series of post-processing steps on these collections of raw data points to produce fully formed "features". In this discussion, we will distinguish "peaks" from "features" with the implication that a peak is a 2D object, while a feature is 3D.

Individual data points recorded from an LC-MS instrument are characterized by three dimensional measurements: mass-to-charge ratio (m/z), retention time (RT) and intensity (I). LC-MS data is inherently discrete (instruments have physical detection and resolution limits), and so individual maxima appear as clusters of raw (m/z, RT, I) data points.

Elements' feature finding strategy seeks to first organize the raw (m/z, RT, I) data points into clusters of points and derive a single (m/z, RT, I) value for each cluster. The (m/z, RT, I) values determined from these point clusters constitute the features or peaks, which from then on will be treated as single (m/z, RT, I) points in all spectral library matching and analyte association steps.

Polarity Grouping

Mass spectrometers can be configured to detect the presence of either positively charged ions or negatively charged ions, but not both at the same time. Operating a mass spectrometer in positive mode produces a set of (m/z, RT, I) data points that all derive from positively charged species; operating a mass spectrometer in negative mode produces a set of (m/z, RT, I) data points that all derive from negatively charged species.

However, it is possible to operate a mass spectrometer in "Mixed mode", alternatively scanning for positively and negatively charged species. In this case, the landscape of (m/z, RT, I) data points in a given sample will derive from a mixture of positively and negatively charged chemical species.

In the "Polarity Grouping" step, all data associated with positive scans are organized together, and all data from negative scans are organized together. Unless the instrument was run in mixed mode, this will result in all of the scans being associated with positive or with negative. If the instrument was run in mixed mode, the scans will be divided appropriately based on polarity, and feature finding will be performed on both the positive and the negative scan sets.

Mass spectrometers generate a large amount of (m/z, RT, I) data points. For this reason, it is common to condense the data into smaller sets of points using a process called "centroiding" (see appendix B). Within a single RT scan, sequential sets of points that form a concave-down curve shape are collapsed into a single point (the centroid of the points comprising the curve). The option to output centroid data is often specified at the time of running a sample through the mass spectrometer Centroiding offers the trade off of generating smaller output files, with the downside of obscuring the pure, raw data points, which can make feature finding more challenging. Because of the nature of centroiding, in our feature finding algorithm, we must handle these two types of feature data differently. It is therefore important

for the algorithm to establish if a dataset is centroid or not (which is commonly also called "profile mode").

The raw data files we read in Elements for Metabolomics contain metadata describing if the machine was operated in centroid or profile mode. We determine this setting from the raw data, and adjust our feature finding approach accordingly.

Determine global m/z and RT grid distances (PROFILE MODE only)

Data that is generated in profile mode tends to show regularity in the distance in (m/z) and (RT) between data points, as though the raw data points existed on a grid.Segment dataset into windows

An effort is made to determine the grid distances (the inherent spacing between raw data points) at the level of the whole data set. This inherent m/z and RT spacing may vary depending on the particular segment of the raw file we are examining, however we estimate a "global" value to help determine if our local estimates are close. If, for any local window, we find the m/z and RT grid constants to vary too widely from the global constants, we will use the global constants.

Determine local m/z and RT grid distances within each window

For each window, the m/z and RT grid constants are re-calculated, as a given window may demonstrate significantly different m/z and RT grid distances than the global.

If the dataset was collected in PROFILE mode, the m/z grid distance is determined in the same way that the global m/z grid distance constant is determined, see "Determine global m/z and RT grid distances (PROFILE MODE only)" on page 167, considering only the data points in the window.

If the dataset was collected in CENTROID mode, the 5 most intense points in the window are collected, and sorted in order of retention time (RT). The distance in m/z between sequential points is determined, and the m/z grid distance for the window is taken to be either 10 times the maximum sequential distance or 0.008 AMU, whichever is smaller.

Organize all (m/z, RT, I) points in windows into point clusters

All (m/z, RT, I) data points in a given window are ordered by decreasing intensity (most intense point first, least intense point last).

1. The most intense point in the window is assigned to a new point cluster.
2. (m/z, RT, I) points from the window are progressively considered and are either:
   - (A) added to an existing point cluster, or
   - (B) considered as the definition of a new point cluster.
3. This process continues until all points in the window have been considered and assigned to clusters.

The criteria that determines if a point is "close enough" in (m/z, RT) space to an existing

cluster to join that cluster, or if it must instead define a new cluster, goes as follows:

CENTROIDED MODE:

- (m/z, RT, I) points that are more than 1 m/z grid units away from the closest m/z edge of a cluster are too far away to join that cluster.
- (m/z, RT, I) points that are more than 2 RT grid units away from the closest RT edge of a cluster are too far away to join that cluster.

Furthermore, once the 8th point is added to a cluster, this triggers a computation to determine the average intra-cluster m/z distance, considering point in sequential RT. The average m/z distance between sequential points of these first 8 data points is noted as the “intra-cluster m/z distance”. (m/z, RT, I) points beyond the 8th point must satisfy the additional constraint that these points may not be more than 4 times the intra-cluster m/z distance to join the cluster.

PROFILE MODE:

- (m/z, RT, I) points that are more than 3 m/z grid units away from the closest m/z edge of a cluster are too far away to join that cluster.
- (m/z, RT, I) points that are more than 5 RT grid units away from the closest RT edge of a cluster are too far away to join that cluster.

Multiple Cluster possibilities

A point may be close enough to join multiple clusters in the window to join any of them. In this case, a series of tie-break criteria are employed to determine which of the clusters the point is assigned to. Each (m/z, RT, I) point may belong to only one point cluster, and all points in the window must be assigned to point clusters.

The criteria to break ties between different candidate clusters a (m/z, RT, I) point might join are as follows:

1. Consider the distance in m/z of the point and the candidate cluster. If the point is located within the m/z bounds of the candidate cluster, the m/z distance is 0. The point joins to candidate cluster with the smallest distance.
2. If the m/z distance is the same among candidates, consider the distance in RT of the point and the candidate cluster. If the point is located within the RT bounds of the candidate cluster, the RT distance is 0. The point joins to the candidate cluster with the smallest distance.
3. If the m/z distance and the RT distance is the same among candidates, consider the distance in m/z between the point and the maximum (m/z, RT, I) point located within the candidate cluster. The point joins to the candidate cluster with the smallest distance.
4. If all distances are still the same, consider the distance in RT between the point and the maximum (m/z, RT, I) point located within the candidate cluster. The point joins to the candidate cluster with the smallest distance.

5. If all four of these distances are the same, consider the intensity of the maximum (m/z, RT, I) point located within the candidate cluster. The point joins to the candidate cluster with the higher intensity.
6. If all distances are the same, and the maximum (m/z, RT, I) point intensity is the same across multiple candidate clusters, the point joins the candidate cluster with the maximum (m/z, RT, I) point with the higher RT.

Merge point clusters within each window together

The set of clusters generated as described in section Organize all (m/z, RT, I) points in windows into point clusters within each window are progressively merged together, until no more merging is possible.

If the feature data was collected in either CENTROID mode or PROFILE mode, two clusters are merged together if the following criteria are satisfied:

1. The two clusters are assessed, and the one with the higher RT is designated the “higher RT” cluster, the one with the lower RT is designated the “lower RT” cluster.
2. The two clusters are merged together if the lower RT boundary of the “higher RT” cluster is located within 4 RT grid units of the upper RT boundary of the “lower RT” cluster, AND one of the following criteria are met:
   - The m/z bounds of one cluster are within the m/z bounds of the other cluster.
   - The lower m/z edge of the higher RT cluster is within 2 m/z grid units of the lower m/z edge of the lower RT cluster.
   - The higher m/z edge of the higher RT cluster is within 2 m/z grid units of the higher m/z edge of the higher RT cluster.
   - The intensity-weighted centroid m/z value of the higher RT cluster is located within the m/z bounds of the lower RT cluster, and the intensity-weighted centroid m/z value of the lower RT cluster is located within the m/z bounds of the higher RT cluster.

   OR

   - The lower RT boundary of the “higher RT” cluster is located within 20 RT grid units of the upper RT boundary of the “lower RT” cluster,
   - AND the m/z edges of the two clusters are very close: the lower m/z edge of the higher RT cluster is within 2 m/z grid units of the lower m/z edge of the lower RT cluster, and the higher m/z edge of the higher RT cluster is within 2 m/z grid units of the higher m/z edge of the higher RT cluster.
3. If the feature data was collected in PROFILE mode, two clusters may also be merged together if three of the four (m/z, RT) edges (min Mz, max Mz, min RT, max RT) of one cluster are located inside the other cluster (merge by engulfing).

Assign clusters as “edge-touchers” or “in-bounds”

Clusters that touch one or more (m/z, RT) edges of the window from which they derive are designated as “edge-touchers”, while clusters that do not touch a m/z or RT edge are designated as “in-bounds”. In-bounds clusters are set aside, and all edge-touchers from all windows are collected, along with information about the window from which they derive.

Merge all edge-toucher clusters (cross-window merging)

The windowing step artificially cuts some clusters into pieces, and point-clustering is performed on whatever signal ends up in each window. The purpose of the cross-window merging step is to properly assemble all artificially cut-apart clusters back together.

The merging step works by using a “propagating waveform” algorithm to compare edge-touchers from a given window to edge-touchers in the window to the right (same RT block unit, one additional m/z block unit) and to the edge-touchers in the window underneath (one additional RT block unit, same m/z block unit), and merge clusters together when it is clear that they belong together.

- **PROFILE Mode**

  If the data is run in profile mode, Scaffold Elements applies to the dataset a series of filters as follows:

  minNumPoints, minNumTotalScans minNumSequentialScans, zeroMzWidth.

  Please see “Process clusters” on page 171 for a more detailed description of these filters.

- **CENTROID Mode**

  If the data is run in centroid mode no filtering is applied.

After the propagating waveform approach is completed, in rare cases some (m/z, RT, I) data points may belong to multiple point clusters, so all point clusters where (m/z, RT, I) data points belong to multiple point clusters are merged together.

Now all (m/z, RT, I) point clusters are guaranteed to contain a unique set (m/z, RT, I) data points.

Process clusters

Many of the point-clusters that emerge from the in-bounds clusters procedure see “Assign clusters as “edge-touchers” or “in-bounds”” on page 170 and cross-window merged edge-touchers procedure, see “Merge all edge-toucher clusters (cross-window merging)” on page 170, are not real features. Either they are spurious clusters of noise, or clusters of multiple features, or a combination of noise and one or more real features.

To extract the real features in the set of point clusters, a series of processing steps is undergone involving both splitting and filtering point clusters to a final set of feature-worthy point clusters. Splitting steps are performed first, followed by filtering steps.

Splitting Steps refer to the division of a single point cluster into one or more smaller point clusters, where no raw data may exist in more than one split point cluster, and all data points

in a starting cluster must be present among the split clusters (data points are not discarded in this step). Splitting steps are applied progressively, until the set of clusters cannot be split any further (based on the criteria that would entail a split). For all splitting steps, a single point cluster is input, and one or more “split” clusters are output.

Filtering Steps refer to examination of a given point cluster, and deciding if that point cluster is a valid feature. If not, the point cluster is removed from further consideration.

Depending on whether the feature data was generated in profile mode or centroid mode, a different processing schedule is employed:

PROFILE Mode

Splitting Steps applied:

- MassTraceLocalMax (rectangle-smoothed, with window size of 5)
- MassTraceGaps (2 grid units)
- XIClocalMax (3x triangle-smoothed, with window sizes of 7, 11, and 21)
- XICGaps

Filtering Steps applied:

- minNumPoints
- minNumTotalScans
- minNumSequentialScans
- zeroMzWidth
- requireXICmax (3x triangle-smoothed, with window sizes of 7, 11, and 21)

CENTROID Mode

Splitting Steps applied:

- XIClocalMax (3x triangle-smoothed, with window sizes of 7, 11, and 21)
- XICGaps

Filtering Steps applied:

- minNumTotalScans
- minNumSequentialScans
- requireXICmax (3x triangle-smoothed, with window sizes of 7, 11, and 21)

What follows is a detailed description of each of the steps within the two processing schedules, note that we heavily rely on the XIC and mass trace, see Feature, Mass Trace and XIC, in the splitting and filtering steps that follow.

Splitting Steps

MassTraceLocalMax

For a given point cluster, the mass trace is assessed. The mass trace is smoothed using a rectangle-smoothing filter, with a window size of 5 points. The resulting waveform is assessed for maxima and minima using a 3-point method: If a given mass trace point p has a value higher than both the p-1 and p+1th points, the point p is a local maximum. If a point p has a value lower than both the p-1th and the p+1th point, the point is a local minimum.

