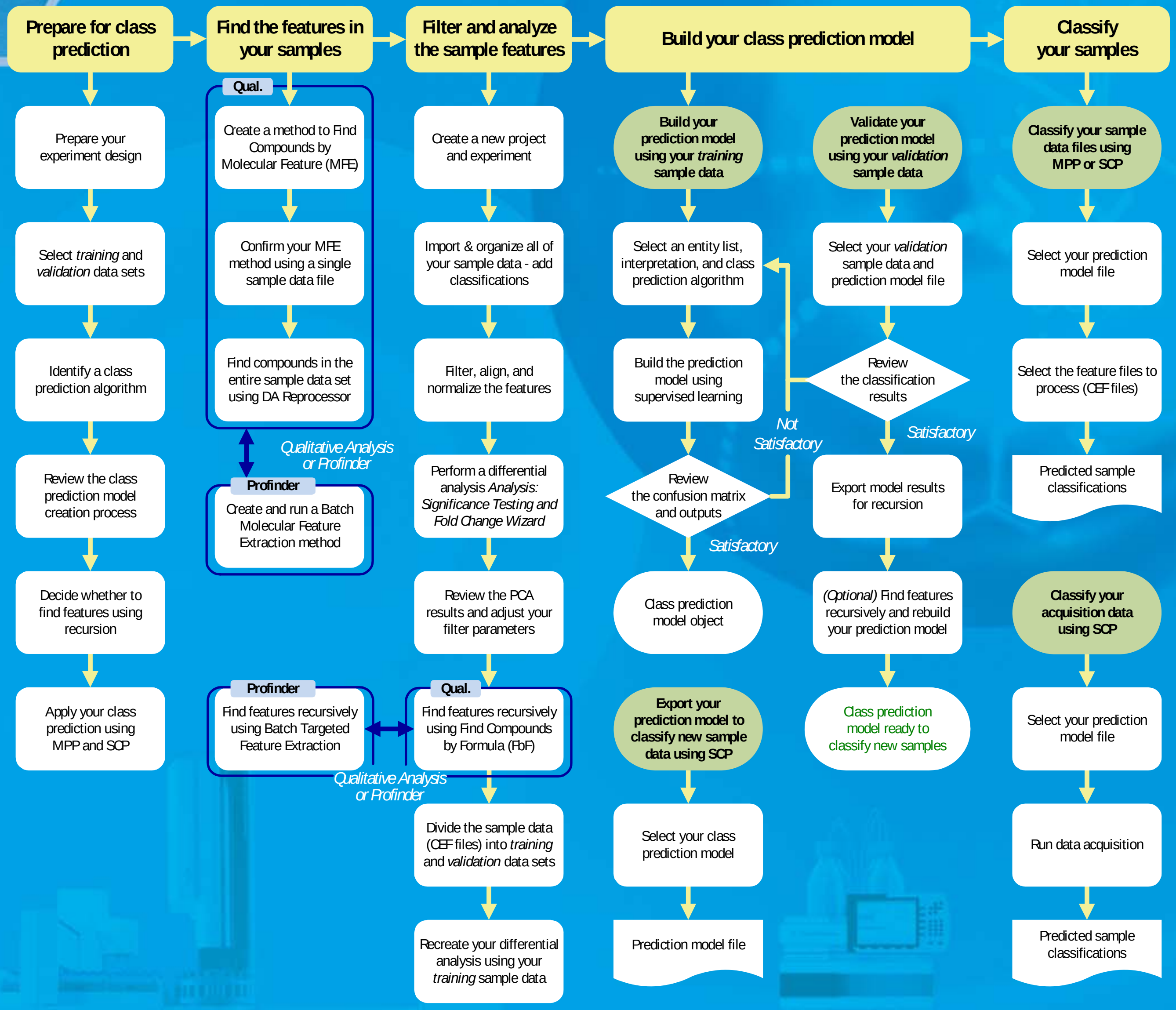
Class Prediction with Agilent Mass Profiler Professional

Workflow Guide



Prepare for class prediction

Prepare your experiment design

Select *training* and *validation* data sets

Identify a class prediction algorithm

Review the class prediction model creation process

Decide whether to find features using recursion

Apply your class prediction using MPP and SCP

Find the features in your samples

Qual.

Create a method to Find Compounds by Molecular Feature (MFE)

Confirm your MFE method using a single sample data file

Find compounds in the entire sample data set using DA Reprocessor

Qualitative Analysis or Profinder

Profinder

Create and run a Batch Molecular Feature Extraction method

Profinder

Find features recursively using Batch Targeted Feature Extraction

Qualitative Analysis or Profinder

Filter and analyze the sample features

Create a new project and experiment

Import & organize all of your sample data - add classifications

Filter, align, and normalize the features

Perform a differential analysis *Analysis: Significance Testing and Fold Change Wizard*

Review the PCA results and adjust your filter parameters

Qual.

Find features recursively using Find Compounds by Formula (FbF)

Divide the sample data (CEF files) into *training* and *validation* data sets

Recreate your differential analysis using your *training* sample data

Build your class prediction model

Build your prediction model using your *training* sample data

Select an entity list, interpretation, and class prediction algorithm

Build the prediction model using supervised learning

Review the confusion matrix and outputs

Not Satisfactory

Satisfactory

Class prediction model object

Export your prediction model to classify new sample data using SCP

Select your class prediction model

Prediction model file

Validate your prediction model using your *validation* sample data

Select your *validation* sample data and prediction model file

Review the classification results

Not Satisfactory

Satisfactory

Export model results for recursion

(Optional) Find features recursively and rebuild your prediction model

Class prediction model ready to classify new samples

Classify your samples

Classify your sample data files using MPP or SCP

Select your prediction model file

Select the feature files to process (CEF files)

Predicted sample classifications

Classify your acquisition data using SCP

Select your prediction model file

Run data acquisition

Predicted sample classifications

Agilent Technologies

Notices

Manual Part Number

5991-1911EN

Edition

Revision A, August 2014

Printed in USA

Agilent Technologies, Inc.
5301 Stevens Creek Blvd.
Santa Clara, CA 95051

Acknowledgements

Microsoft is either a registered trademark or trademark of Microsoft Corporation in the United States and/or other countries.

Adobe is a trademark of Adobe Systems Incorporated.

Warranty

The material contained in this document is provided "as is," and is subject to being changed, without notice, in future editions. Further, to the maximum extent permitted by applicable law, Agilent disclaims all warranties, either express or implied, with regard to this manual and any information contained herein, including but not limited to the implied warranties of merchantability and fitness for a particular purpose. Agilent shall not be liable for errors or for incidental or consequential damages in connection with the furnishing, use, or performance of this document or of any information contained herein. Should Agilent and the user have a separate written agreement with warranty terms covering the material in this document that conflict with these terms, the warranty terms in the separate agreement shall control.

Technology Licenses

The hardware and/or software described in this document are furnished under a license and may be used or copied only in accordance with the terms of such license.

Restricted Rights

If software is for use in the performance of a U.S. Government prime contract or subcontract, Software is delivered and licensed as "Commercial computer software" as defined in DFAR 252.227-7014 (June 1995), or as a "commercial item" as defined in FAR 2.101(a) or as "Restricted computer software" as defined in FAR 52.227-19 (June 1987) or any equivalent agency regulation or contract clause. Use, duplication or disclosure of Software is subject to Agilent Technologies' standard commercial license terms, and non-DOD Departments and Agencies of the U.S. Government will receive no greater than Restricted Rights as defined in FAR 52.227-19(c)(1-2) (June 1987). U.S. Government users will receive no greater than Limited Rights as defined in FAR 52.227-14 (June 1987) or DFAR 252.227-7015 (b)(2) (November 1995), as applicable in any technical data.

Safety Notices

CAUTION

A **CAUTION** notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in damage to the product or loss of important data. Do not proceed beyond a **CAUTION** notice until the indicated conditions are fully understood and met.

WARNING

A WARNING notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in personal injury or death. Do not proceed beyond a WARNING notice until the indicated conditions are fully understood and met.

Contents

About this guide

- This class prediction workflow guide is part of a series of workflow guides developed to help you to analyze your sample data using Mass Profiler Professional. Other workflow guides available in this series include:

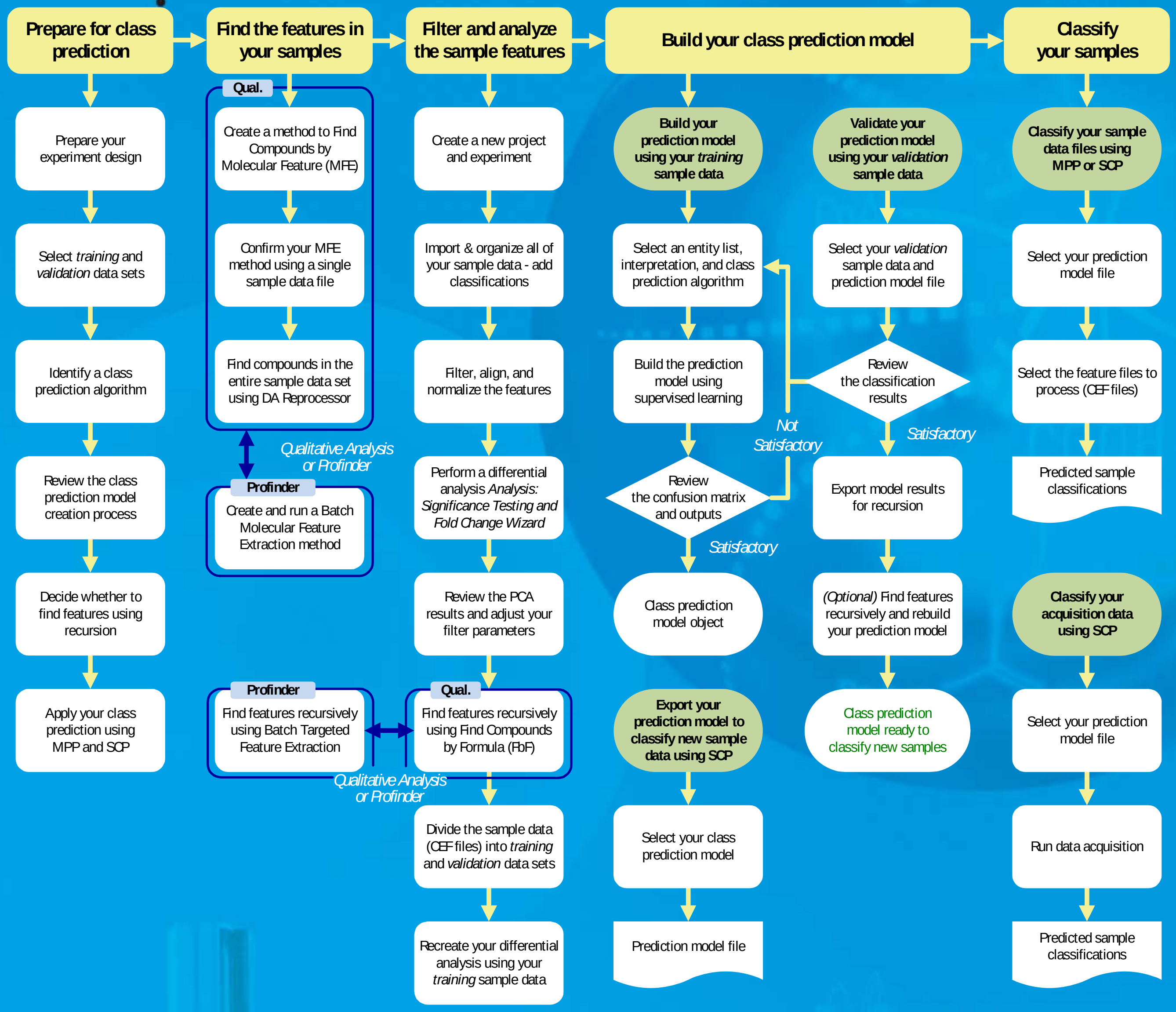
  The *Agilent Metabolomics Workflow - Discovery Workflow Guide* (5990-7067EN, Revision B) and