Once all local maxima and minima have been identified, the intensity ratio between the local maximum and its two surrounding local minima is assessed. If both the intensity ratio of the local maximum to its preceding local minimum and the intensity ratio of the local maximum to its subsequent local minimum is greater than or equal to 2, the point cluster is split at the subsequent local minimum point.

This process continues until all local maxima and local minima have been assessed.

MassTraceGaps

For a given point cluster, the mass trace is assessed. If no intensity is recorded in more than 2 sequential mz grid units, a gap is said to have occurred in the mass trace. The point cluster is split at the point of the gap.

The process continues until all mass trace points have been assessed.

XIClocalMax

The following procedure is applied 3 times, each with a different window size: 7, 11, and 21 points.

For a given point cluster, the XIC is assessed. The XIC is smoothed with a triangle filter, with the above window sizes (7, 11, and 21 points). The resulting waveform is assessed for maxima and minima using a 5-point method:

- If a given XIC point *p* has a value higher than both the *p-1th* and the *p+1th* point, and the *p-1th* point is higher than the *p-2nd* point, the point p is a local maximum.
- If an XIC point p has a value lower than both the p-1th and p+1th point, and the p-1th point is lower than the p-2nd point, and the p+1th point is lower than the p+2nd point, the XIC point p is a local minimum.

Once all local maxima and minima have been identified, the intensity ratio between the local maximum and its two surrounding minima is assessed. If both the intensity ratio of the local maximum to its preceding local minimum and the intensity ratio of the local maximum to its subsequent local minimum is greater than or equal to 5, the point cluster is split at the subsequent local minimum point.

This process continues until all local maxima and local minima have been assessed.

XICGaps

For a given point cluster, the XIC is assessed. If no intensity is recorded in more than 5 sequential RT grid units, a gap is said to have occurred in the XIC. The point cluster is split

at the point of the gap. The process continues until all mass trace points have been assessed.

Filtering Steps

minNumPoints

If a cluster contains fewer than 20 raw (m/z, RT, I) data points, the cluster is filtered out.

minNumTotalScans

If a cluster contains fewer than 4 RT scans, the cluster is filtered out. In other words, if only 3 RT values are represented amongst all raw (m/z, RT, I) points, the cluster is filtered out.

minNumSequentialScans

If a cluster contains fewer than 3 sequential RT scans, the cluster is filtered out.

zeroMzWidth

If a cluster contains only one (m/z) value for each RT value represented, the cluster is filtered out.

requireXICmax

For a given point cluster, the XIC is assessed. The XIC is smoothed with a triangle filter, with the above window sizes (7, 11, and 21 points). The resulting waveform is assessed for maxima and minima using a 5-point method: If a given XIC point p has a value higher both the p-1th and the p+1th point, and the p-1th point is higher than the p-2nd point, and the p+1th point is higher than the p+2nd point, the point p is a local maximum. If an XIC point p has a value lower than both the p-1th and p+1th point, and the p-1th point is lower than the p-2nd point, and the p+1th point is lower than the p+2nd point, the XIC point p is a local minimum.

Once all local maxima have been identified, the intensity ratio between the local maximum and the first and last XIC point is assessed. If both the intensity ratio of the local maximum to the first point in the XIC and the intensity ratio of the local maximum to the last point in the XIC is greater than or equal to 6, the point cluster is retained.

Only one of the 3 smoothing filter window sizes need satisfy the above criterion for the cluster to be retained. If no local maximum with a sufficiently high XIC to bounds ratio is discovered for all 3 smoothing filter window sizes, the cluster is discarded.

Appendix F.Isotopic Clustering

If two or more atomic species contain the same number of protons and electrons, but vary in the number of neutrons contained in their respective nuclei, they are called "isotopes". The variance in neutron number causes a variation in the atomic weight, and correspondingly, molecules that contain different isotopic forms of the same atomic elements will vary in weight. The relative abundance of different isotopic forms of all naturally occurring elements is approximately constant, which allows us to predict the relative abundance of different atomic weights of the same molecule that might occur if the isotopic form of each component atom were selected randomly based on its known relative abundance.

In Elements, features are organized into **isotopic clusters**, each of which contains the monoisotopic feature, also known as the **[M+0]**, and, if they exist, the **[M+1] (**which contains one additional neutron) and **[M+2]** (with two extra neutrons) features. Elements also searches for the **[M+3]** form if the **[M+1]** and **[M+2]** forms have been detected. The **[M+1], [M+2]** and **[M+3]** forms are then removed from the set of valid features.

Elements searches for isotopic peaks at predicted m/z offsets from the **[M+0]** peak, searching a region from the **minimum m/z of the[M+0] + offset** to the **maximum m/z of [M+0] + offset**. The offsets are calculated so that they will include isotopic peaks attributable to the presence of 13C, 15N, and 33S (for **[M+1]**), and 34S, 13C+15N, 13C+33S, 18O, 13C+13C (for **[M+2]**). The signal from all isotopic peaks detected in the searched region is combined to represent the total intensity of the **[M+1]** or **[M+2]** isotopes.

Isotopic clustering accomplishes three goals:

- Non-monoisotopic features are not searched against spectral libraries. This step is called **de-isotoping**.
- The observed relative abundance of the **[M+1]** and **[M+2]** features is compared to the theoretically predicted relative abundance of **[M+1]** and **[M+2]** features (based on the known relative abundance of naturally occurring elements). This comparison is incorporated into the scoring algorithm (please see “Isotopic Distribution Score” on page 214).
- The charge of the feature is deduced from the pattern of the isotopic distribution.

Isotopic Clustering Algorithm

This algorithm is applied to a single MS Sample, containing a large collection of raw (m/z, RT, Intensity) MS1 data points. If the data was collected in mixed mode, the raw MS1 data in the file is divided, with each ionization mode considered separately.

As a result of the feature finding step, all raw (m/z, RT, Intensity) data points deriving from a single ionization mode (either positive or negative) are organized into non-overlapping point clusters.

The complete set of point clusters is sorted according to the intensity of the most intense (m/z, RT, Intensity) point in the point cluster, so that the first entry in the list contains the point

cluster containing the most intense single raw data point in the file. This list of sorted point clusters is called the **available point cluster list**.

Point clusters from the **available point cluster list** are compared in order, with the assumption that the identified point cluster is associated with the monoisotopic form (also known as **[M+0]**). A cluster, while under investigation, is called the **query cluster**.

The **query cluster** is assumed to correspond to an ion with a charge number of either +1, +2, or +3 for positive mode data or -1, -2 or -3 for negative mode data.

All of the unassigned point clusters in the **available point cluster list** are compared to the **query cluster**.

For each of the charge numbers of 1, 2, and 3, an **[M+1]** and an **[M+2]** cluster are sought for the **query cluster** using the following process:

For [M+1]:

All unassigned features in the file that have an m/z value between the **minimum m/z bound of [M+0] + 0.9958/charge** and the **maximum m/z bound of [M+0] + 1.0042/charge**, with an **RT** value not more than 5 seconds from the **[M+0] RT** value are collected.

For [M+2]:

All unassigned features in the file that have an m/z value between the **minimum m/z bound of [M+0] + 1.9916/charge** and the **maximum m/z bound of [M+0] + 2.0084/charge** with an **RT** value that differs from the **[M+0] RT** by not more than 5 seconds are considered potential **[M+2]** peaks. To avoid misclassifying a feature that is actually the monoisotopic peak of a different analyte, if the difference between the m/z of a candidate **[M+2]** peak and that of the **[M+0]** peak is larger than the mass shift attributable to two C13 atoms (2.0067096756/charge) and is close (within 0.002977/charge) to the mass shift of two H atoms (2.0156006448/charge), the peak is not considered a true **[M+2]** and is not removed from the **available point cluster list**.

All **[M+1]** clusters are merged together, removed from the **available point cluster list** and assigned as the **[M+1]** for the **query cluster**. All **[M+2]** clusters are merged together, removed from the **available point cluster list** and assigned as the **[M+2]** for the **query cluster**.

The charge state of a feature is assigned according to the number of detected features associated with each charge number. If neither an **[M+1]** nor an **[M+2]** peak is detected for any charge number, the charge state is set to 1. Otherwise the charge number with the largest number of detected isotopic peaks is used. In case of a tie, the lower charge number is assigned.

If both an **[M+1]** and an **[M+2]** cluster are found in the **available point cluster list**, an attempt is made to find the **[M+3]** cluster from the set of available point clusters, using the matching procedure described above, but using bounds of **minimum m/z bound of [M+0] + 2.9874/charge** to **maximum m/z bound of [M+0] +3.0126/charge**, and an RT value not more than 5 seconds from the **[M+0] RT**. Identified **[M+3]** clusters are removed from the **available point cluster list**.

If an **[M+1]** or **[M+2]** cluster is not detected, that cluster is scheduled for reextraction.

The process of cluster assignment continues until all clusters in the **available point cluster list** have been assigned.

At this point, the reextraction step is carried out for all clusters scheduled for reextraction. This may require reextraction of the **[M+1]**, the **[M+2]**, or both the **[M+1]** and **[M+2]**.

Isotopic Reextraction:

Isotopic reextraction works by taking all signal at a predicted m/z offset from the **[M+0]** within the region from the **minimum m/z bound of [M+0] + offset/charge** to the **maximum m/z bound of [M+0] + offset/charge**. The offsets searched are the mass shifts of **13C**, **15N**, and **33S** (for **[M+1]**), and **34S**, **13C+15N**, **13C+33S**, **18O**, **13C+13C** (for **[M+2]**). The RT ranges used for **[M+1]** reextraction are the **[M+0] minRt** and **maxRt**. For **[M+2]** reextraction, if the **[M+1]** peak was present, the **[M+1] minRt** and **maxRt** are used, while if the **[M+1]** peak was not present (i.e. was also reextracted), then the **[M+0] minRt** and **maxRt** are used.

- If reextraction bounds would overlap with the bounds defined by an existing feature, the reextraction is not performed.
- Reextraction is performed with no noise thresholding of the raw data.

Appendix G.Forming Consensus MS1 peak groups

When all samples undergo similar chromatographic procedures, an analyte would be expected to elute at essentially the same time in each sample in a study. Further, in most cases, true coelution is rare, so coeluting ions are likely to be different ion forms of the same analyte. As a result, it can be helpful to group ions that elute together into MS1 peak groups before attempting to identify them. In Scaffold Elements, therefore, the important unit for identification and quantification is the consensus, or cross-sample, MS1 peak group.

The algorithm for forming Consensus MS1 peak groups in Elements consists of a series of steps:

Formation of Single-Sample MS1 Peak Groups

Retention Time Alignment

Consensus MS1 Peak Group Formation

Analyte Grouping and Clustering

Formation of Single-Sample MS1 Peak Groups

Initially, each sample is considered independently. First, features are grouped by retention time with other features with the same polarity. Next, if the sample was analyzed in mixed polarity mode, MS1 peak groups of opposing polarities with the same retention times are combined.

Retention Time Alignment

Retention time alignment is performed across samples using these single-sample MS1 peak groups. Consensus MS1 peak groups are formed using a high RT tolerance threshold and matching features from MS1 peak groups across samples.

A subset of this initial set of consensus MS1 peak groups is designated as “anchor spectra.” Anchor spectra must be featured in at least a certain portion of the samples. This portion is designated the **RT alignment spectrum min reproducibility** parameter and may be adjusted by the user via the Advanced Tab in the Workflow Dialog. The sample which contains the largest number of anchor spectra is selected as the reference sample. From the anchor spectra, a monotonic mapping is formed between each sample and the reference sample. Linear interpolation is used to align the remaining features.

Retention time alignment is only performed between samples with the same polarity. Positive samples may be aligned with other positive or with mixed mode samples. Similarly, negative samples may be aligned with other negative or with mixed mode samples. If mixed mode samples are present, they are preferred for use as reference samples in the RT alignment.

Consensus MS1 Peak Group Formation

Following retention time alignment, consensus (cross-sample) MS1 peak groups are carefully formed from the raw data. MS1 peak groups from all samples whose aligned retention times fall within a certain tolerance of each other are considered for inclusion in a consensus spectrum. A representative peak is selected from each single-sample spectrum. The representative peak must have approximately the same m/z value in each spectrum, and must be the most intense peak in the spectrum in at least one sample. In considering whether spectra should be joined, the representative peak is compared with respect to mass accuracy, isotopic distribution, peak shape and similarity of MS2 spectra. In addition, the most intense peak of each spectrum must be found in all others to which it joins, and the overall similarity of the MS1 peak groups is evaluated and must meet a threshold. No more than one MS1 peak group from each sample may be included in a consensus MS1.