  The *Integrated Biology with Agilent Mass Profiler Professional - Workflow Guide* (5991-1909EN, Revision A, June 2013)

- This workflow guide presents the steps to use MassHunter Qualitative Analysis and MassHunter Profinder to find untargeted and targeted (recursive) features in your sample data files.
- Class prediction uses many concepts that are described in the Metabolomics Workflow, and the Metabolomics Workflow presents steps that precede the operations used in the Integrated Biology Workflow.
- The Mass Profiler Professional and Sample Class Predictor wizards and workflow images are based on version 12.61.
- Formatting of text that appears in the left-hand margin helps guide you through the operations.
- Operations are illustrated with flow charts that show you how the wizards are navigated based on your experiment and selections.

Before you begin

Make sure you read and understand the information in this chapter and have the necessary computer equipment and software before you start your class prediction analysis.



Prepare for class prediction

Prepare your experiment design

Select *training* and *validation* data sets

Identify a class prediction algorithm

Review the class prediction model creation process

Decide whether to find features using recursion

Apply your class prediction using MPP and SCP

Find the features in your samples

Qual.

Create a method to Find Compounds by Molecular Feature (MFE)

Confirm your MFE method using a single sample data file

Find compounds in the entire sample data set using DA Reprocessor

Qualitative Analysis or Profinder

Profinder

Create and run a Batch Molecular Feature Extraction method

Profinder

Find features recursively using Batch Targeted Feature Extraction

Qualitative Analysis or Profinder

Filter and analyze the sample features

Create a new project and experiment

Import & organize all of your sample data - add classifications

Filter, align, and normalize the features

Perform a differential analysis *Analysis: Significance Testing and Fold Change Wizard*

Review the PCA results and adjust your filter parameters

Qual.

Find features recursively using Find Compounds by Formula (FbF)

Divide the sample data (CEF files) into *training* and *validation* data sets

Recreate your differential analysis using your *training* sample data

Build your class prediction model

Build your prediction model using your *training* sample data

Select an entity list, interpretation, and class prediction algorithm

Build the prediction model using supervised learning

Review the confusion matrix and outputs

Not Satisfactory

Satisfactory

Class prediction model object

Export your prediction model to classify new sample data using SCP

Select your class prediction model

Prediction model file

Validate your prediction model using your *validation* sample data

Select your *validation* sample data and prediction model file

Review the classification results

Satisfactory

Export model results for recursion

(Optional) Find features recursively and rebuild your prediction model

Class prediction model ready to classify new samples

Classify your samples

Classify your sample data files using MPP or SCP

Select your prediction model file

Select the feature files to process (CEF files)

Predicted sample classifications

Classify your acquisition data using SCP

Select your prediction model file

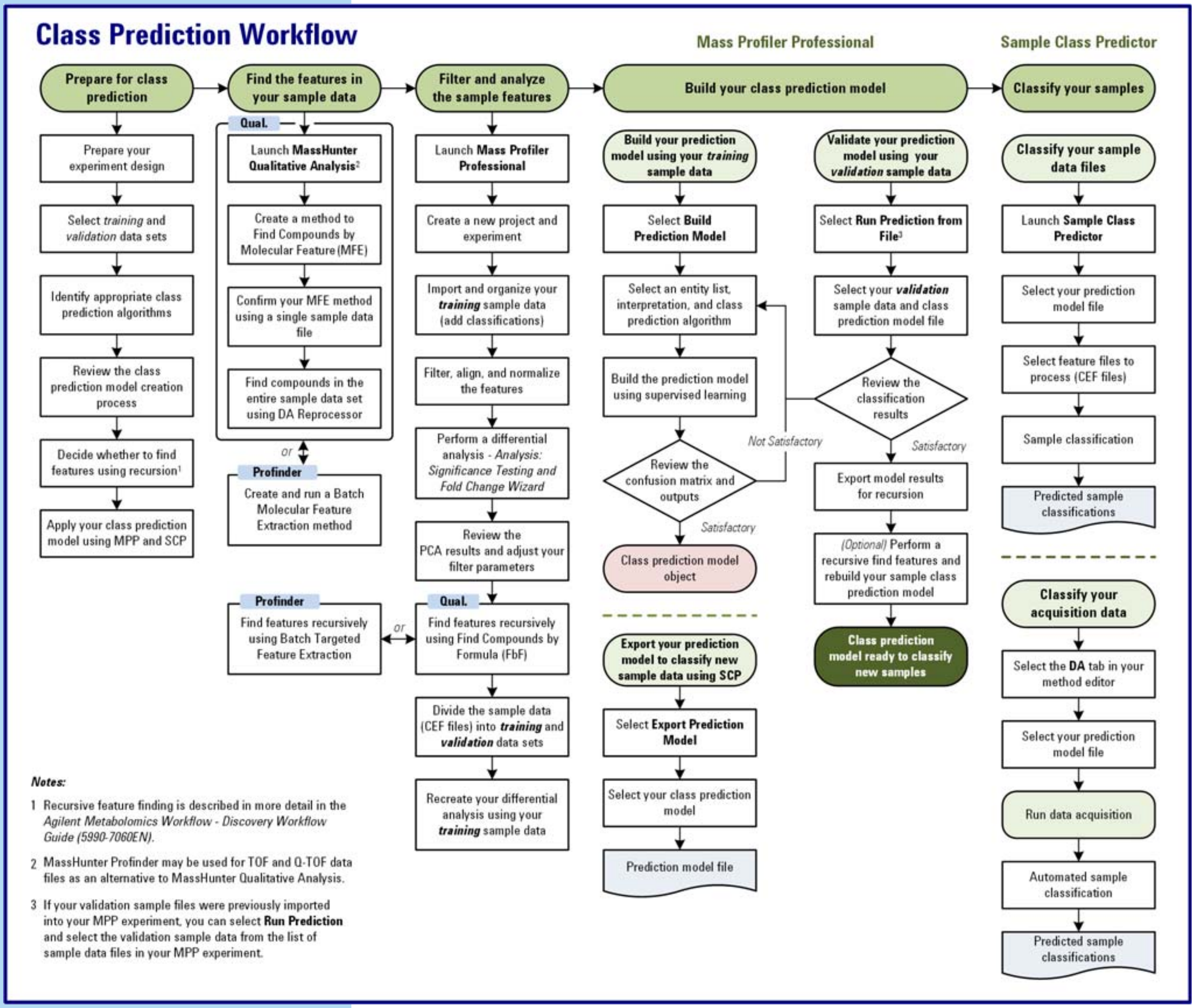
Run data acquisition

Predicted sample classifications

Agilent Technologies

Introduction

Class prediction is the process of building a statistical model that is used to predict the class membership of unknown samples based on the prior analysis and interpretation of abundance profiles of the features in samples with known class membership. The chapters of this workflow guide follow the flow illustrated in Figure 1.


Class Prediction Workflow
Mass Profiler Professional
Sample Class Predictor
Prepare for class prediction
Find the features in your sample data
Filter and analyze the sample features
Build your class prediction model
Classify your samples
Prepare your experiment design
Select training and validation data sets
Identify appropriate class prediction algorithms
Review the class prediction model creation process
Decide whether to find features using recursion¹
Apply your class prediction model using MPP and SCP
Qual.
Launch MassHunter Qualitative Analysis²
Create a method to Find Compounds by Molecular Feature (MFE)
Confirm your MFE method using a single sample data file
Find compounds in the entire sample data set using DA Reprocessor
or
Profinder
Create and run a Batch Molecular Feature Extraction method
Profinder
Find features recursively using Batch Targeted Feature Extraction
or
Launch Mass Profiler Professional
Create a new project and experiment
Import and organize your training sample data (add classifications)
Filter, align, and normalize the features
Perform a differential analysis - Analysis: Significance Testing and Fold Change Wizard
Review the PCA results and adjust your filter parameters
Qual.
Find features recursively using Find Compounds by Formula (FbF)
Divide the sample data (CEF files) into training and validation data sets
Recreate your differential analysis using your training sample data
Build your prediction model using your training sample data
Select Build Prediction Model
Select an entity list, interpretation, and class prediction algorithm
Build the prediction model using supervised learning
Review the confusion matrix and outputs
Not Satisfactory
Satisfactory
Class prediction model object
Export your prediction model to classify new sample data using SCP
Select Export Prediction Model
Select your class prediction model
Prediction model file
Validate your prediction model using your validation sample data
Select Run Prediction from File³
Select your validation sample data and class prediction model file
Review the classification results
Satisfactory
Export model results for recursion
(Optional) Perform a recursive find features and rebuild your sample class prediction model
Class prediction model ready to classify new samples
Classify your sample data files
Launch Sample Class Predictor
Select your prediction model file
Select feature files to process (CEF files)
Sample classification
Predicted sample classifications
Classify your acquisition data
Select the DA tab in your method editor
Select your prediction model file
Run data acquisition
Automated sample classification
Predicted sample classifications

Notes:

1 Recursive feature finding is described in more detail in the *Agilent Metabolomics Workflow - Discovery Workflow Guide (5990-7060EN).*

2 MassHunter Profinder may be used for TOF and Q-TOF data files as an alternative to MassHunter Qualitative Analysis.

3 If your validation sample files were previously imported into your MPP experiment, you can select **Run Prediction** and select the validation sample data from the list of sample data files in your MPP experiment.

Figure 1 *Class prediction workflow*

Agilent MassHunter Mass Profiler Professional is used in conjunction with Agilent MassHunter Qualitative Analysis/DA Reprocessor or Agilent MassHunter Profinder to build and validate class prediction models. After you create your class prediction model using Mass Profiler Professional, you can use Mass Profiler Professional to manually run class prediction to classify new samples.

Agilent MassHunter Sample Class Predictor adds value by automating your sample class prediction with acquisition and helping you save time and money in your analyses by providing real-time quality assurance and control of your data. With Sample Class Predictor you can integrate the class prediction models you create using Mass

Profiler Professional within Agilent Data Acquisition and automatically classify new samples.

This class prediction workflow guide is part of a series of workflow guides developed to help you to analyze your sample data using Mass Profiler Professional. Other workflow guides available in this series include the *Agilent Metabolomics Workflow - Discovery Workflow Guide* (5990-7067EN) and the *Integrated Biology with Agilent Mass Profiler Professional - Workflow Guide* (5991-1909EN).