After consensus MS1 peak groups have been formed using these strict criteria, the remaining single-sample MS1 peak groups are re-evaluated for inclusion in these consensus MS1's. This time, the spectrum need not contain the most intense peaks of the other spectra, and inclusion is based on the similarity (Jaccard index) of the MS1 peak groups.

Next, the individual features in each sample are organized into cross-sample consensus features based on MS2 spectral similarity, isotopic distribution scores, mass accuracy, peak shape similarity and proximity in aligned retention time. The set of all consensus MS1 peak groups are compared to this set of consensus features, to ensure that no consensus feature is assigned to different MS1 peak groups in different samples. If a feature had been inconsistently assigned, it is assigned to the consensus MS1 in which it appears in the largest number of samples and removed from all others. In case a feature matches more than one consensus MS1 in an equal number of samples, peak shape similarity is used as a tie-breaker.

If the experiment contains separate positive and negative samples, consensus MS1 peak groups are merged across charge states. Consensus MS1 peak groups are merged if there exists a feature pair between two consensus MS1 peak groups that could correspond to an {M+H] and {M-H] pair.

Filtering

The user may elect to retain only MS1 peak groups which contain at least one feature that has an MS2 spectrum. If the "Only retain analytes with MS2 data option" in the Search tab of the Workflow dialog is checked, the MS1 peak groups are filtered at this point.

Reextraction

Once consensus MS1 peak groups have been formed, if the user has checked the box selecting the parameter "Perform cross-sample feature reextraction", the program attempts to extract signal to replace missing values. In cases in which an analyte is found in some, but not all, samples, the program extracts the signal from the region at the aligned retention time corresponding to the elution time of the analyte in the other samples.

Analyte Grouping and Clustering

All ions assigned to the same consensus MS1 peak group are organized together into a single

analyte cluster. Analyte grouping, however, depends on the option selected for the “Treat each MS1 peak group as a single analyte” parameter in the Search tab of the Workflow dialog.

If “Treat each MS1 peak group as a single analyte” is checked, the ions are also organized into a single analyte group. This is the default behavior, because unless there is a reason to expect coeluting ions, it is generally the case that all of the ions that elute at the same retention time are different forms derived from the same analyte.

If “Treat each MS1 peak group as a single analyte” is not checked, then only analyte identifications made with identical ions are organized into the same analyte group. Unidentified coeluting ions are treated as separate analytes (if “Retain unknown analytes” is selected).

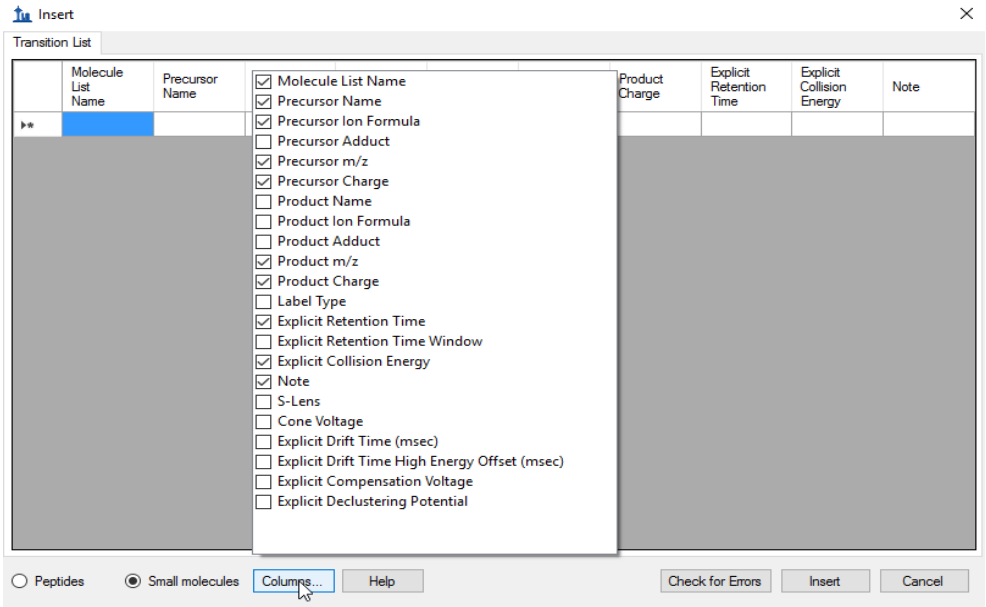
Appendix H.Exporting a Transition List to Skyline

Elements provides the ability to export a spectral library as a transition list which can then be used in Skyline [2].

Steps:

- Create a spectral library containing the identifications to be exported to Skyline.
- Open the library in the Library View, click “Export to Skyline” and save the file.
- In Skyline, select Edit>Insert>Transition List to open the Transition List dialog. Select “Small molecules” using the radio button at the bottom. Click “Columns” and check the following options:

Figure 6: Transition List column selection in Skyline

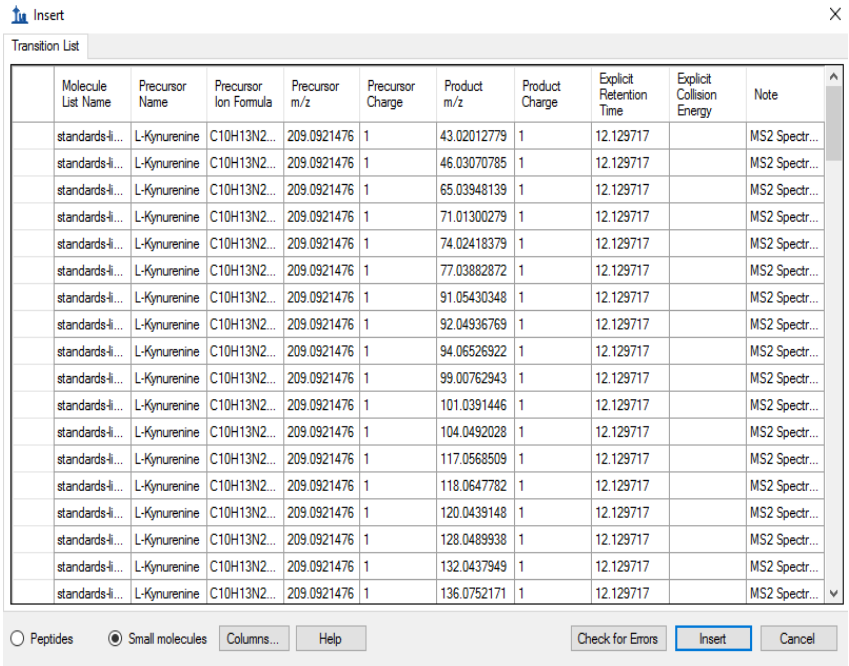


- Click outside the column name dialog to close it.
- Open the .CSV file that was exported from Elements with Excel. Copy all values, excluding the column headers.
- Return to the Skyline Transition List dialog, hover over the first cell, which should be highlighted in blue, and type CTRL V to paste the information copied from Excel into the

2. MacLean, Bioinformatics 2010 article: Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments

table. Verify that all cells are filled and that all of the information has not been pasted into a single cell. Click “Insert”.

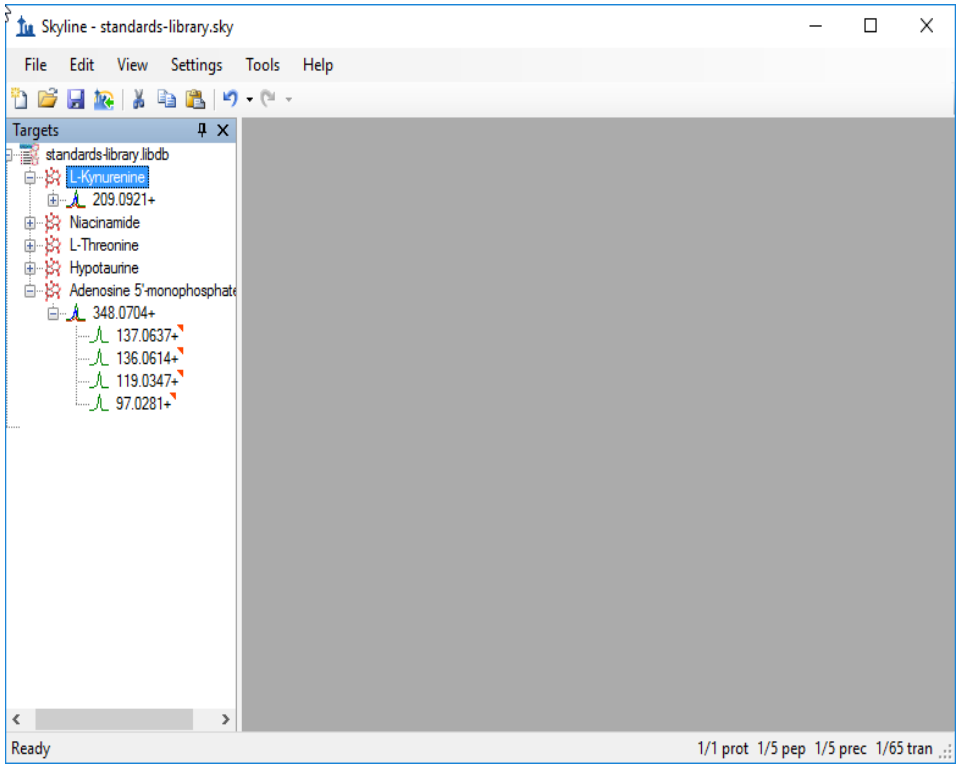
Figure 7: Data pasted into Skyline’s Insert Transition List dialog



Molecule List Name	Precursor Name	Precursor Ion Formula	Precursor m/z	Precursor Charge	Product m/z	Product Charge	Explicit Retention Time	Explicit Collision Energy	Note
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	43.02012779	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	46.03070785	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	65.03948139	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	71.01300279	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	74.02418379	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	77.03882872	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	91.05430348	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	92.04936769	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	94.06526922	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	99.00762943	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	101.0391446	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	104.0492028	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	117.0568509	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	118.0647782	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	120.0439148	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	128.0489938	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	132.0437949	1	12.129717		MS2 Spectr...
standards-li...	L-Kynurenine	C10H13N2...	209.0921476	1	136.0752171	1	12.129717		MS2 Spectr...

Skyline will import the data and display it as follows:

Figure 8: Imported Transition List in Skyline



Note: Before importing Elements exports into Skyline, a few settings may need to be changed through Settings -> Transition Settings.

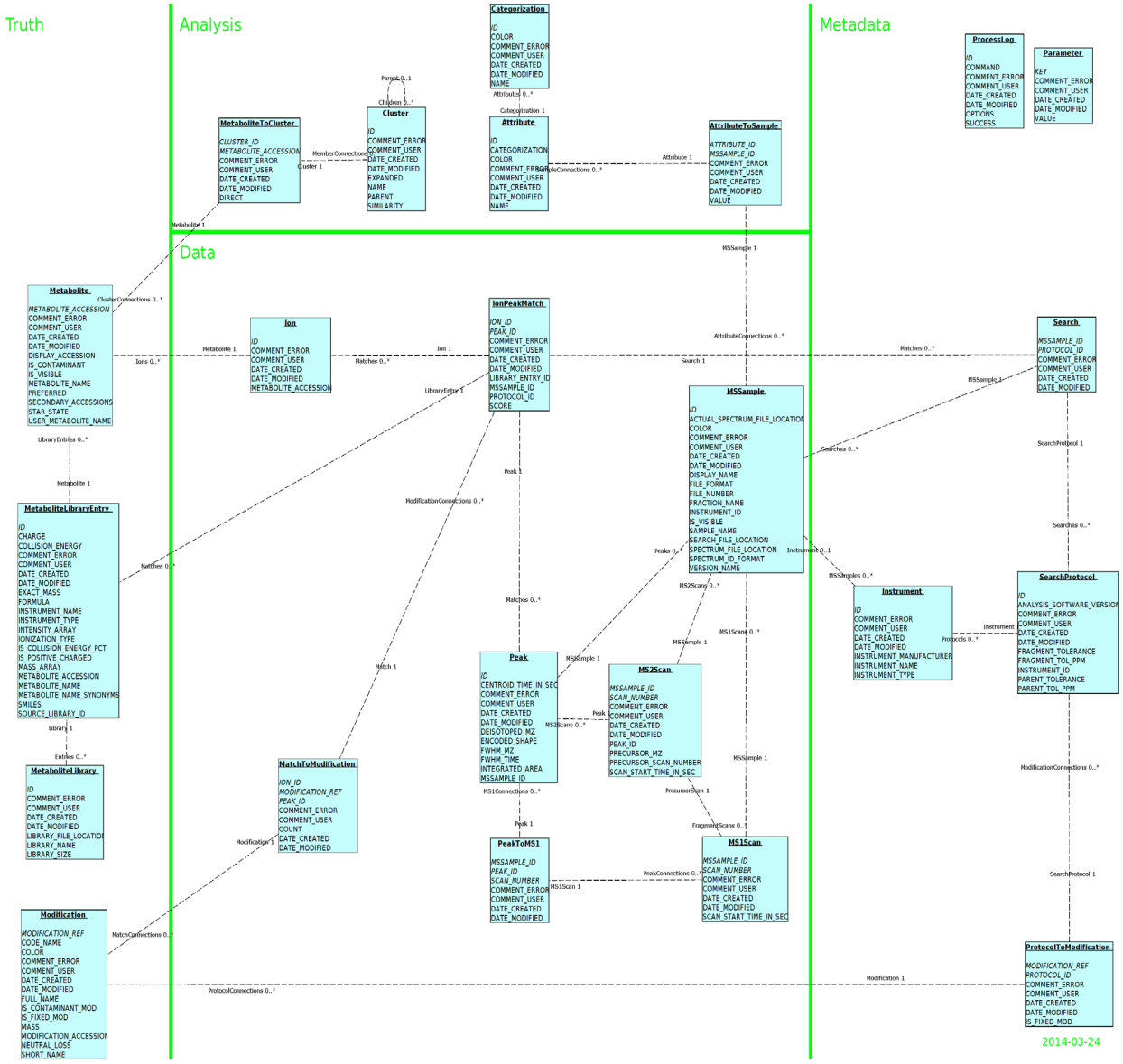
- First, under "full scan" select the desired MS1 filtering. Generally selecting isotope peaks "Count" and setting peaks to “3” gives results similar to Elements’.
- Second, under "filter" ensure that the desired precursor and ion charges are in the list. Generally, both fields should be "1" unless there are multiply-charged ions in the data. Ensure that "ion types" is "p". Ensure that "auto select all matching ions" is selected.

Appendix I.Structure of Scaffold Elements files (*.metdb)

Structure of Scaffold Elements files (*.metdb)

Scaffold Elements experiment files, *.METDB, are SQLite files. Figure 9 shows the schema of a *.METDB file. SQlite queries of this database can be submitted through the Export SQL table dialog box that can be opened from the menu **Export > Export SQL Table to Excel**. It is also possible to query the database directly in Elements from the Publish View> SQL Report tab, under the ‘SQL’ box.

Figure 9: Scaffold Elements database schema



-

Appendix J.Terminology

Blocking

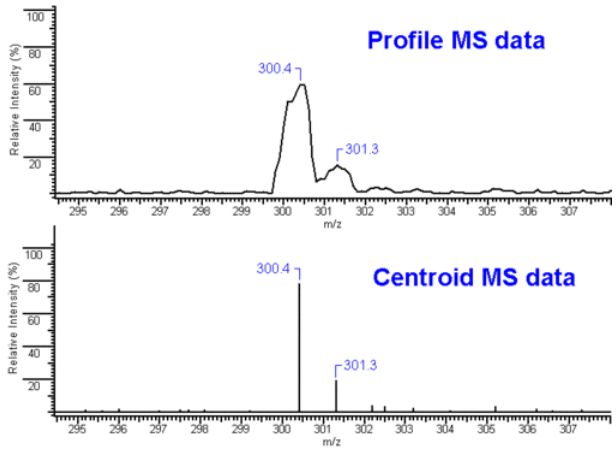
When groups of experimental units are similar, it is often a good idea to gather them together into blocks. By blocking the variability attributable to the differences between the blocks is isolated so that the differences caused by the treatments appear clearer.

Contingency table

In statistics, a contingency table (also referred to as cross tabulation or cross tab) is a type of table in a matrix format that displays the (multivariate) Frequency Table or distribution of the variables.

Profile vs.Centroid MS data

Data obtained from a mass spectrometer is collected in either profile or centroid mode. Data collected in profile mode produces features with width in the m/z dimension for a single retention time. Finding the 1D centroid of these waveforms is known as “centroiding”. When data is collected in centroid mode, the mass spectrometer automatically produces the centroided waveforms. Shown below are two mass spectra illustrating an ion cluster for profile data and a centroid mass spectrum created from the profile data.



Dendrogram

Or tree diagram, is a common method of graphically displaying the output of hierarchical clustering.

Experiment (a statistical definition)

An experiment manipulates factor levels to create treatments, randomly assigns subjects to these treatments levels, and compares the responses of the subject groups across treatment levels.

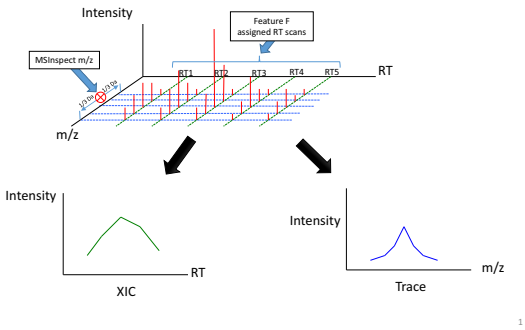
-

Factor

A variable whose levels are manipulated by the experimenter. Experiments attempt to discover the effects that differences in factor levels may have on the responses of the experimental units.

Feature, Mass Trace and XIC

In the context of LC-MS, a feature is a collection of raw (m/z, RT, intensity) data points that typically resembles a three-dimensional Gaussian waveform. An XIC (extracted ion chromatogram) is produced from the feature by summing all data points with the same retention time (RT) together, producing a 2D plot of RT versus intensity. A mass trace is produced from the feature by summing all data points with the same m/z together, producing a 2D plot of m/z versus intensity. (see Figure below).



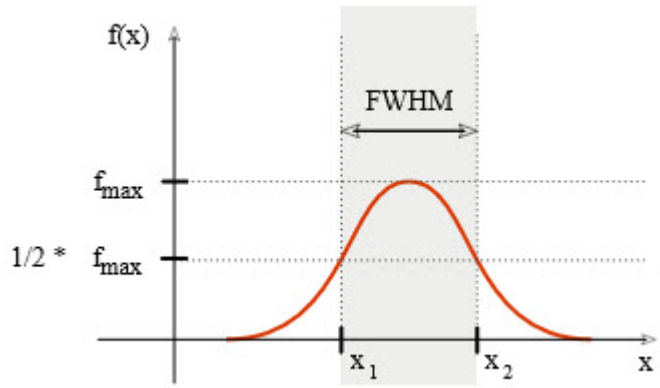
Frequency Table

In statistics, a frequency table is a table that displays the frequency of various outcomes in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample. Bivariate joint frequency distributions are often presented as (two-way) Contingency tables.

FWHM

Full Width at Half Maximum (FWHM) is a measure applied to 2D features: the width of the feature taken from a height of ½ of the maximum of the feature. Elements provides this measure for monoisotopic features in both retention time and m/z.

-



Treatment

The process, intervention or other controlled circumstance applied to randomly assigned experimental units. Treatments are the different levels of a single factor or are made up of combinations of levels of two or more factors.

-

Appendix K.Heat map clustering

The goal of reordering the columns and rows of a data matrix is to place analyte and samples with similar characteristics close to each other. Generally the reordering in heat maps is typically done using an agglomerative hierarchical clustering algorithm that groups similar data contained in a matrix or table. The clustering information is then displayed using a Dendrogram.

An agglomerative hierarchical clustering algorithm on n objects begins by considering each object to be a cluster of its on containing 1 object. At each step, the two closest groups are merged together until n objects are in a single group. In the case of a data matrix an object is typically a multidimensional vector whose components are given by the data listed in a row or a column.

There are a number of possible algorithms used to create agglomerative hierarchical clusters. Their differences mainly pertain to the definition of closeness or similarity between two objects or clusters before they are merged and to the agglomeration process by which clusters are merged into larger clusters.

Similarity or closeness is typically represented by the measurement of a distance *d(A,B)* between every pair *A* and *B* of objects. Typically *A, B* are multidimensional vectors that contain a series of numbers belonging to any two rows or columns depending on the group that is being clustered, i.e. either among the columns or rows of the data matrix.

Measuring the distance between clusters that have to be agglomerated into larger clusters is more complicated than measuring distances between single vectors. Different algorithms take different approaches in defining which is the link between clusters being considered as a measure of closeness when performing agglomeration.

Scaffold Elements uses an agglomerative hierarchical algorithm called **Single-Linkage clustering** with a Euclidean distance metric. The distance metric is applied to the coordinate-wise rank vectors of the rows or columns of the data matrix containing values that depend on the selected display type in the Samples View. The ranking is done over the whole ensemble of data included in the data matrix.

This distance metric will tend to associate measurements that rank at close levels.

$$d_{Euclidean}(r_A,r_B) = \sqrt{(r_{A1} - r_{B1})^2 + \ldots + (r_{An} - r_{Bn})^2}$$

Where:

- *A* and *B* are vectors whose coordinates characterize two rows or two columns of the data matrix at a selected Display Type and summarization level, see Summarization Bar; and r_A and r_B are coordinate-wise vectors of A and B. The components of the vector are given by the displayed values shown in the Samples Table at specific summarization levels, filtering and thresholding conditions. When considering the clustering of rows, the summarization level determines how many values for a selected statistics are shown in a row and consequently defines the dimensions of *A* and *B*. When considering the clustering

-

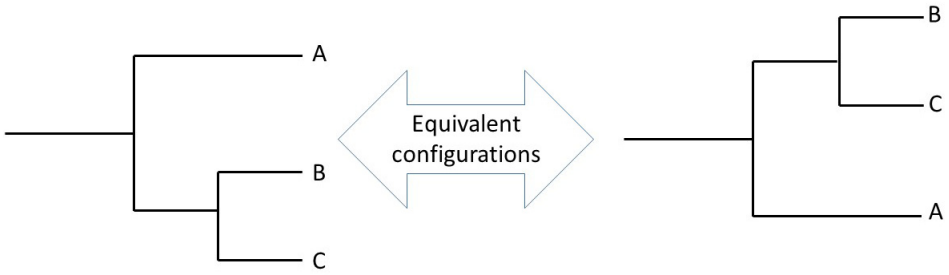
of the columns in the data matrix the dimension of the vectors that represent the columns is defined by filtering and thresholding applied to the Samples table.

In single-linkage, clustering agglomeration is made based on a single element pair, namely those two elements (one in each cluster) that are closest to each other. The shortest of these links that remains at any step causes the merging of the two clusters whose elements are involved. This method is also known as nearest neighbor clustering.

In Scaffold Elements clustering agglomeration is also based on a single pair element but the element selected for each cluster to be used to evaluate the shortest metric distance is the multidimensional vector that represents the center of mass among the group of vectors or points belonging to a cluster.

A common method used to graphically displaying the output of hierarchical clustering is to draw a dendrogram of the linkages among different clusters. At the bottom of the graph, each line corresponds to each object (cluster of size 1). When two clusters are merged, a line is drawn connecting the two clusters at a height corresponding to how similar the clusters are. The order of the objects is chosen to ensure that at the point where two clusters are merged, no other clusters are between them, but this ordering is not unique. When two clusters are merged, the choice of which of them is on the left or on the right side is arbitrary, this feature is called binary switching, see Figure 10.

Figure 10: Binary switching example



The Heat map in Scaffold Elements

Scaffold Elements constructs a heat map using information and data available in the Samples Tableand displays it in the The Visualize View.

Whatever thresholding and filtering are applied to the Samples Table determine the number of rows and columns considered for the data matrix used to create the Heat map shown in The Visualize View.However the rows listed include only groups of analytes even if any type of clustering might have been applied to the Samples table. Each column contains data from any MS sample or selected level of summarization.

In Scaffold Elements the quantitative values used for performing the agglomeration is determined by the selected Display type. This means that every Display type will show a

-

different ordering of the Heat map.

The result of the clustering is visualized as a dendrogram, which shows the sequence of cluster nodes and the distance at which each node is created.

-

Appendix L.Techniques to Control the Family-wise Error Rate

Currently Scaffold Elements supports control of the family-wise error rate, FWER, using Holm's step-down procedure and Hochberg's step-up procedure.

The way Scaffold Elements develops the two procedures is described in the following publication: Y. Huang *et al.* Biometrika (2007), 94,4,pp.965–975.

The two methods make the same type of comparisons, but Holm starts at the smallest p-value and works down the list until one fails the bound, while Hochberg starts at the largest p-value and works up the list until one passes the bound (and then declares that everything below that passes. Hence the Holm bound is in general more conservative than the Hochberg.