To increase your confidence in obtaining reliable and statistically significant results, review the chapter "Prepare for an experiment" in the *Agilent Metabolomics Workflow - Discovery Workflow Guide* and make sure your analysis includes a carefully thought-out experimental design that includes the collection of replicate samples.

More information

The *Class Prediction with Agilent Mass Profiler Professional - Workflow Guide* is part of the collection of Agilent manuals, help, application notes, and videos. The current collection of manuals and help are valuable to users who understand class prediction and the metabolomics workflow and who may require familiarization with the Agilent software tools. Video tutorials for MPP provide step-by-step instructions to analyze example GC/MS and LC/MS data files. This workflow provides a step-by-step overview of performing class prediction using MassHunter Qualitative Analysis, Profinder, and Mass Profiler Professional.

The following selection of publications provides materials related to class prediction, metabolomics, and MassHunter software used to analyze your sample data:

- *Manual:* Agilent G3835AA MassHunter Profinder Software - Quick Start Guide *(G3835-90014, Revision A, December 2013)*
- *Manual:* Integrated Biology with Agilent Mass Profiler Professional - Workflow Guide (5991-1909EN, Revision A, June 2013)
- *Manual:* Integrated Biology with Agilent Mass Profiler Professional - Workflow Guide Overview (5991-1910EN, Revision A, June 2013)
- *Manual:* Agilent Metabolomics Workflow - Discovery Workflow Guide (5990-7067EN, Revision B, October 2012)
- *Manual:* Agilent Metabolomics Workflow - Discovery Workflow Overview (5990-7069EN, Revision B, October 2012)
- *Manual:* Agilent G3835AA MassHunter Mass Profiler Professional - Quick Start Guide (G3835-90009, Revision A, November 2012)
- *Manual:* Agilent G3835AA MassHunter Mass Profiler Professional - Familiarization Guide (G3835-90010, Revision A, November 2012)
- *Manual:* Agilent G3835AA MassHunter Mass Profiler Professional - Application Guide (G3835-90011, Revision A, November 2012)
- *Presentation:* Advances in Instrumentation and Software for Metabolomics Research (Advances in Instrumentation and Software for Metabolomics.pdf, September 18, 2012)
- *Presentation:* Two Workflows to Support Automated Class Prediction with Complex Samples (WP20_405_Two_Workflows_Support_Automated_Class_Prediction.pdf, June 25, 2012)
- *Presentation:* Predictive Classification of Contaminants Encountered During the Distillation of Shochu, a Distilled Beverage Native to Japan (ASMS_2011_ThP_316.pdf, June 23, 2011)

- *Brochure:* Agilent Solutions for Metabolomics (5990-6048EN, April 30, 2012)
- *Brochure:* Agilent Mass Profiler Professional Software (5990-4164EN, April 27, 2012)
- *Application:* Detecting Contamination in Shochu Using the Agilent GC/MSD, Mass Profiler Professional, and Sample Class Prediction Models (5991-0975EN, August 2, 2012)
- *Application:* Metabolomic Profiling of Wines using LC/QTOF MS and MassHunter Data Mining and Statistical Tools (5990-8451EN, June 22, 2011)

A complete list of references may be found in “References” on page 172.

NOTE

This manual gives publication numbers to most references. You can easily download the documents from the Agilent literature library.

Go to the Agilent literature library at www.chem.agilent.com/en-US/Search/Library and type the publication number or the publication title in the search box.

Then click the **Search** 🔍 button.

“Definitions” on page 160 contains a list of terms and their definitions as used in this workflow.

Required items

To build a class prediction model

- Data files from an Agilent GC/MS or LC/MS system
- PC running Windows 7
- Qualitative Analysis/DA Reprocessor or Profinder
- Mass Profiler Professional
- ID Browser

To automate class prediction

- PC running Windows
- MassHunter Data Acquisition or GC/MSD ChemStation Software
- Sample Class Predictor Software

The Class Prediction with Mass Profiler Professional and Sample Class Predictor workflow performs best when using the hardware and software described in the "required" sections below. The required hardware and software is used to perform the data acquisition and analysis tasks shown in Figure 2.

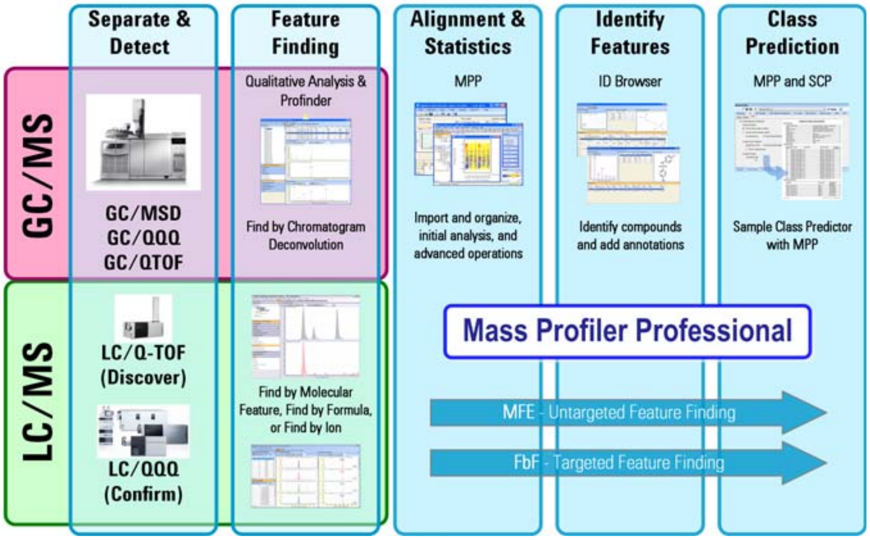

Separate & Detect
Feature Finding
Alignment & Statistics
Identify Features
Class Prediction
GC/MS
LC/MS
GC/MSD
GC/QQQ
GC/QTOF
Qualitative Analysis & Profinder
Find by Chromatogram Deconvolution
MPP
Import and organize, initial analysis, and advanced operations
ID Browser
Identify compounds and add annotations
MPP and SCP
Sample Class Predictor with MPP
LC/Q-TOF (Discover)
LC/QQQ (Confirm)
Find by Molecular Feature, Find by Formula, or Find by Ion
Mass Profiler Professional
MFE - Untargeted Feature Finding
FbF - Targeted Feature Finding

Figure 2 *Agilent hardware and software used to acquire and analyze your samples following the class prediction workflow. Sample separation to class prediction typically involves either or both GC/MS and LC/MS analyses.*

Required hardware

- PC running Windows
  - *Minimum:* Windows 7 (32-bit or 64-bit) with 4 GB of RAM
  - *Recommended:* Windows 7 (64-bit) with 8 GB or more of RAM
  - At least 50 GB of free space on the C:\ partition of the hard drive
- An Agilent chromatography mass spectrometry system (for example, GC/MS, LC/MS, and CE/MS) to generate the sample data files used in this workflow.

Required software

- Mass Profiler Professional Software B.12.00 or later

  Mass Profiler Professional software is a chemometrics software package designed to exploit the high information content of mass spectrometry data. Researchers can easily import, analyze and visualize GC/MS, LC/MS, CE/MS and ICP-MS data from large sample sets and complex MS data sets.

  Mass Profiler Professional integrates smoothly with MassHunter Workstation and ChemStation software, and is designed for analyzing data from any MS-based application where you need to determine relationships among sample groups and variables, including metabolomics, proteomics, food safety, environmental, forensics and toxicology.

For metabolomics and proteomics studies, the optional Agilent Pathway Architect software helps you evaluate MS data in biological context.

- MassHunter Qualitative Analysis software, Version B.06.00 SP1 or later

  MassHunter Qualitative Analysis software automatically finds and extracts all spectral and chromatographic information from a sample, even when the components are not fully resolved. Powerful data navigation capabilities permit you to browse through compound-specific information in a single sample and compare chromatograms and spectra among multiple samples. The software also includes a customizable user interface and the capability to save, export or copy results into other applications.

- MassHunter Profinder B.06.00 or later

  MassHunter Profinder software is a feature finding application that is optimized specifically for sample data files from TOF and Q-TOF LC/MS instruments and allows you to easily visualize, review and edit compound results across multiple different samples. Profinder speeds up your differential analysis workflows by providing a batch feature extraction software tool for raw mass spectrometric data - for TOF and Q-TOF sample data files Profinder may be used in place of Qualitative Analysis and DA Reprocessor.

- MassHunter Data Acquisition software, Version B.06.00 or later (this includes MassHunter DA Reprocessor)

  The DA Reprocessor program automates the application of your data analysis methods to multiple sample data files by running a worklist which starts the Qualitative Analysis program for each file and method in the worklist. DA Reprocessor is a utility that is shipped with MassHunter Data Acquisition software. DA Reprocessor is included on the Data Acquisition Utilities disk. See the *Data Acquisition Installation Guide* for information on installing this program. The version B.0X of the MassHunter Data Acquisition software must match the version of MassHunter Qualitative Analysis software.

- MassHunter ID Browser B.05.00 or later

  ID Browser, which is built into Mass Profiler Professional and Qualitative Analysis, performs compound identification using:
  - LC/MS Personal Compound Database (METLIN, pesticides, forensics)
  - GC/MS libraries (NIST and Fiehn library)
  - Empirical Formula Calculation using the Agilent Molecular Formula Generator (MFG) algorithm

Compounds can be quickly and easily identified through integration within the Mass Profiler Professional environment. ID Browser automatically annotates the entity list and puts the compound names onto any of the various visualization and pathway analysis tools.

Optional software

- MassHunter Quantitative Analysis software, Version B.06.00 or later

  The MassHunter Quantitative Analysis software supports simple and efficient review of large multi-compound quantitation batches. A graphical “Batch-at-a-Glance” interface facilitates navigating results by compound or sample, or switching between the two approaches. A sophisticated quantitation engine helps you set up over 20 different outlier criteria, and a parameter-less integrator facilitates reliable unsupervised quantitation. The ability to filter results and focus on outliers or questionable peak integrations significantly reduces the data review time for large multi-compound batches. A method task editor and “Curve-Fit Assistant” provide for simple method and multi-level calibration setup.

- Agilent ChemStation software

  ChemStation handles a wide variety of separation techniques such as GC, GC/MS, LC, LC/MS, CE and CE/MS. ChemStation is a scalable data system ideally suited for applications in all industries ranging from early product development to quality control.

- AMDIS

  AMDIS is an acronym for the automated mass spectral deconvolution and identification system developed by NIST. (http://www.amdis.net) AMDIS helps analyze GC-MS data of complex mixtures, even data with strong background ions and coeluting peaks. AMDIS is not for use with data collected in SIM mode. AMDIS automatically extracts pure (background free) component mass spectra from highly complex GC-MS data files and uses these purified spectra when searching a mass spectral library.