For example, let's suppose we have (m=5) analytes A, B, C, D, E with p-values 0.030, 0.014, 0.013, 0.060, and 0.009 respectively, and want to reject the null hypothesis at alpha = 0.05. Let's sort the p-values and make the following table:

k	p-value	$\alpha/(m+1-k)$	p-value < $\alpha/(m+1-k)$
1	0.009	0.01	yes
2	0.013	0.0125	no
3	0.014	0.0167	yes
4	0.030	0.025	no
5	0.060	0.05	no

- The Holm step-down procedure would start at k=1 and reject $H_0(1)$ but it would stop at k=2 since the p-value is larger than the bound.
- The Hochberg step-up procedure would start at k=5, go to k=4, go to k=3, see that the bound passes and stop, accepting $H_0(1)$, $H_0(2)$, and $H_0(3)$.

Instead of simply saying whether null hypotheses are rejected or not, we report the lowest $\alpha/(m+1-k)$ bound value. When p-values are lower than the reported bound the Null hypothesis can be rejected.

This means that in the example described above, the bound for Holm would be 0.0125, and for Hochberg 0.025 and the bound reported would be 0.0125.

The reason we are reporting both methods is related to the fact that technically the Hochberg procedure should only be used if the hypothesis tests are independent (which they are certainly not for Fisher's Exact Test, and not usually really for the other tests as well).

-

Appendix M.Using Principal Component Analysis in Scaffold Elements

Using Principal Component Analysis in Scaffold Elements

Principal Component Analysis is a tool for identifying the underlying sources of variation in a data set. PCA looks for patterns of expression among the analytes that can be used to group samples in meaningful ways. When used in combination with the flexible summarization offered in Scaffold Elements, this provides a powerful tool for exploring the biological meaning of quantitative differences observed in an experiment.

An Example

This example was performed in Scaffold LFQ. It uses demo file spectral_counting.sfdb. The data used in this example comes from a study to measure the effects of thermal processing on allergens in English walnuts[1] and was obtained from the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org)[2] via the PRIDE partner repository, with the dataset identifier PXD000907.

To begin, we create Categories corresponding to the variables in the experiment and apply Attributes to the samples to represent the experimental design.

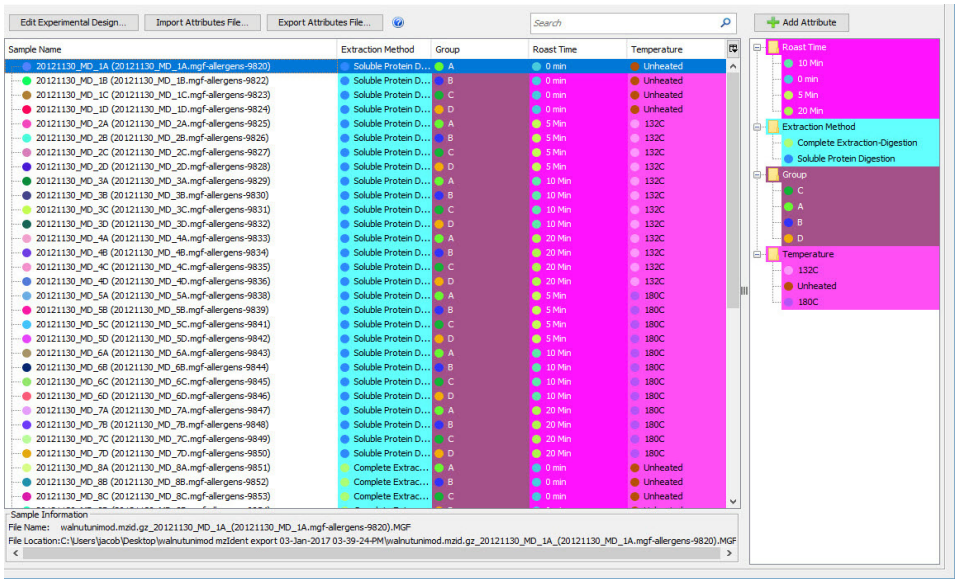
This study used a block design, in which four replicate samples were divided and subsamples of each underwent several treatments. The analytes from each were then extracted in two different ways.

In Scaffold Elements, we create a Category for the Replicate Group Number, and one for each of the variables studied. These are Protein Extraction Method, Roasting Time, Roasting Temperature. We create the appropriate Attributes to represent the different values each of these variables may take and assign the values to the samples in the Organize View.

1. Downs ML, Baumert JL, Taylor SL, Mills EN. Mass spectrometric analysis of allergens in roasted walnuts. J Proteomics. 2016 May 2. pii: S1874-3919(16)30177-4 PubMed: 27150359.
2. Vizcaíno JA, Csordas A, del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T, Xu QW, Wang R, Hermjakob H. 2016 update of the PRIDE database and related tools. Nucleic Acids Res. 2016 Jan 1;44(D1): D447-D456. PubMed PMID:26527722.

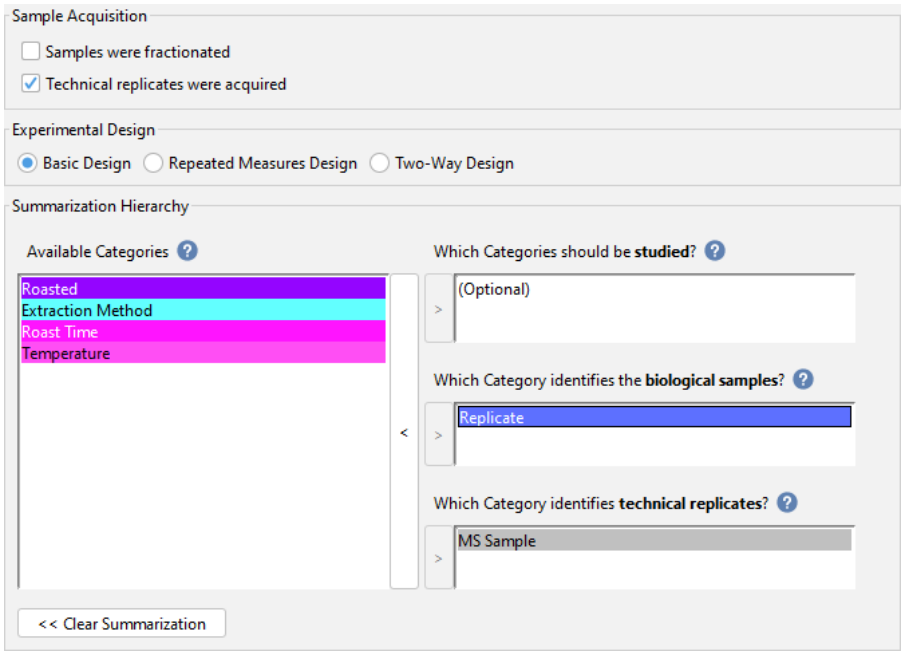
-

Figure 11: Samples with Attributes assigned



We might initially set the Summarization Hierarchy to specify our technical and biological replicates:

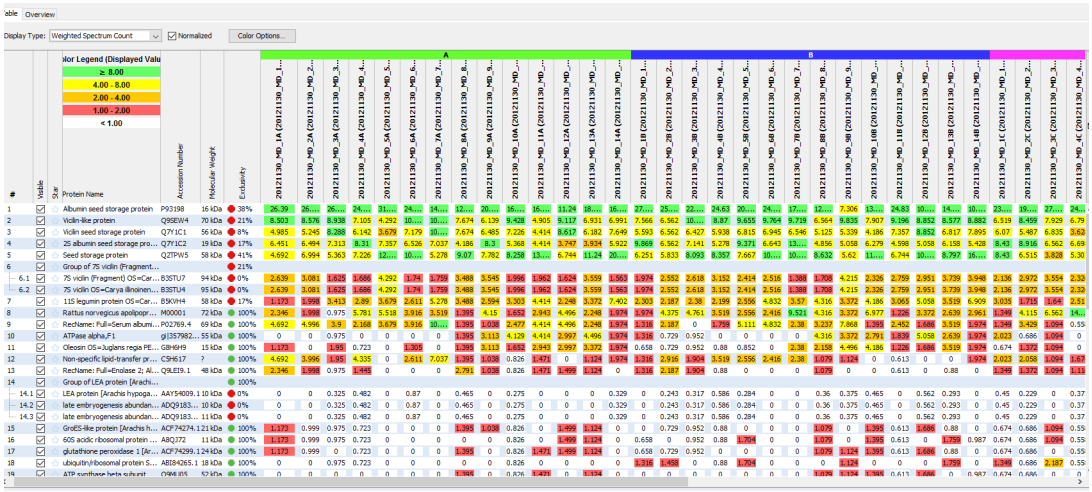
Figure 12: The Initial Summarization Hierarchy



The Samples View shows spectral counts for 53 analytes in the various MS Samples, but it is difficult to discern any patterns at this point:

-

Figure 13: The Samples View Initially



We must apply the treatment-related Categories to make this display meaningful, but there are several different treatments and we do not yet know which are important and how they interact. PCA can guide us in making these determinations.

PCA analyzes data in order to find patterns in the expression levels of the various analytes that differentiate samples or groups of samples. It constructs a weighting function that, when applied to the quantitative values of the analytes, results in the greatest separation of the samples, or, put another way, explains most of the variation between samples. This is Principal Component 1. The algorithm then continues to find other independent functions that also separate the samples in different ways, although they may account for somewhat less of the variation. These become Principal Component 2, 3, etc.

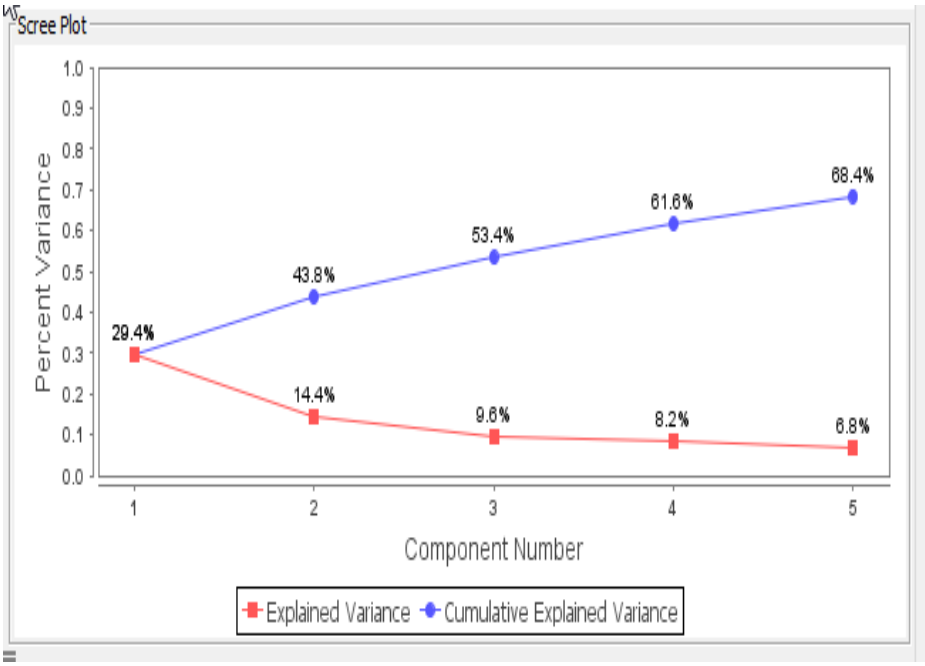
The Principal Component Analysis tab in the Visualize View provides several plots to help us interpret the results of PCA.

Scree Plot

The Scree Plot indicates the percentage of variation in the data explained by each Principal Component. This may help in determining which and how many of the factors in the study need to be considered.

-

Figure 14: The Scree Plot

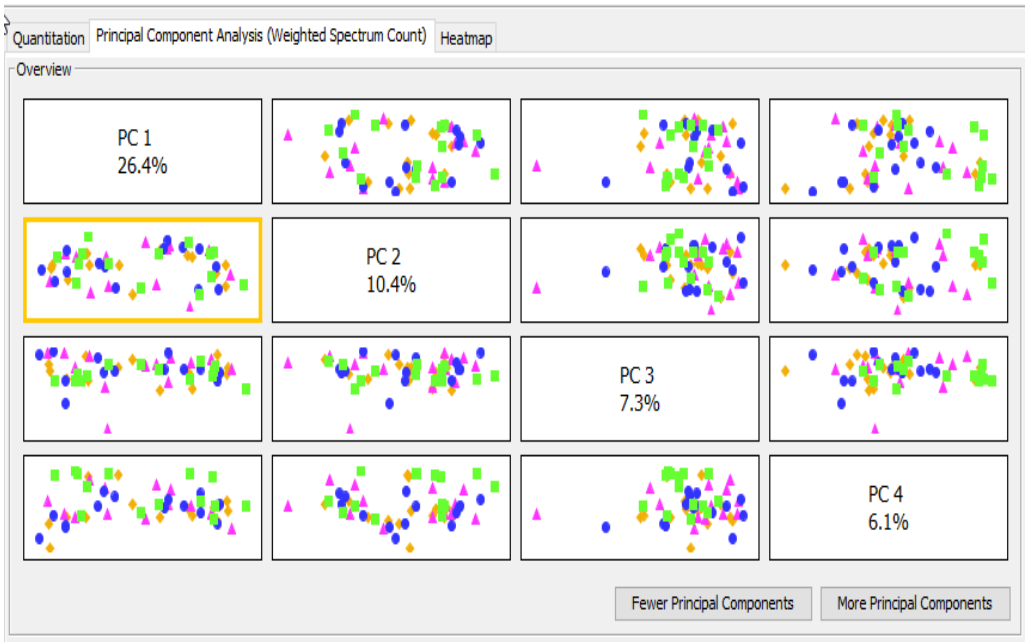


Overview

The Overview is a series of graphs where one Principal Component is plotted against another. The points in these graphs represent samples, and the X and Y coordinates are the values computed from the corresponding Principal Component functions. We can see that the samples tend to cluster in different ways depending on the Principal Components applied.

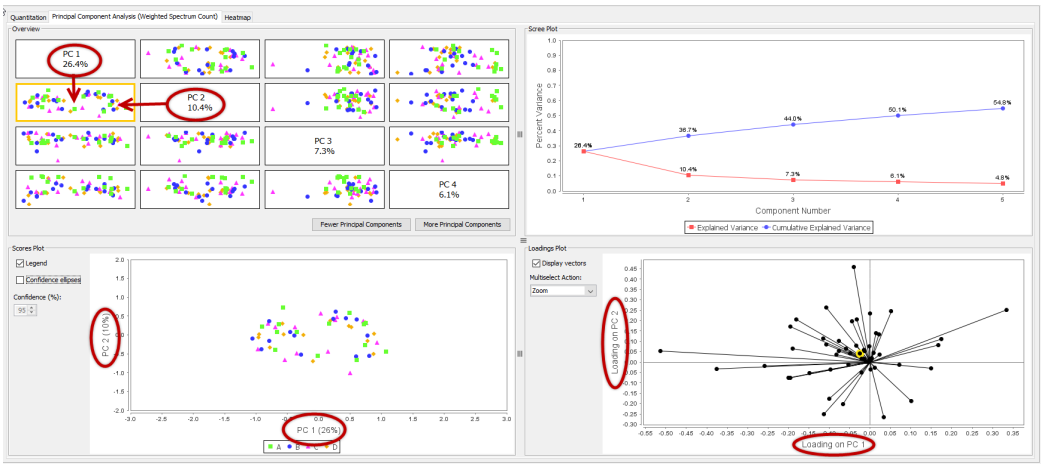
-

Figure 15: The Overview Plot



Clicking on a graph in the Overview selects the combination of Principal Components for display in greater detail in the Loadings and Scores Plots below.

Figure 16: Selecting Principal Components with the Overview

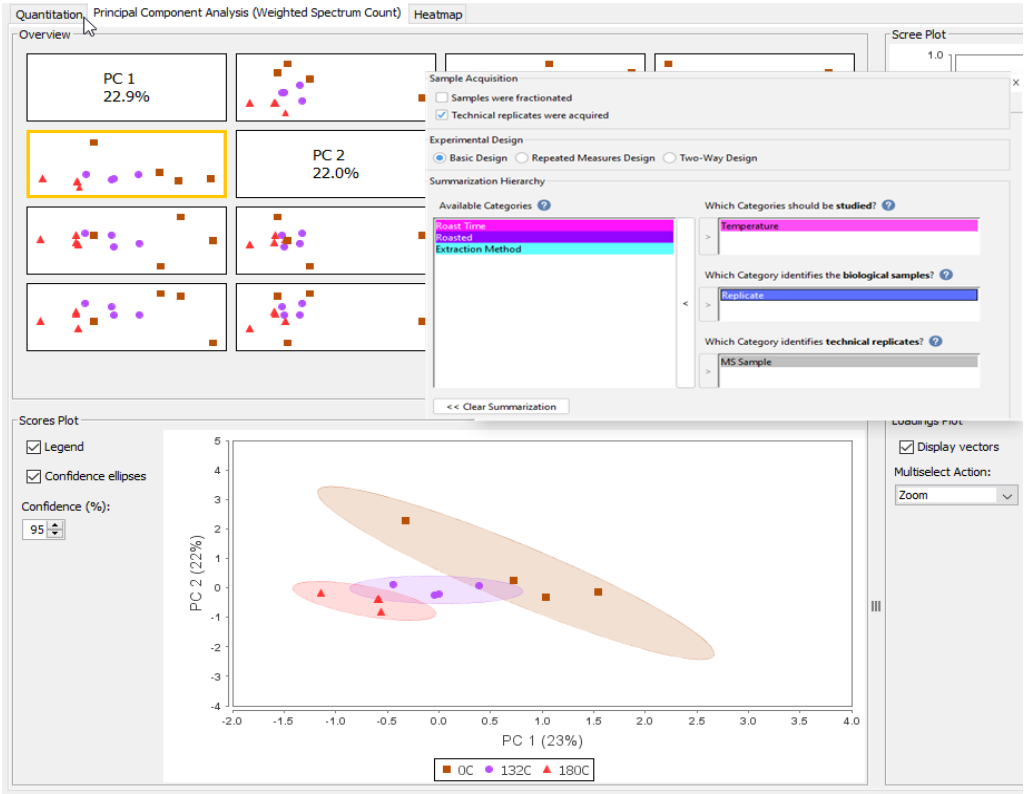


In the plot of PC2 vs. PC1, we can see that there appears to be clustering; to determine whether one or more of the treatments applied are responsible for the variation in the data, we will try applying different Categories.

First, we try Temperature:

-

Figure 17: Exploring the relationship of Temperature and PC1 and PC2



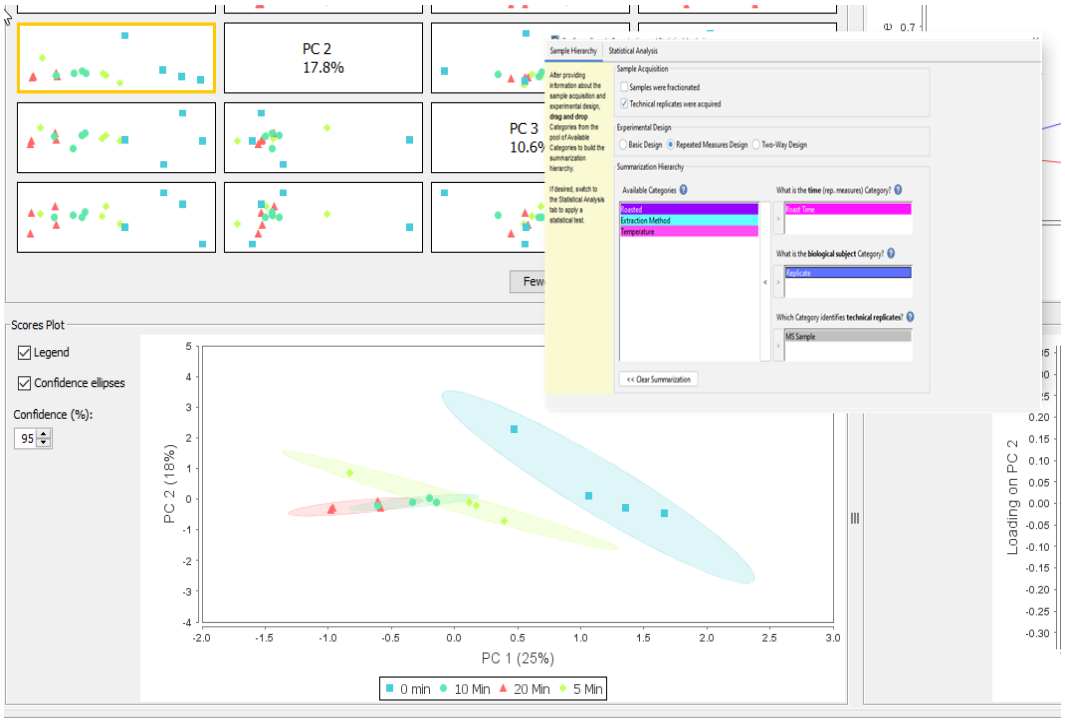
Here we see clustering in the Scores Plot. Roasting Temperature correlates with PC1, as the samples subjected to higher temperatures appear to the left in the chart (lower PC1 values), those roasted at lower temperatures are in the middle, and unroasted samples are on the right (higher PC1 values).

The temperature also appears to be contributing to some degree to PC2, as the samples roasted at the highest temperature appear slightly lower in the plot (lower PC2 values).

Looking at Roasting Time, we see:

-

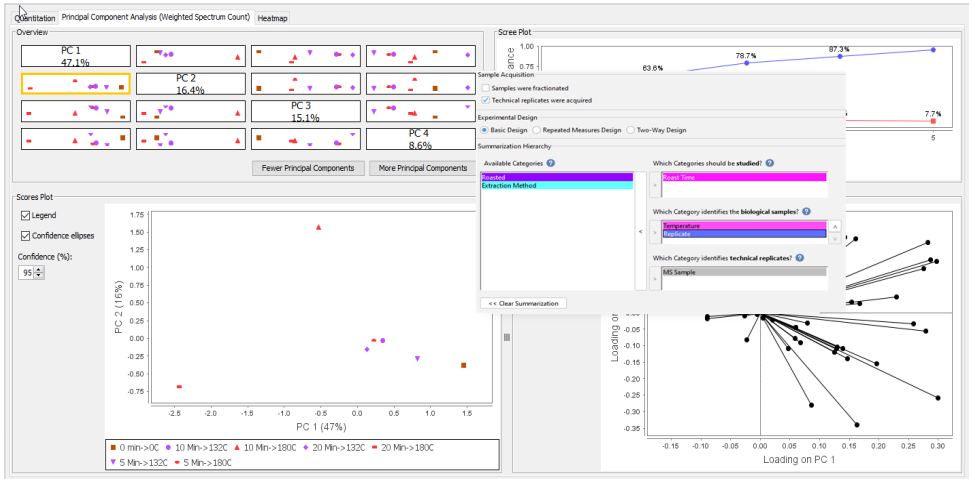
Figure 18: Exploring the Relationship of Roasting Temperature with PC1 and PC2



Once again, we see a similar pattern, with unroasted samples to the right, and the samples roasted for the longest period to the left, but there is more overlap between the different time groups. It appears that Roasting Time is correlated with PC1, but that the analyte changes occur at various time points in different samples. This is probably because of interactions between roasting time and roasting temperature.

If we examine these two variables together, we see that the roasting for 10-20 minutes at 132C is similar to roasting for 5 minutes at 180C.

Figure 19: Relationship between Time and Temperature

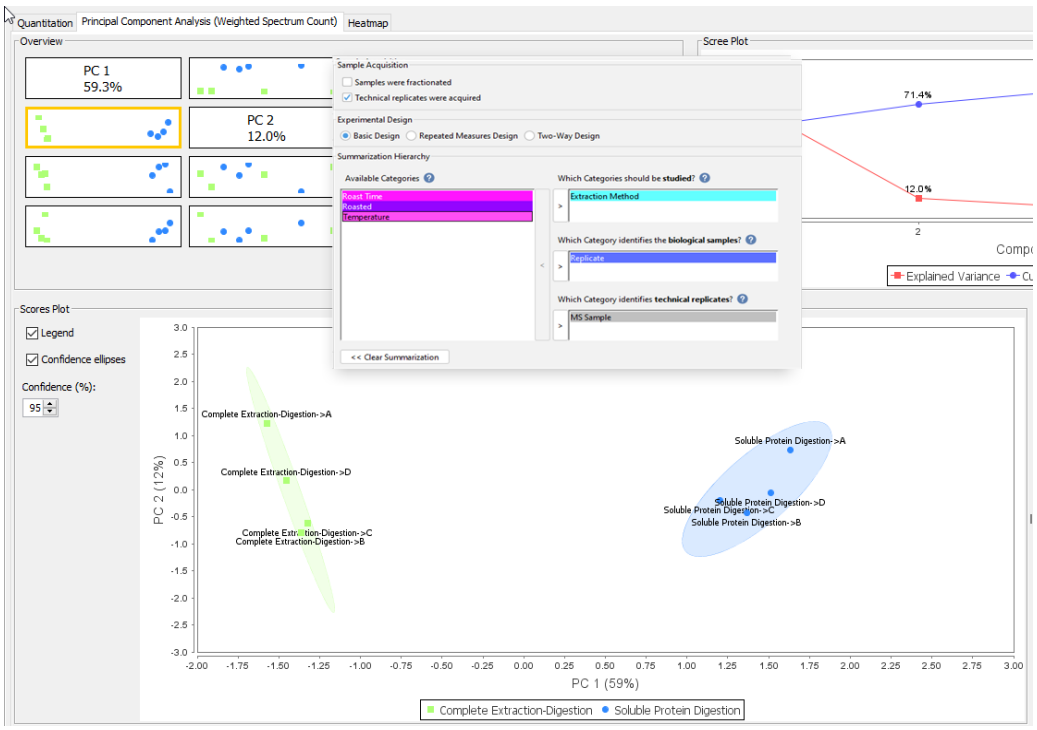


-

It appears, then, that the variation is explained by how thoroughly the nuts are roasted, which is governed by a combination of time and temperature. Since temperature gave clearer results, we will use temperature as the measure of degree of roasting. Another alternative would be to create a new attribute that captures the combination of time and temperature.