- METLIN Personal Compound Database and Library

  METLIN personal compound database and library (PCDL) contains over 25,000 compounds, including 8,000 lipids with retention times for about 700 standards. Used with TOF and Q-TOF data, identification is enabled using accurate mass and/or retention time database searching. Searching the MS/MS spectral library with more than 2,200 compounds enables more confident identification. PCDL represents a data management system designed to assist in a broad array of metabolite research and metabolite identification by providing public access to its repository of current and comprehensive mass spectral metabolite data. (http://metlin.scripps.edu/)

- Agilent Fiehn GC/MS Metabolomics Library

  The Fiehn GC/MS Metabolomics RTL Library is a growing metabolomics-specific library that contains searchable EI spectra and retention-time indexes for approximately 800 common metabolites. The Fiehn library integrates with Agilent's other software tools for GC/MS metabolomics.

Compliance

21 CFR Part 11 is a result of the efforts of the US Food and Drug Administration (FDA) and members of the pharmaceutical industry to establish a uniform and enforceable standard by which the FDA considers electronic records equivalent to paper records and electronic signatures equivalent to traditional handwritten signatures. For more information, see

http://www.fda.gov/RegulatoryInformation/Guidances/ucm125067.htm

MassHunter Data Acquisition Compliance Software includes the following features which support 21 CFR Part 11 compliance:

- Hash Signature for data files let you check the integrity of files during a compliance audit
- Roles that restrict actions to certain users
- Method Audit Trail Viewer

MassHunter Quantitative Analysis Compliance Software includes the following features which support 21 CFR Part 11 compliance:

- Security measures ensuring the integrity of acquired data, analysis, and report results
- Comprehensive audit-trail features for quantitative analysis, using a flexible and configurable audit-trail map
- Customizable user roles and groups let an administrator individualize user access to processing tasks

Before you begin creating methods and submitting studies, you may decide to install MassHunter Data Acquisition Compliance Software and MassHunter Quantitative Analysis Compliance Software.

The Quantitative Analysis Compliance program is installed separately from the Quantitative Analysis program. See *Agilent MassHunter Quantitative Analysis Compliance Software Quick Start Guide* (G3335-90099, Revision A, February 2011) for instructions on installing the Compliance program.

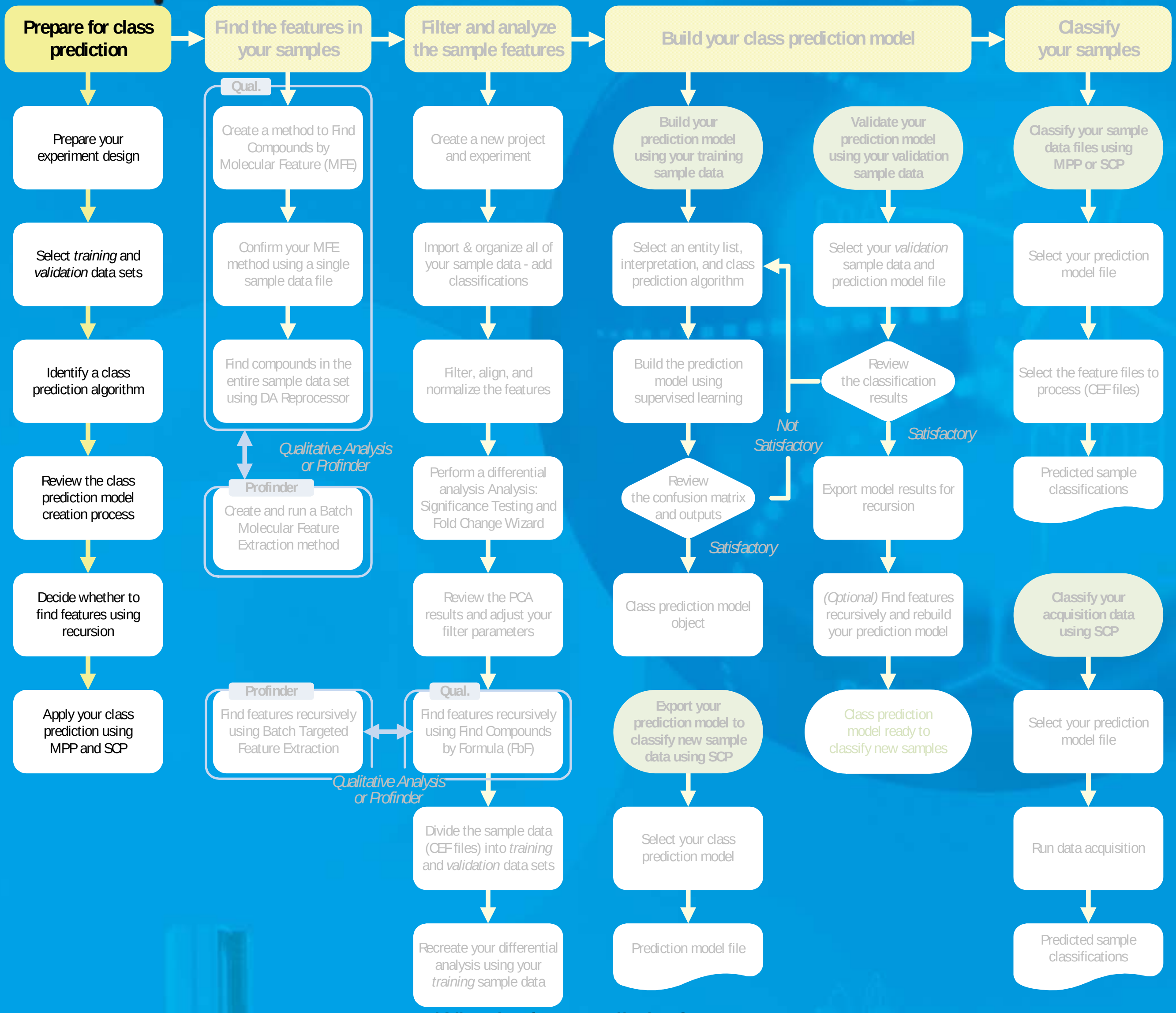
The Data Acquisition Compliance program is installed automatically with the MassHunter Data Acquisition software. See *Agilent MassHunter Data Acquisition Compliance Software Quick Start Guide* (G3335-90098, Revision A, February 2011) for instructions on enabling and using the MassHunter Compliance Software.

Roles

When Compliance is enabled, only certain users can perform certain actions. For example, the user that logs on to the system to submit a study needs to have certain Quantitative Analysis privileges to automatically build the quantitative analysis method.

Prepare for class prediction

This section presents a sequence of steps that provide training material and an overview that is important to understand before you perform the class prediction workflow. An introduction is presented to the statistical algorithms available in class prediction.



Prepare for class prediction

Prepare your experiment design

Select *training* and *validation* data sets

Identify a class prediction algorithm

Review the class prediction model creation process

Decide whether to find features using recursion

Apply your class prediction using MPP and SCP

Find the features in your samples

Qual.

Create a method to Find Compounds by Molecular Feature (MFE)

Confirm your MFE method using a single sample data file

Find compounds in the entire sample data set using DA Reprocessor

Qualitative Analysis or Profinder

Profinder

Create and run a Batch Molecular Feature Extraction method

Profinder

Find features recursively using Batch Targeted Feature Extraction

Filter and analyze the sample features

Create a new project and experiment

Import & organize all of your sample data - add classifications

Filter, align, and normalize the features

Perform a differential analysis Analysis: Significance Testing and Fold Change Wizard

Review the PCA results and adjust your filter parameters

Qual.

Find features recursively using Find Compounds by Formula (FbF)

Qualitative Analysis or Profinder

Divide the sample data (CEF files) into *training* and *validation* data sets

Recreate your differential analysis using your *training* sample data

Build your class prediction model

Build your prediction model using your training sample data

Select an entity list, interpretation, and class prediction algorithm

Build the prediction model using supervised learning

Review the confusion matrix and outputs

Satisfactory

Class prediction model object

Export your prediction model to classify new sample data using SCP

Select your class prediction model

Prediction model file

Validate your prediction model using your validation sample data

Select your *validation* sample data and prediction model file

Review the classification results

Not Satisfactory

Satisfactory

Export model results for recursion

(*Optional*) Find features recursively and rebuild your prediction model

Class prediction model ready to classify new samples

Classify your samples

Classify your sample data files using MPP or SCP

Select your prediction model file

Select the feature files to process (CEF files)

Predicted sample classifications

Classify your acquisition data using SCP

Select your prediction model file

Run data acquisition

Predicted sample classifications

Agilent Technologies

What is class prediction?

Class prediction is the process you use within Mass Profiler Professional (MPP) to build, validate, test, and export a class prediction model that is developed based on the abundance profiles of the features in samples with known classification. The class prediction model created within MPP is subsequently used by Sample Class Predictor (SCP) to integrate class prediction within data acquisition. Using this latter, separately licensed program, you automate the prediction model for real-time QA/QC of your samples. If you do not have a Sample Class Predictor license you can manually process all of your data within Mass Profiler Professional.

This workflow guide helps you use Mass Profiler Professional to build a prediction model that is used to classify samples acquired using chromatography/mass spectrometry. The available prediction model algorithms learn from samples that have known functional class membership (training and validation data sets) to build a prediction model that classifies new samples (test data sets) into the known classes. Class prediction involves ten main steps:

(1) Prepare your experiment design to include a large number of replicates for each of the known classifications so that the sample data cover a range of variables such as operator, instrument condition, run order, sample preparation, and subject,

(2) Find the molecular features in *all* of your sample data files using Qualitative Analysis/DA Reprocessor or Profinder,

(3) Create a differential analysis using Mass Profiler Professional,

(4) Recursively find the targeted features in all of your sample data files using Qualitative Analysis/DA Reprocessor or Profinder,

(5) Partition your sample data into *training* and *validation* data sets,

(6) Recreate your differential analysis with the *training* data set using Mass Profiler Professional,

(7) Select one or more prediction model algorithms that support your hypothesis, experiment design, and expected interrelationships of the features among the classifications,

(8) Build your class prediction model using the *training* data set and supervised learning,

(9) Validate your class prediction model using the *validation* data set, known samples that were not used during the model creation, and

(10) Apply your prediction model to samples with unknown classification.

Class prediction assigns functional classifications to your samples based on the abundance profile of the features (compounds) identified in your training data set. Possible classifications can include material origin, biological material maturity, production quality, clinical and/or sample treatments, diseases, and other conditions. Class prediction helps you:

- Predict the class membership (parameter value) of a sample
- Identify chemical (metabolite) signatures that discriminate well among classes
- Identify samples that could be potential outliers

A simple view of the class prediction workflow is illustrated in Figure 3 on page 15 starting with data acquisition through to class prediction involving either LC/MS or

GC/MS analyses. Find by Molecular Feature, also referred to as molecular feature extraction (MFE) and Find by Formula (FbF) are two different algorithms used by MassHunter Qualitative Analysis for finding compounds. All results files generated by Agilent analytical platforms can be imported into Mass Profiler Professional for quality control, statistical analysis and visualization, and interpretation.