As can be seen below, Extraction Method produces the clearest clustering of all in PC2 vs PC1:

Figure 20: Exploring the Relationship between Extraction Method and PC1 and PC2

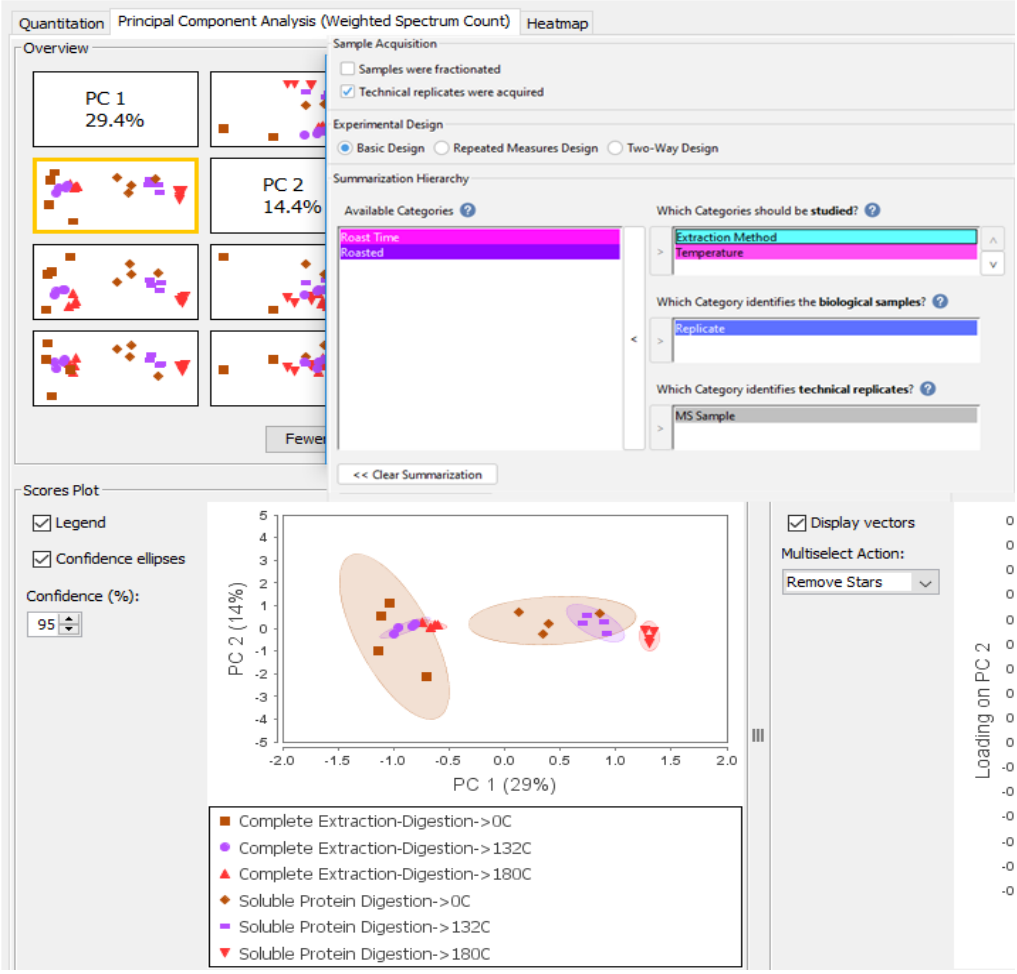


By examination of the labels, we can see that PC2 is probably based on differences among Replicates, since the Replicates appear in the same order in each extraction method and separate along the PC2 axis.

Exploring combinations of the factors produces some very clear clustering:

-

Figure 21: Roasting Temperature and Extraction Method



Here, the samples cluster by a combination of Extraction Method and Roasting Temperature. PC1 appears to represent a combination of Extraction Method and thoroughness of roasting.

Once we have established which treatments have a significant effect on analyte content and levels, we may wish to determine which specific s are most affected by them. This can help in answering questions such as which pathways are implicated in a disorder, which s are affected by a treatment, or which s might be useful in developing assays. for a certain condition. To move from samples to s, we examine the Loadings Plot.

Loadings Plot

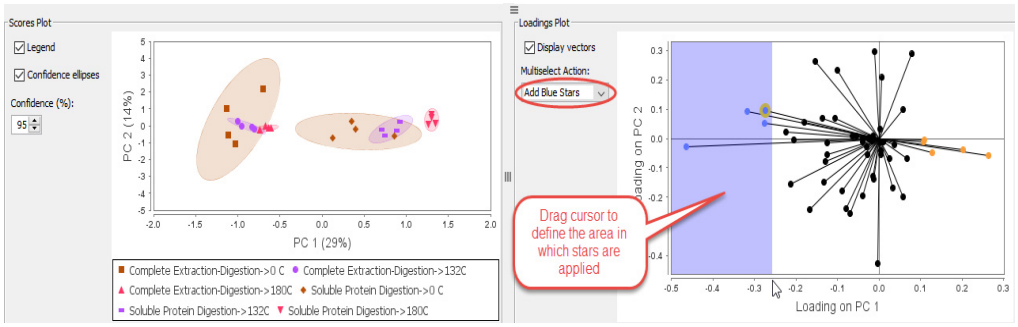
In the Loadings Plot, each point represents a . The coordinates of each point are a measure of the contributions of that to each of the components in the plot. For example, if the plot displays PC1 on the x-axis and PC2 on the y-axis, points far to the left and right represent s that contribute strongly to Principal Component 1. s near the top and bottom contribute strongly to PC2. As a result, corresponding locations in the Scores and Loading plots are

-

related.

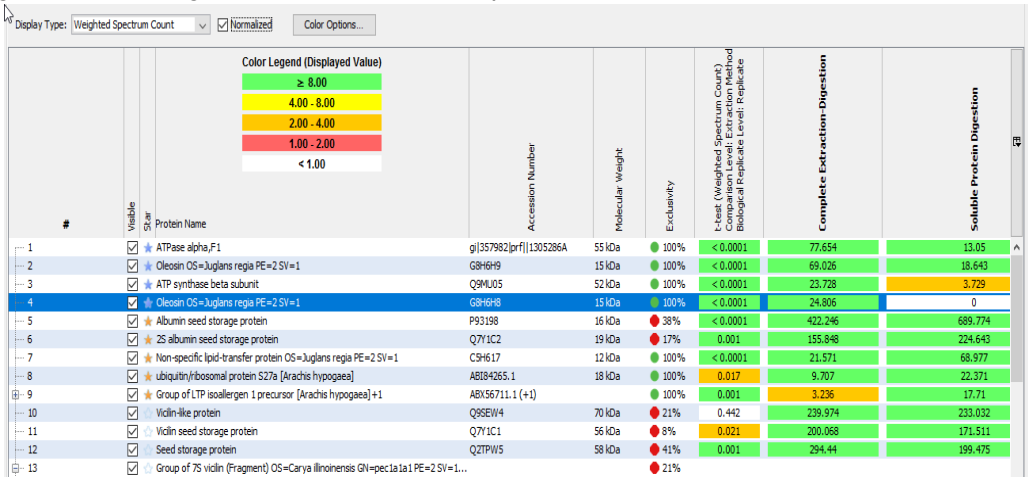
We can mark s through the Scores plot that may prove useful in identifying samples that are particularly effective at differentiating samples based on certain criteria. For example, we place orange stars on the s to the right in the Loadings Plot, and blue stars on s to the left:

Figure 22: Starring s of Interest through the Loadings Plot



In the Samples View, after summarizing by extraction method and applying a statistical test, we can see that indeed the orange-starred s are significantly higher in the soluble extraction while the blue-starred s are significantly higher in the complete extraction.

Figure 23: Viewing the starred s in the Samples View



#	Visible	Star	Protein Name	Accession Number	Molecular Weight	Exclusivity	t-test (Weighted Spectrum Count) Comparison Level: Extraction Method Biological Replicate Level: Replicate	Complete Extraction-Digestion	Soluble Protein Digestion
1	☑	★	ATPase alpha,F1	gi|357982|prf||1305286A	55 kDa	100%	< 0.0001	77.654	13.05
2	☑	★	Oleosin OS=Juglans regia PE=2 SV=1	G8H6H9	15 kDa	100%	< 0.0001	69.026	18.643
3	☑	★	ATP synthase beta subunit	Q9MU05	52 kDa	100%	< 0.0001	23.728	3.729
4	☑	★	Oleosin OS=Juglans regia PE=2 SV=1	G8H6H8	15 kDa	100%	< 0.0001	24.806	0
5	☑	★	Albumin seed storage protein	P93198	16 kDa	38%	< 0.0001	422.246	689.774
6	☑	★	2S albumin seed storage protein	Q7Y1C2	19 kDa	17%	0.001	155.848	224.643
7	☑	★	Non-specific lipid-transfer protein OS=Juglans regia PE=2 SV=1	C5H617	12 kDa	100%	< 0.0001	21.571	68.977
8	☑	★	ubiquitin/ribosomal protein S27a [Arachis hypogaea]	ABI84265.1	18 kDa	100%	0.017	9.707	22.371
9	☑	★	Group of LTP isoallergen 1 precursor [Arachis hypogaea]+1	ABX56711.1 (+1)		100%	0.001	3.236	17.71
10	☑	☆	Vicilin-like protein	Q9SEW4	70 kDa	21%	0.442	239.974	233.032
11	☑	☆	Vicilin seed storage protein	Q7Y1C1	56 kDa	8%	0.021	200.068	171.511
12	☑	☆	Seed storage protein	Q2TPW5	58 kDa	41%	0.001	294.44	199.475
13	☑	☆	Group of 7S vicilin (Fragment) OS=Carya illinoinensis GN=pec1a1a1 PE=2 SV=1...			21%			

In summary, by combining the insights gained through PCA analysis with flexible summarization and statistical analysis, we can gain insight into the biologically significant patterns in the data.

-

Appendix N.How PCA is Performed in Scaffold Elements

Principal Component Analysis (PCA) is a classical dimension-reduction technique based on linear algebra. The idea is to find the “underlying processes” that explain the variance in the data. In PCA, these “underlying processes” consist of linear combinations of the original variables.

The principal basis vectors are chosen one at a time in such away that each vector chosen

- is perpendicular to all the previously chosen principal basis vectors
- is one unit long
- points in the direction that explains the most variation of the data (given the constraints)

Dimension reduction is achieved by projecting the original vectors into the space spanned by some subset of the principal basis vectors.

In Scaffold Elements, the variables we consider are the (thresholded and filtered) s that are currently viewable in the Samples View. These s’ intensities are measured across the samples at the Biological Replicate Level. We can consider this as a collection of vectors

$$\begin{aligned}
\vec{I_1} &= (I_{11}, I_{12}, I_{13}, \ldots, I_{1m}) \\
\vec{I_2} &= (I_{21}, I_{22}, I_{23}, \ldots, I_{2m}) \\
&\vdots \\
\vec{I_n} &= (I_{n1}, I_{n2}, I_{n3}, \ldots, I_{nm})
\end{aligned}$$

where there are n samples and m s.

Intensity data is generally roughly log normal, that is, after applying a log transformation it becomes roughly normally distributed. There is a large wrinkle introduced with this idea of applying a logarithm, however, namely, how to deal with missing values.