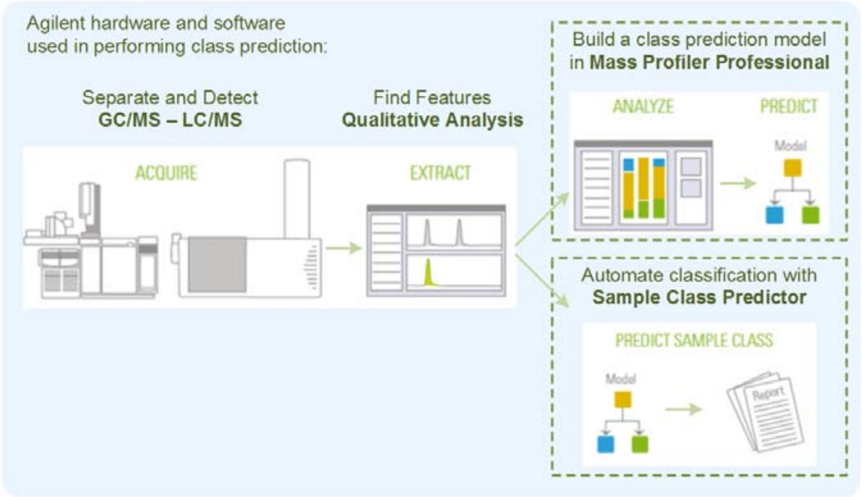

Agilent hardware and software used in performing class prediction:
Separate and Detect
GC/MS – LC/MS
Find Features
Qualitative Analysis
ACQUIRE
EXTRACT
Build a class prediction model in Mass Profiler Professional
ANALYZE
PREDICT
Model
Automate classification with Sample Class Predictor
PREDICT SAMPLE CLASS
Model
Report

Figure 3 *An Agilent class prediction workflow from separation to classification involving either GC/MS or LC/MS analyses.*

Experiment variables

Experiment variables and classifications are derived from your experiment. Manipulated attributes of the state of the organism are referred to as independent variables. The biological response to the change in the attributes may manifest in a change in the metabolic profile. Each metabolite that undergoes a change in expressed concentration is referred to as a dependent variable. Metabolites that do not show any change with respect to the independent variable may be valuable as control or reference signals.

The features in a sample may be individually referred to as a **compound**, **descriptor**, **element**, **entity**, **feature**, or **metabolite** during the various steps of the class prediction workflow. When hundreds to thousands of dependent variables (e.g., metabolites) are available, chemometric data analysis is employed to reveal accurate and statistically meaningful correlations between the attributes (independent variables) and the metabolic profile (dependent variables). Meaningful information learned from the metabolite responses can subsequently be used for clinical diagnostics, for understanding the onset and progression of human diseases, and for treatment assessment. Therefore, metabolomic analyses are poised to answer questions related to causality and relationship as applied to chemically complex systems, such as organisms.

Sample sets

Robust prediction models are developed using large sample sets. Each sample set should contain replicate samples from all of the known classifications so that the sample data cover a range of variables such as operator, instrument condition, run order, sample preparation, and subject. Models developed from samples that cover a sufficiently large range of classifications and contain a large number of replicates can be considered to be generic to the population. The class prediction model developed using these samples can therefore be expected to be able to classify any biological sample taken from the population.

Before collecting your samples, you first determine the range of phenotypes that your prediction model spans and then establish the number of replicate samples to collect for each combination of phenotype and parameter condition to produce a model that is able to handle a wide range of sample quality.

To train a prediction model, you assign classifications to each sample, find features, and perform differential analyses until you have identified the entities and model algorithm that is able to differentiate your training sample data into the known classifications.

After you have collected your sample data, the next step is to find the molecular features in all of your sample data.

Prepare your experiment design

You start with a sound experiment design that includes a large sample set containing replicate samples from all of the known classifications. If your sample preparation includes sample treatments that can improve expression of the analytes for your study, include replicate samples for each of the sample treatments in your sample set. A robust prediction model is generated by replicate samples that cover a range of variables such as operator, instrument condition, run order, sample preparation, and subject.

Your data set should (1) cover the entire analytical space, (2) account for wide variations in sample quality where the data spans from good to poor due to variations in sampling (study quality), sample preparation, and instrumental performance, and (3) include a large number of replicates.

To increase your understanding of preparing your experiment design review, the chapter “Prepare for an experiment” in the *Agilent Metabolomics Workflow - Discovery Workflow Guide*. Since an understanding of natural variability and replicates is particularly important to class prediction, these sections are also presented in this chapter.

Natural variability

Before any statistical analysis is begun, it is important to understand how a sample taken from any one specimen represents the population as a whole and how increasing the sample size improves the accuracy of the sample set in describing characteristics of the population.

Under identical conditions, all life systems produce a range of results. Specimens taken from the population may show one of the following characteristics:

- Results comparable to the mean of the population (i.e., characteristics shown by the majority of the population), for example results within ±1 standard deviation (~68%) from the mean.
- Results that differ significantly from those shown by the majority of the population (i.e., characteristics that are not shown by the majority of the population), for example results beyond ±3 standard deviations (~99.7%) from the mean.
- Results anywhere in between ±1 to ±3 standard deviations from the mean, and beyond.

In many biological and biochemical systems characteristics are found to show a probability of variation referred to as a normal distribution. Figure 4 on page 18, for instance, shows a normal distribution of a characteristic within a population where 68% of the sampled population would be shown to have the mean characteristic plus or minus one standard deviation (s). This natural variation of the population response to identical conditions is referred to as natural variability. Natural variability thus means that any single sample specimen taken from a population is not guaranteed to reflect the mean characteristics of the population.

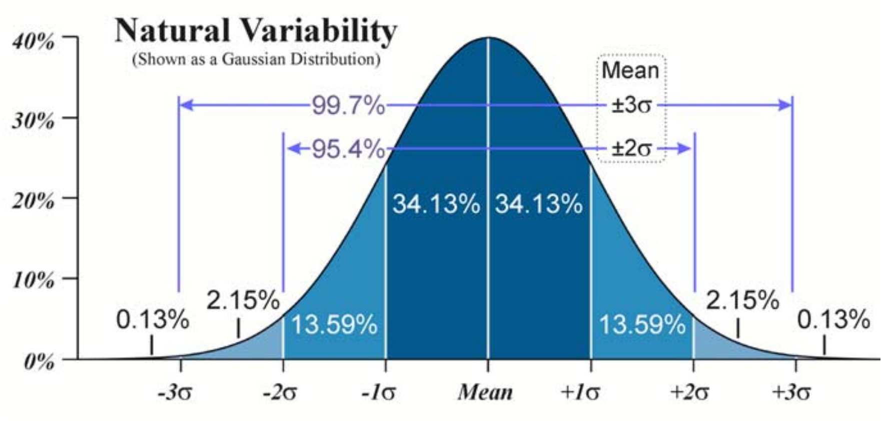


Figure 4 *Natural variability shown as a Gaussian (normal) distribution. Depending on the predefined requirement for significance, if the mean of a sample set is beyond ±2s from the natural variation there may be a significant effect. Similarly, if a particular observation routinely falls beyond ±2s from the natural variability of the data the change producing the effect may be considered significant.*

Natural variability occurs from inherent randomness and unpredictability in the natural world. Natural variability is found in all life and natural sciences and in all forms of engineering. For example, a population of plants grown under identical conditions of illumination, precipitation, and nutrient availability shows a range in growth mass per day. This range of variable growth mass may be expressed as a mean where 95% of the population is expected to show a natural range of variability within two standard deviations of the mean (see Figure 4).

In other words, for a set of fixed attributes (independent variables), a representative set of samples taken from the population of plants shows a natural variability in the dependent variable, such as daily growth mass. When an experiment is undertaken where plants from the same population are subject to variations in the fixed attributes, the plants response shows a change in growth mass in addition to their natural variability in growth mass. Thus if the entire population is sampled, two adjacent normal distributions are obtained with means reflecting the plant growth mass under the two conditions (see Figure 5). Such unpredictability in the measurable variability of any biochemical expression must not be mistakenly correlated with deliberate variations of an independent variable.

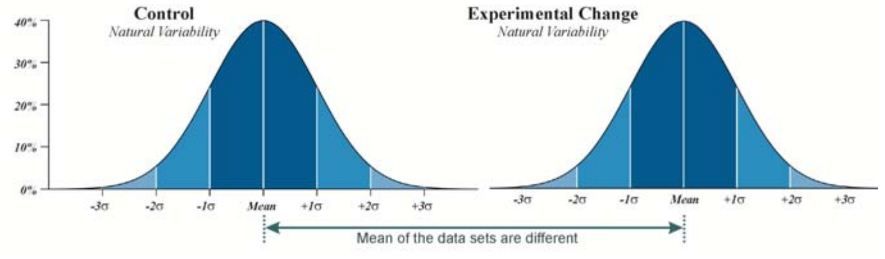


Figure 5 *Natural variability of populations that are subject to two different experimental conditions where the mean of each data set falls outside of the mean ±3s of the other data set.*

During your class prediction experiment the natural variability of the data representing a population must be understood in order to confidently express your experimental correlation. The investment of time and resources in performing statistical analyses and class prediction requires that the natural variability of the subject specimen be known or reasonably estimated so that the results of the analysis may be conclusively shown to be either within the natural variability (no correlation) or outside of the natural variability and therefore provide for a degree of correlation with the independent variable(s).

Experimental data collection that does not incorporate consideration of the natural variability of the data does not yield meaningful results. Thus, crucial to the class prediction workflow, as with all statistical data treatments, is an understanding and well planned collection of the data; without that, the results follow the adage “garbage in, garbage out.”

Replicate data

Replicate sampling and measurement of many specimens from the population is the only way to estimate the natural variability of your data. No guarantee exists that a single sample specimen from a population represents the mean of the population. Therefore to create the most confident class prediction model possible many sample specimens from the population are necessary. Any single sample from a population with a natural variability shown in Figure 4 on page 18 has a 99.99% chance that it lies within four standard deviations (±4*s*) of the mean of the true population, but in fact that single sample may on a rare occasion fall even further from the population mean.

If ten (10) samples are taken from the population, the mean of these samples produces a statistically more accurate approximation of the true mean of the population. The accuracy of the approximation of the true population mean proportionally improves with more samples. The true value of the population mean is achieved only if the entire population is sampled. However, sampling the entire population is not typically feasible because of constraints imposed by time, resources, and finances. On the other hand, evaluating fewer samples increases the chance of false negative and false positive correlations from your experiment.

Too few samples may lead to an incorrect conclusion; you may have an inaccurate class prediction model. Figure 6 on page 20 shows that if too few samples are evaluated, and if these samples just happen to be samples lying far from the mean because of natural variability, an incorrect conclusion may be drawn that the change in the independent variable produced no significant change in the response. The estimate of the standard deviation of the sample mean estimate of a population mean (standard error) is equal to the standard deviation of the samples divided by the square root of the number of samples (Equation 1).

Equation 1 $$SE = \sigma/(\sqrt{N})$$

where SE = standard error of the population; σ = standard deviation; and N = number of samples.

Large sample sizes lead to more confident conclusions; you create a more accurate class prediction model. As the sample size increases, the likelihood that the data approximates the true response of the population increases. The standard deviation of the sample may become smaller, and the likelihood of making a correct correlation between cause and effect is improved (Equation 2).