In order to mitigate this problem, we have opted to apply a generalized logarithm (glog) instead of a regular logarithm. We use a generalized logarithm very similar to that used by Durbin[3] which is also used by MetaboAnalyst[4].This allows us to impute missing values as intensity I=0, and still apply the transformation. Explicitly, the transformation is:

3. Durbin BP, Hardin JS, Hawkins DM, Rocke DM. A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics. 2002;18(Suppl. 1):S105–S110.
4. Xia, J., Sinelnikov, I., Han, B., and Wishart, D.S. (2015) MetaboAnalyst 3.0 - making metabolomics more meaningful . Nucl. Acids Res. 43, W251-257.

-

$$\mathrm{glog}(I) = \log\left(\frac{I + \sqrt{I^2+1}}{2}\right).$$

Note that when I is large, glog(I)≈log(I), while for I near 0, glog(I) is perfectly well defined and approximately linear.

After applying glog to all intensities,

$$a_{ij} = \mathrm{glog}(I_{ij}),$$

we form the matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & & a_{nm} \end{bmatrix}.$$

The rows of A correspond to the samples S1, S2, ..., Sn, while the columns correspond to the s prot1, prot2,..., protm.

For spectral counts, the same transformation is applied, but with Count substituted for I.

Now, since we are interested in the variance of this “cloud” of vectors, it makes sense to center them by subtracting out the column means. This moves the “cloud” so that it is around the origin. Call this centered matrix X.

At this point in PCA, one must make a choice between using the covariance or the correlation matrix. In the second case, one would further scale each column of the centered matrix by the standard deviation of that column. This scaling is a good choice for those whose variables are not comparable to each other, being measured on different scales, it puts everyone on equal footing. However, in this case the variables, being s measured in the same way on the same machine, etc. are comparable in scale to each other so we opt to use the covariance matrix, that is:

$$\Sigma = \frac{1}{n-1} X^T X.$$

The entries in the matrix Σ measure the covariance of the variables (s).

-

Now, since Σ is a real symmetric matrix, it can always be diagonalized :

$$\Sigma = VDV^T$$

where

$$D = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & & \lambda_m \end{bmatrix}$$

is a diagonal matrix consisting of the eigenvalues of Σ arranged so that λ1≥λ2≥λ3≥⋯≥λm≥0,and V is an m×m matrix whose ith column, vi, is an eigenvector corresponding to λi. (That means: Σ·vi=λivi.) It turns out that these eigenvectors v1,v2,...,vm are exactly the principal basis vectors we are seeking, and satisfy the desired bullet points.

A.Interpretation

Each principal component points in turn at the direction of greatest remaining variation. Moreover, the eigenvalues measure how much variation is accounted for by each principal component.

Percent explained variance

The percentage of variance explained by the ith principal component is given by the formula:

$$\%\ \text{explained variance} = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_m}.$$

Interpretation of scores

How does dimension reduction work? Recall that each sample has a vector of its values across the s. We can project this vector onto the space spanned by, say, the first two principal components. This will give us a 2-dimensional understanding of how the samples differ. The plot of these 2-dimensional projections is called the 2D Scores Plot.

-

Interpretation of loadings

The dual question is how do the principal basis vectors correspond to the s? Take the first two principal basis vectors:

$$\vec{v_1} = (v_{11}, v_{12}, \ldots, v_{1m})$$
$$\vec{v_2} = (v_{21}, v_{22}, \ldots, v_{2m})$$

The coordinates of these vectors are called the loadings[5] of the s on to the principal components. Each v1j is a measure of how much the jth contributes to the first principal component (while v2j measures how much it contributes to the second principal component). Note these are unit length, so

$$v_{i1}^2 + v_{i2}^2 + \cdots + v_{im}^2 = 1.$$

plotting the points (v1j,v2j) for j=1,2,...,m gives the 2D loadings plot. Each point corresponds to a . If a 's point is close to (1,0) or (−1,0) on the loadings plot, it means that this is “mostly responsible" for the first principal component hence it must explain a great deal of the variation among the samples. If it is close to (0,1) or (0,-1) it means that this is “mostly responsible" for the second principal component.

Example

Suppose we have the s prot1, prot2, prot3, and prot4 and samples S1,S2, S3, S4, and S5, and that the following table shows the glogged intensities of the s in the samples:

	prot1	prot2	prot3	prot4
S_1	6	7	10	3
S_2	4	6	8	4
S_3	5	5	6	5
S_4	3	4	4	6
S_5	7	3	3	7

This table basically shows the matrix A. Note that prot3 behaves a lot like prot2, and that prot4 also has a similar expression profile to prot2 except reversed.

This sort of observation, though tricky to see here, will become exceedingly clear after PCA decomposition.

---

5. Actually `loading' is a loaded term in the literature; it sometimes means the coordinates of a scaled version of the basis vector. This is more often done when using the correlation matrix instead of the covariance matrix.

-

Already the trend is a bit more clear after we compute the means of 5, 5, 6, and 5 respectively and subtract these from the columns to get the matrix X:

$$X = \begin{bmatrix} 1 & 2 & 4 & -2 \\ -1 & 1 & 2 & -1 \\ 0 & 0 & 0 & 0 \\ -2 & -1 & -2 & 1 \\ 2 & -2 & -4 & 2 \end{bmatrix}.$$

The covariance matrix is:

$$\Sigma = \frac{1}{4}\begin{bmatrix} 10 & -1 & -2 & 1 \\ -1 & 10 & 20 & -10 \\ -2 & 20 & 40 & -20 \\ 1 & -10 & -20 & 10 \end{bmatrix} \qquad (2)$$

We can see that this matrix shows that s 2, 3, and 4 are highly correlated (large values off the diagonal except in the first row/column), while 1 is not correlated with the others. We can diagonalize - (we will skip the details of how), to figure out that the principal basis vectors in this case are:

$$\vec{v_1} = \begin{pmatrix} 0.04 \\ -0.41 \\ -0.82 \\ 0.41 \end{pmatrix} \text{ and } \vec{v_2} = \begin{pmatrix} 0.99 \\ 0.02 \\ 0.04 \\ -0.02 \end{pmatrix},$$

$$\lambda_1 = 15.0$$

$$\lambda_2 = 2.5$$

(The third and fourth eigenvalues are both 0.)

Let us interpret these results. The first principal basis vector shows the linear relationship between s 2, 3, and 4. In particular, the component for the 3rd is twice that of the 2nd and 4th, and going in the same direction as the 2nd. The second principal basis vector shows that all of the remaining variation is basically occurring with 1.

The percentage of variance explained by the first principal component is

-

$$\frac{15.0}{15.0 + 2.5 + 0 + 0} = 85.9$$

In this case, the second principal component explains the remaining 13:1% of the variation.

Interface in Scaffold Elements

Users will find the Principal Component Analysis tab in the Visualize View. The tab shows four components.

Overview Chart

The Overview Chart allows an initial view into the first 3, 4, or 5 principal components. The squares along the diagonal denote the principal components (PCs) and show the percent explained variance for each. Off the diagonal, each square is a 2D scores plot whose axes are determined by the PC for the corresponding row and column. For details on interpreting scores plots, see section 4.3 below.

The Overview Chart allows the user to select the axes for the scores and loadings chart. Simply mouse-over the square corresponding to the desired axes and click to select those axes for the other charts in the PCA view.

Scree Plot

The Scree Plot gives a graphical display of the percent explained variance by the first 5 principal components. The lower curve shows the percent explained by each individual principal component, while the upper curve show the cumulative percent explained variance.

Scores Plot

The scores plot shows the scores: the projections of the original vectors onto the space spanned by the selected principal components. The samples, taken from the Biological Replicate Level, are denoted as dots which are colored according the attribute to which they correspond in the Comparison Level.

The scores plot also shows the 95%-confidence ellipses for each attribute. (Actually the p%-confidence ellipses where p can be specified by the user.) These ellipses show the region where 95% of the data points will lie assuming their distributions are independent and normally distributed in the dimensions being plotted. These ellipses can be used to see if attributes separate well in the currently examined dimensions.

-

Loadings Plot

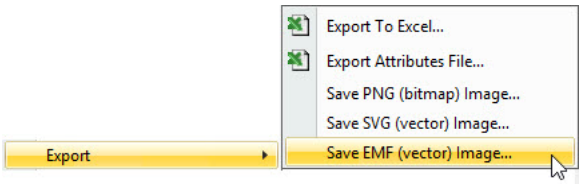
The 2D Loadings Plot shows the loadings as described above. The plot is interactive. In addition to allowing zooming, one can also use the plot to select the current (with a single click), or select a and switch to the s View with a double click.

-

Appendix O.Description of Mouse Right Click Context Menu Commands

- **Collapse All** - Collapses the samples tree table.
- **Copy >** - Provides a number of option for copying data from a table

  - *Copy Image* - Copies the image of the current table.
  - *Copy Selected Cell* - Copies data contained in the selected cell of the current table.
  - *Copy Selected Row* - Copies data contained in the selected row of the current table.
  - *Copy All Data* - Copies data contained in the current table.
- **Edit Sample Name** - Activates the sample name text box for editing purposes. Equivalent to double clicking over the sample's name.
- **Expand All** - Expands the whole samples tree table
- **Export** -Provides access to a couple of exports and three different ways to export an image of the table.



  - *Export To Excel* - Generates a tab delimited text file of the currently selected table. The file can be opened and viewed in Excel.
  - *Export Attribute File* - Generates a tab delimited text file of the meta-data attributes assigned to each MS sample in the current experiment. The file can be opened and viewed in Excel.
  - *Save PNG (Bitmap) Image* - Saves a PNG image of the selected table.
  - *Save SVG (Vector) Image* - Saves a PNG image of the selected table.
  - *Save EMF (Vector) Image* - Saves a EMF image of the selected table.
- **Find** - Opens the Find dialog.
- **Print** - Prints an image of the current table.

-

Index

A
Add Spectrum to Library 115
Advanced Tab 43
Applying filters to the Metabolite List 102
Attributes File 89
B
box plot 115
C
Confidence Thresholds 102
Consensus MS1 Spectrum Formation 24
Custom Spectral Library 156
D
Data processing workflow 20
Data value tags
    Missing Ref. 103
    Missing Values 103
    No Values 103
DIA Data 23
Display pane in the Samples View 102
Display Type 102
E
Elements
    Installation instructions 8
    Licensing 10
Elements feature identification 24
Elements Features extraction 22
Elements for Metabolomics
    licensing for 10
Experimental designs supported in Elements 19
F
Family-wise Error Rate 173
Feature Matching Selected Ion pane in Metabolites View 107
Features Extraction 22
G
Grouping and Clustering 25, 26
H
Heat Map
    reordering 161
I
ID Score 166
Identification of In-source Fragments 162
in-source fragments 104
Installation instructions 8
Installing Elements 8

Ions pane in Metabolites View 107
Isotopic Distribution Score 163
L
Layout type in Metabolites View 106
Library View 128
License key registration
    License key renewal 14
License keys 10
licensing for Elements for Metabolomics 10
Lock Mass Correction 35
M
Mass Accuracy Score 162
metabolites
    hiding in the Samples View 101
Metabolites List bar 106
Metabolites View 105
    Feature Matching Selected Ion pane 107
    Ions pane 107
    Layout type 106
    Metabolites List bar 106
    Visualization pane 108
Metabolomic Flux Analysis 141
Minimum system requirements 15
Missing Ref., Data Value tag 103
Missing Values, Data Value tag 103
MS1 Annotation Score 166
MS1 peak group 25
MS2 Score 165
N
No Values, Data Value tag 103
Normalized check box 104
O
Organizing data 82
P
PCA 181, 182
Penalty for lack of MS2 match 167
percentage of missing values 115
Personal Spectral Library 179
Precursor intensity quantitation
    Calculations 143
Precursor Intensity tags 102
present in analogous labeled sample 142
Publish View 136
R
Rainbow 85
Recommended system requirements 15

Reextraction
177
Registering license key
No Internet connection 13, 14
Renewal 14
Renewing time based license key 14
Require incorporation number higher than 2 in analogous labeled sample 142
Rolling up of quantitative values 103
S
Samples View
Display pane 102
hiding metabolites 101
sorting feature 97
Saving Indexed Peak Files 40
Scoring algorithms 162
Searches against spectra libraries 24
Skyline 130
sorting feature
Samples View 97
Star filters 61
Streamline metabolomics studies with Elements 19
Summarization Bar 59
Supported experimental designs 19
System Requirements
Minimum requirements 15
Recommended requirements 15
System requirements 15
T
Table Tab Display pane
Display Type 102
Tagging Metabolites of Interest
The star function 101
Terminology 154
The Flux Report 142
The METLIN Mass Spectral Library 52
The Multi-Select Action 118
Time based license key renewal 14
U
Use only the most intense ion in each metabolite 142
V
Visualization pane in the Metabolites View 108
X
XIC Score 165