Equation 2
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})}$$

where σ = standard deviation; N = number of samples; x_i = the value of an individual sample; and $\bar{X}$ = mean (or average) of all N samples.

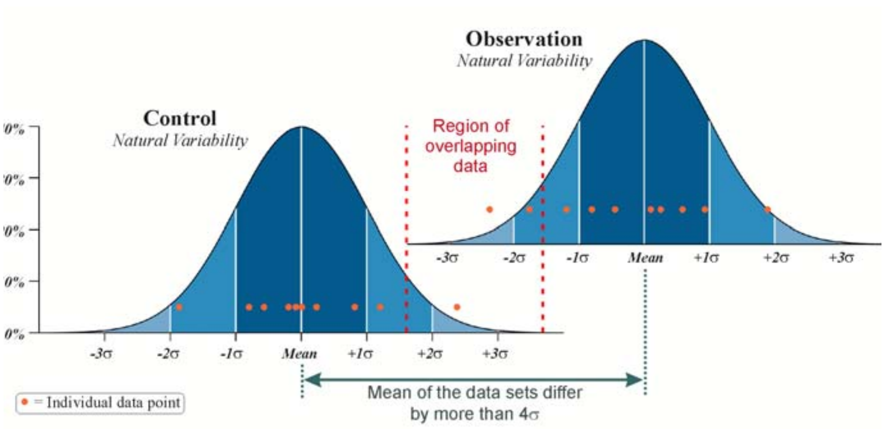


Figure 6 *Replicate data are necessary to distinguish whether the represented populations actually show significant differences. The three data points in the region of overlap of the natural variability may, if too few replicates are selected, lead to a result suggesting a less significant difference in the populations.*

Successful application of class prediction analyses depends on the availability of sufficient replicate samples and specimens. Coupled with an understanding of the systems under study and a well planned collection of the samples and concomitant data, the statistical data treatment of the replicate samples is the backbone of the class prediction workflow. A sufficient set of replicate data, ten (10) or more replicates, may provide a significant answer to the hypothesis and prevent a total loss of time and resources invested in performing the described statistical analyses.

Acquire your sample data

For your class prediction model it is recommended to acquire at least ten (10) replicates for each combination of phenotype and parameter condition (within an independent variable or within each permutation of parameters when more than one independent variable exists). Too few samples increase the chance for obtaining a false negative or false positive sample classification result.

Features of the example experiment

This class prediction workflow is illustrated using a metabolomics experiment with one independent variable (parameter name) containing four conditions (parameter values). The class prediction workflow used to generate results from this experiment helps you use MassHunter Qualitative Analysis, DA Reprocessor, Profinder and Mass Profiler Professional to perform class prediction with your own sample data files.

The example experiment used in this workflow guide presents an analysis of a metabolomic response to changes in a single independent variable, also referred to as a parameter. The data was acquired using four (4) parameter values for the independent variable. The parameter values consist of a single control data set that represents the organism without perturbation and data sets from three variations where the organism is subject to one of three conditions established by the experiment design.

An ideal experiment involves at least ten (10) replicates for each parameter value. Thus, an ideal experiment with a single parameter and four parameter values has a data sample size of at least forty (40) samples. In this example the minimum sampling conditions are met with ten replicates per parameter value. The sample data files are organized in a set of folders under **C:\MassHunter\Data** as illustrated in Figure 7.

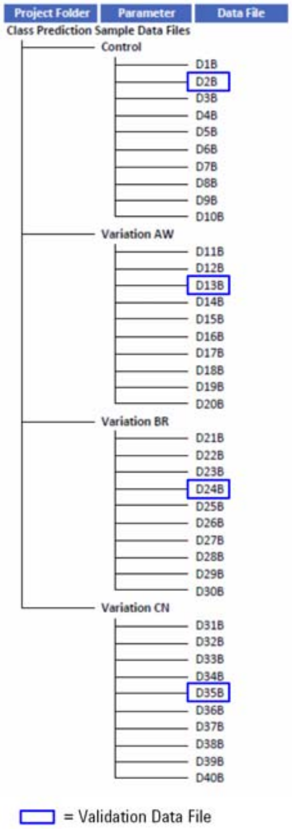


Figure 7 *Example file list used in this workflow guide. Four files are selected to use to validate the class prediction model.*

Select training and validation data sets

The process of learning feature abundance profiles from known sample data is called ***training***. Training requires a data set in which class membership of the samples are known. To build a class prediction model you start with a sound experiment design that includes a large sample set containing replicate samples from all of the known classifications. Provided that you have sufficient replicates for each of your classes, you begin training by setting aside a subset of your sample data, typically one or more sample data files representing each of the classifications, to use as a *validation data set*. The remainder of the data, the *training data set*, is used to train your prediction model. The training sample data files are used to build your class prediction model; the validation data files are used as a final quality inspection of your class prediction model.

Supervised learning and final quality inspection

Two types of validation are performed in the class prediction workflow - supervised learning and final quality inspection.

During ***supervised learning*** the *training data set* is partitioned into a number of sample groups containing at least one sample data file from each of the classes. One sample group is treated as a validation data set; the remaining sample groups are treated as the training data set and used to build a class prediction model. The sample group representing the validation data set is then used to evaluate the performance of the class prediction model. Then the model/validation process is repeated using a different sample group as the validation data set and the remaining sample groups to train the class prediction model. When all of the sample groups have been treated in turn as the validation data set the model/ validation process is complete, you can review the model results via a confusion matrix (a data table that is used to assess the ability of a prediction model to correctly predict known classifications).

When you are satisfied with the results from the supervised learning, you then use the *validation data set* that you set aside earlier for the ***final quality inspection*** of your class prediction model. Once your model is validated, the abundance profiles, also referred to as signatures, can be used to predict the membership of new samples.

Identify appropriate class prediction algorithms

Class prediction involves a process of assigning a condition (also referred to as an attribute value, parameter value, or class) to a new sample on the basis of a mathematical/statistical model created using a training sample set of data whose conditions are known. Therefore class prediction is an instance of machine learning that is based on supervised learning, i.e., creating mathematical representations of class membership by applying a class prediction algorithm to a set of correctly identified samples. Cluster analysis (referred to as clustering in MPP), on the other hand, is an example of unsupervised learning where the samples are grouped into categories based on some measure of inherent similarity without any prior knowledge of sample identification. To build a class prediction model you must select a class prediction algorithm for the supervised learning that is appropriate to your experiment.

Supervised learning

Supervised learning is a process employed in data analysis that uses knowledge of the phenotype to simplify the data (for example, reduce the number of entities) to retain the entities that provide the best correlation to the characteristics (conditions) involved in the particular analysis. The goal of supervised learning is to optimize a mathematical relationship that accurately associates the entities in your samples to the conditions in your interpretation; for example, when you are evaluating qualitative samples representing disease versus healthy samples, or when you are evaluating quantitative samples representing degree of disease progression or response to therapy. In order to perform supervised learning your training data must be properly classified.

Variable definitions

Variables are derived from your experiment design. When one or more of the attributes of the state of the organism under study are manipulated, those attributes are referred to as *independent variables*. The biological response to the change in the attributes may manifest in a change in the metabolic profile expressed by the organism. Each metabolite that undergoes a change in expressed concentration is referred to as a *dependent variable*. Metabolites that do not show any change with respect to the independent variable may be valuable as control or reference signals.

Throughout this workflow guide an *independent variable* is referred to as a parameter name. The attribute values within an independent variable are referred to as parameter values and at times may be referred to as a condition, an attribute, a parameter value, or class.

Similarly, throughout this workflow guide, a *dependent variable* may at times be referred to as a feature, compound, metabolite, element, descriptor, or entity.

Class prediction algorithms

MPP supports five different class prediction algorithms (machine learning algorithms). The advantages for using any one of the supervised learning algorithms is often a matter of subjective opinion and your personal experience in applying the particular algorithm to your experiment and experimental parameters. The available class prediction algorithms are:

1. Partial Least Squares Discrimination
2. Support Vector Machine
3. Naïve Bayes
4. Decision Tree
5. Neural Network

Except for Partial Least Squares Discrimination, the algorithms used by Class Prediction are also available within the advanced operation Find Minimal Entities.

You can access these algorithms in the Workflow Browser by selecting **Build Prediction Model**. Each algorithm creates a class prediction model when you complete the training. You can use the class prediction models to predict the functional class membership of new samples and different experiments by selecting **Run Prediction** in the Workflow Browser.

In general, each of the class prediction algorithms, also referred to as supervised learning algorithms, have the following features. The first paragraph for each algorithm presents a general summary of the best application of the algorithm.

Partial Least Squares Discrimination (PLS-DA)

PLS-DA is best suited for making classifications where all of the parameter values are measurable with little error, where the number of samples is smaller than the number of parameter values, and where there may be a simple model structure among the classifications and their attributes. This approach is applicable when an interpretation contains categorical parameter values.

PLS-DA is an extension of partial least squares regression (PLSR). Partial least squares analyses seek to find latent variables that summarize the sample variability among the attribute values and that are highly predictive of the classification attributes. Latent variables are parameters in a mathematical model that are often not themselves observable or measurable in your experiment (also referred to as hidden variables). The latent variables may represent multiple physical or measurable characteristics of your sample that reduce the dimensionality of your analysis and therefore help present a less complex representation of the underlying relationships among your classifications.

Partial least squares regression is a statistical method that is similar to principal components analysis, but instead of finding hyperplanes of minimum variance between the independent and dependent variables, PLSR finds a linear regression model by projecting the predicted attribute variables and the observed attribute variables onto a new space (Figure 8 on page 25). Because both the X and Y data used for the regression analysis are projected to new spaces, the partial least squares family of methods are known as bilinear factor models. PLS-DA is a variant of PLSR where the Y values used in regression are categorical rather than continuous.

Partial least squares data analyses benefit from simultaneously performing dimension reduction and regression analyses. However, a drawback to partial least squares analyses is that they are drawn to the independent variables with the most variability. As a result outliers in the samples may have significant influence on the directions, resulting scores, and relationship with the response. Specifically, outliers can make it appear that such that:

- no relationship between the predictors and response when there truly is a relationship, or
- a relationship between the predictors and response when there truly is no relationship

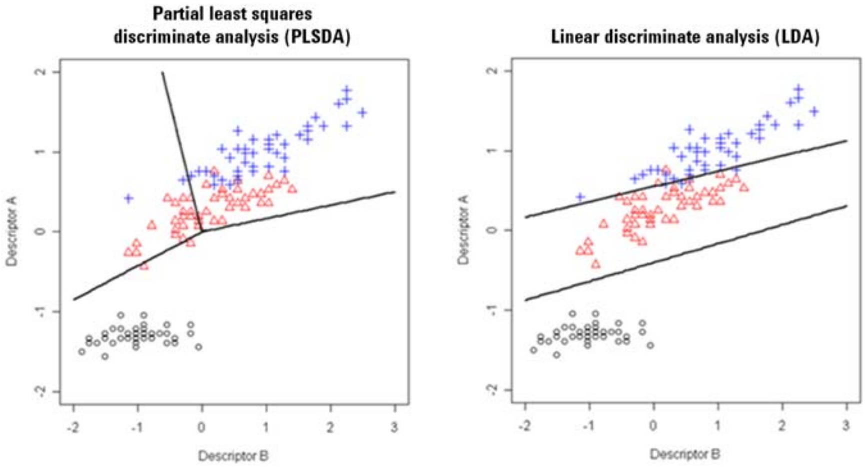


Figure 8 *Partial least squares discriminant analysis seeks to find linear combinations of the variables that are highly correlated, but each linear combination does not need to be separated linearly across the entire variable space as in the case of linear discriminate analysis. An entity in this analysis is referred to as a descriptor.*

Since PLS-DA is best for making classifications where all of the attribute values are measurable with little error and where there may be a simple model structure among the classifications and their attributes, PLS-DA is well suited for chromatography mass spectrometry. Chromatography mass spectrometry provides an accurate measurement of retention time, mass, and abundance. More-or-less traditional mathematical relationships can be applied to the data, and simple models can be assessed through a least squares regression fit of the data to find the model that provides the best fit.

Support Vector Machine (SVM)

SVM is best suited for making classifications where the attributes within the samples are intertwined and can benefit from transforming the input attribute space into additional dimensionality to identify classification separation planes. The algorithm looks for differentiation among your classifications in pairs and can work with smaller entity lists.

SVM attempts to separate samples representing two classes by imagining each sample as a point positioned in a two or three dimensional space and then calculating the parameters for a line or plane (linear or non-linear) that separates the samples into each classification. While several possible separating planes within a three-dimensional feature space may exist, the SVM algorithm finds the separator that maximizes the separation between the classes encompassing the sample points. The power of SVM stems from the fact that SVM supervised learning can effectively separate samples using non-linear functions and can therefore separate out samples containing intertwined feature sets. SVM therefore can efficiently classify samples using non-linear classifications by mapping the input into high-dimensional feature spaces.

Examples of SVM separation of samples representing two classifications are shown in Figure 9 and Figure 10 on page 26.

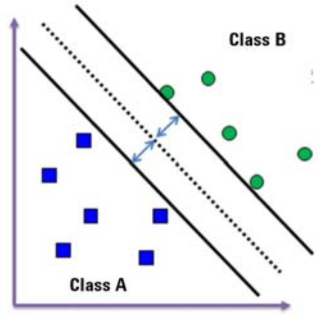


Figure 9 *Support vector machine represents the samples as points in space, divided by a clear gap that is as wide as possible. The gap is defined by a set of hyper-planes. New samples are predicted to belong to a category based on which side of the gap they fall.*

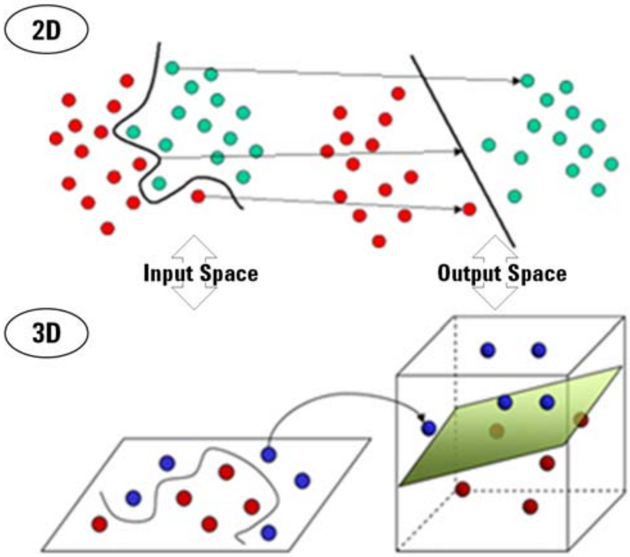


Figure 10 *Illustrations of support vector machine separating samples representing two classifications within a feature space of two and three dimensions.*

Naïve Bayesian (NB)

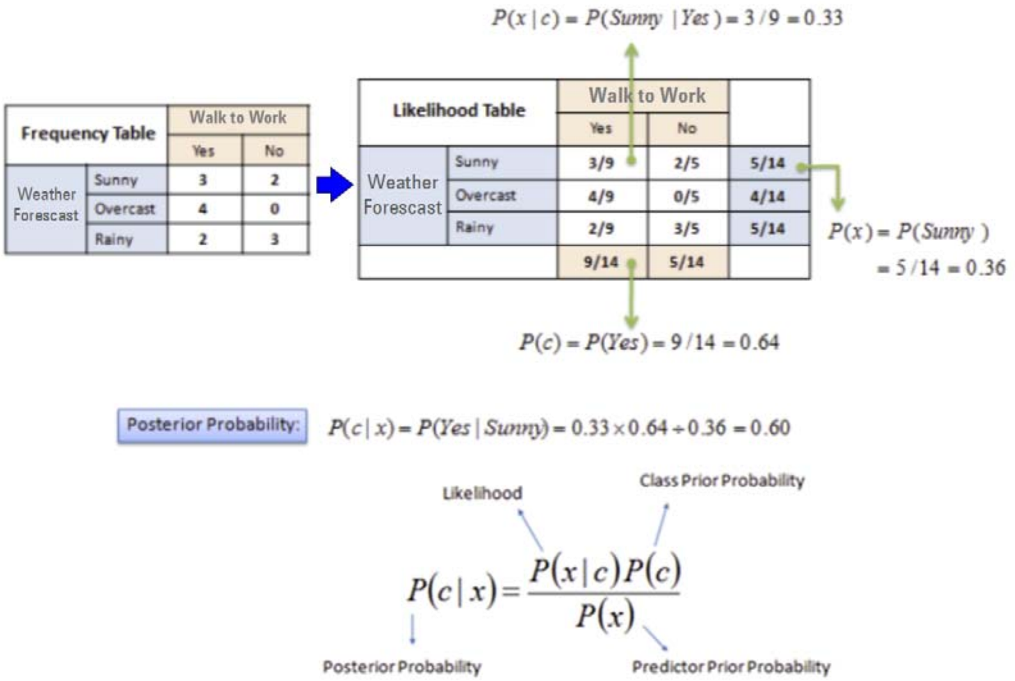
NB is best suited for making classifications where the attributes within a sample are independent from each other. The algorithm looks for differentiation among entities in your entity list by assuming that the change in the appearance of any one entity is unrelated to other entities in the entity list. This algorithm can work with smaller entity lists.

Bayesian classifiers are parameter based, probabilistic multi-class classifiers and can handle both continuous and categorical variables. Bayes' theorem leads to a statistically significant result (that is, an important result) through the mathematical manipulation of conditional probabilities. The statistical inference obtained by applying naïve Bayesian supervised learning is derived from the strong (naïve) assumption that the presence or absence of any particular feature is unrelated to the presence or absence of any other feature, given the class variable.

To predict the probability that a sample belongs to a certain class, the naïve Bayesian classifier assumes that the effect of an attribute of a given class is independent of the value of other attributes within the same class. For example, a fruit may be considered to be a peach if it is orange/red in color, round, has a fuzzy skin texture, and is about 3" in diameter. A naïve Bayes classifier considers that each of these features contribute independently to the probability that the fruit under study is a peach, regardless of the presence or absence of the other features. This assumption of feature independence is called the class conditional independence. The naïve Bayesian model is built based on the probability distribution function of the training data along each feature. The model is then used to classify a data point based on the learned probability density functions for each class.

The naïve Bayesian supervised learning approach provides an advantage in that it only requires a small amount of training sample data to estimate the necessary classification parameters (means and variances of the variables). Because naïve Bayesian assumes that the attributes do not have any dependencies on each other, only the variances of the attributes for each class needs to be determined, and not the entire covariance matrix

The Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). As mentioned before, the effect of the value of an attribute (x) on a given class (c) is independent of the values of other attributes. A simple example of converting the observations recorded in a frequency table to probabilities in a likelihood table is show in Figure 11.



$P(x|c) = P(Sunny \mid Yes) = 3/9 = 0.33$

Frequency Table		Walk to Work	
		Yes	No
Weather Forecast	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Likelihood Table		Walk to Work		
		Yes	No	
Weather Forecast	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$P(x) = P(Sunny) = 5/14 = 0.36$

$P(c) = P(Yes) = 9/14 = 0.64$

Posterior Probability: $P(c|x) = P(Yes \mid Sunny) = 0.33 \times 0.64 \div 0.36 = 0.60$

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Figure 11 *Naïve Bayesian calculation of the posterior probability of walking to work based on a class (c, yes or no) for a given attribute (x, weather forecast of sunny, overcast, or rainy).*

Decision Tree (DT)

DT is best suited for making classifications when the attributes are present in all or nearly all of the samples.

A Decision Tree is created through a linear progression of if-then-otherwise decisions based on the attribute values of the samples. Consider, for example, three samples belonging to classes A, B, and C and that the rows containing the feature and attribute values (class label) corresponding to these samples have values as shown in Table 1.

Table 1 *Example data set used to create the decision tree in Figure 12*

	Feature 1	Feature 2	Feature 3	Class Label
Sample 1	4	6	7	**A**
Sample 2	0	12	9	**B**
Sample 3	0	5	7	**C**

The following sequence of decisions classifies the samples.

> If Feature 1 is greater than or equal to 4, then the sample is part of Class A. Otherwise, if Feature 1 is less than 4, and
>
> > If Feature 2 is greater than or equal to 10, then the sample is part of Class B. Otherwise, if feature 2 is less than 10, then the sample is part of Class C.

This sequence of if-then-otherwise decisions can be arranged as a tree as shown in Figure 12. This tree is called a decision tree, and in the example classification only two of the three available features were necessary to separate the classes.

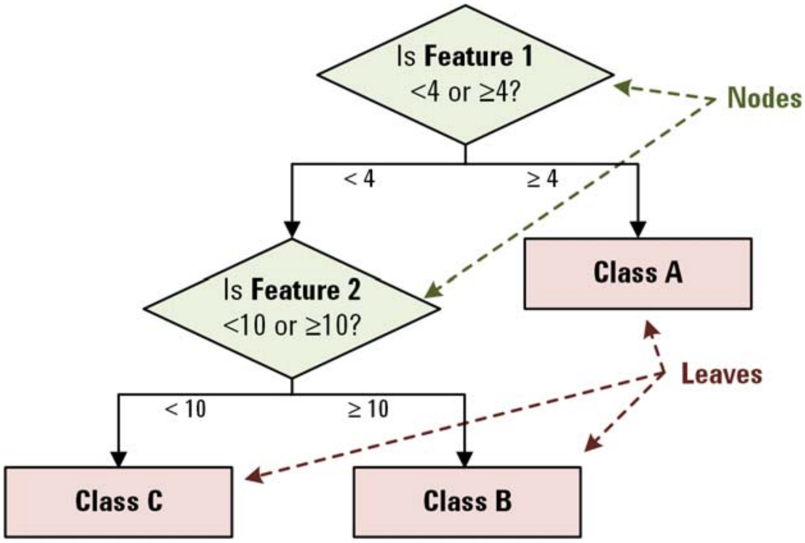


Figure 12 *Features of a Decision Tree: nodes are where a decision is made regarding the attributes of a feature, a branch is the path from one node to another node or one node to a leaf, and a leaf is a classification at the end of a branch. This decision tree is created from the data in Table 1.*

Class prediction models implement axis parallel decision trees as shown in Figure 12 on page 28. In an axis parallel tree, decisions at each node are made using one single feature of the many features present, e.g. a decision of the form "if Feature 2 is less than 10." A sample is classified by following the appropriate branches down the decision tree. All samples that follow the same branch, or path, down the decision tree are said to be at the same leaf. The tree building process continues until each leaf has a purity above a certain specified threshold. Purity is when all samples associated with this leaf, at least a certain fraction, come from one class. Once the tree building process is done, a pruning process is used to prune off portions of the tree to reduce chances of over-fitting.

Neural network (NN)

NN is best suited for making classifications when there likely exists a complex and unknown, or intermediate, layer of relationships that link that attributes within a sample to the classification. The intermediate layer of relationships behaves in a manner similar to how neurons create neural networks in a central nervous system - applying adaptive weights to the numerical parameters and non-linear approximations to the inputs. This algorithm can only be used with interpretations that contain numerical parameter values (parameter values that are not categorical) and requires a lot of computation time. NN is suitable when the data set involves many sample classifications.

A NN can handle multi-class problems, where more than two classes are represented in the data. The neural network implementation in MPP is the multi-layer perceptron trained using the back-propagation algorithm. It consists of layers of neurons. As illustrated in Figure 13 on page 30, the first layer is called the input layer, into which features for the samples to be classified are fed. The last layer is the output layer, which has an output node for each class in the dataset. Each neuron in an intermediate layer is interconnected with all the neurons in the adjacent layers. The strength of the interconnections between adjacent layers is given by a set of weights which are continuously modified during the training stage using an iterative process.

In simpler terms, there must exist in the samples a means to mathematically connect the attributes to classes. The means to make such connectivity among the features within the samples is referred to as the intermediate layer. The intermediate layer basically contains complicated mathematics.

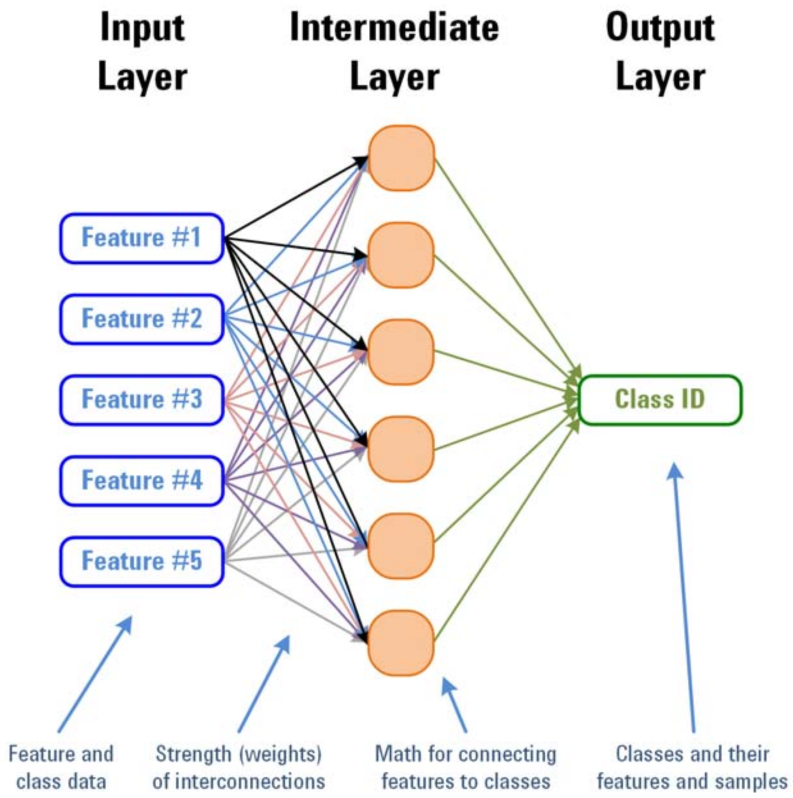


Figure 13 *Simple illustration of the layers of a neural network model.*

Review the steps to create a class prediction model

MPP guides you through a sequence of five steps to build your class prediction model. During the class prediction model development supervised learning additionally employs one of two methods to train and evaluate your selected class prediction algorithm. Each training/evaluation operation results in a confusion matrix. You review the confusion matrix (a table that shows the prediction accuracy of the model algorithm for each classification), decide whether to make further changes to the parameters and repeat the supervised learning, and finally select the prediction model that best meets your experimental requirements.

1. Enter input parameters.

Build your prediction model by first selecting an interpretation, an entity list, and a class prediction algorithm.

The advantages for using any one of the class prediction algorithms is often a matter of subjective opinion and your personal experience in applying the particular algorithm to your experimental parameters.

To build a class prediction model you should use a reliable, large entity list. A reliable entity list to use for class prediction is an entity list that has been filtered to contain entities that appear in no less than one half of the training sample data files, and preferably the entities in your input entity list appear in all of the training sample data files.

2. Enter validation parameters.

Select the class prediction algorithm parameters and select a supervised learning validation type.

Supervised learning is important to assess the ability of the prediction model to properly classify your samples, but it is also an important tool to avoid underfitting or overfitting your model on the training data as illustrated in Figure 14. A model that is improperly fit typically produces low accuracy with new data because the model fails to fit the underlying functions that classify the data.

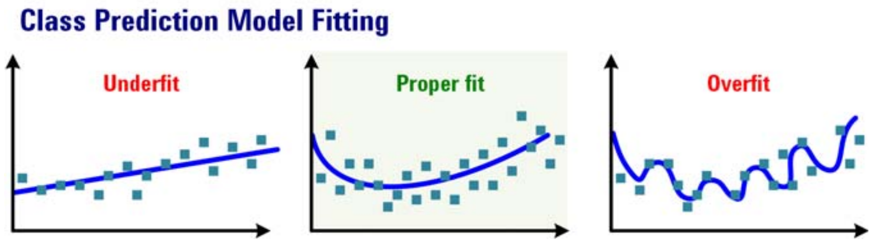


Figure 14 *Examples of model fitting in class prediction*

Supervised learning is performed using the sample data set that you use to train your model; your training sample data is divided into a uniform set of data whereby each set contains a sample representing each classification and each of these sets in turn is treated as a validation data set to a model that is generated using the remaining sample data sets. The results of the supervised learning are presented in a confusion matrix.

Supervised learning validation types

In MPP the prediction accuracy of your class prediction models are evaluated using one of two **validation types**. The validation types organize your training data files into smaller groups and then in turn use the smaller groups within the model development to train and validate the model many times to produce statistical prediction accuracy metrics. This validation approach is most successful when your experiment design includes many replicates for each parameter value. Each validation type presents the performance of the prediction model using a confusion matrix. The available validation type during your prediction model development may vary based on the selected prediction algorithm.

Leave One Out: The replicate samples in your training data set are arranged in columns representing each of the known classifications. All of the data, with the exception of one row containing one sample from each classification, is used to train the learning algorithm and generate a prediction model. The prediction model is then used to classify the remaining row of data as a temporary validation data set. The process is repeated for every row in the dataset and a confusion matrix is generated. During the Leave One Out validation each row of data is sequentially treated as a temporary validation data set for a corresponding model that is trained using all of the other rows of data. Leave One Out validation does not require any parameters necessary to enter or adjust.

N-Fold: The total number of samples associated with the training data set are randomly divided into N equal parts. The samples representing N-1 parts are used for training your prediction model and the remaining samples, representing the remaining one part, are used for validating the prediction model. The process is repeated N times, with a different part of the training data set being used to validate the model developed using the samples representing the N-1 parts. Using N-fold, like Leave One Out, each sample is used at least once to validate the prediction model. A confusion matrix is generated from the results. The default values of three-fold validation and one repeat are sufficient for most analyses. For results with greater confidence, you can use a ten-fold validation with three repeats. For large data sets, the latter may take significantly more computing time.

3. Review validation algorithm outputs.

Review the class prediction results and the confusion matrix generated from the internal validation. Remember that this validation is performed by segregating your training data sets into smaller groups as described in step 2 on page 31.

The prediction results report provides details of the prediction versus actual classification for each condition. The Confusion Matrix results show a cumulative Confusion Matrix, which is the sum of confusion matrices for individual runs of the learning algorithm and a summary of the expected efficacy of the prediction model. After you review the confusion matrix to assess the accuracy of the prediction model, decide whether to make further changes to the parameters and repeat the supervised learning.

4. Review training algorithm outputs.

The performance results include prediction results, a confusion matrix for the training model on the whole entity list, and a Lorenz curve showing the efficacy of classification and the prediction model. The Confusion Matrix results in this step show the result of applying the final prediction model to the training sample data. Wherever appropriate, a visual output of the classification model is presented.

5. Review the Class Prediction Model.

Review the summary information for your class prediction model and save the class prediction model.

This is the conclusion of building your prediction model. Your prediction model is now saved as an object in MPP and can be exported to use with MassHunter Sample Class Predictor to classify your new samples during data acquisition. Consider building more than one class prediction model using a different algorithm to develop experience and to identify the best model for your experiment.

Review recursive feature finding

During the class prediction workflow you recursively find the features in your original sample data files. Combined with collecting replicate samples in your experiment, recursive feature finding improves the statistical accuracy (confidence) of your analysis and reduces the potential for obtaining a false positive or a false negative answer to your hypothesis and sample classification.

Recursive feature finding (Find Compounds by Formula (FbF)) is performed to improve the statistical probability of finding specific features (targeted features) in your sample data files. In particular, you want to specifically find the most important features that add value to your analysis.

Recursive feature finding

During the workflow you recursively find the features in your sample data files after you perform a satisfactory differential analysis. From the total number of untargeted features originally found in your sample data files, the initial differential analysis identifies a subset of features that contribute to the best separation of your experimental conditions (classifications). By using the entity list containing these significant features to perform a targeted feature finding, you improve the statistical accuracy of your differential analysis and can improve the accuracy of your subsequent class prediction model development. Recursive finding of the significant features from your differential analysis is optional depending on the nature of the features; you are encouraged to review “Recursion within the workflow” to help you decide whether to perform recursion at this first opportunity.

You can also recursively find the features in your sample data files after you create a satisfactory class prediction model. By using the entity list that is now known to provide a satisfactory measure of confidence in predicting class membership, to perform a targeted feature finding, you improve the statistical accuracy of your entire class prediction analysis. Recursive finding of the prediction model features from your class prediction model is optional depending on the nature of the features; it is highly recommended to recursively find your features after building your class prediction model if you did not recursively find the features after your initial differential analysis.

Recursion within the workflow

For the best measure of confidence in your *differential analyses* and *class prediction* you find the features in your sample data files using recursion. The following steps are typical for differential analysis and class prediction and are presented to help you understand where recursion is performed.

a **Begin an MPP-based workflow.** MPP experiments typically begin with analyzing untargeted features (Find Compounds by Molecular Feature, MFE) that were subsequently binned together during the alignment in the MS Experiment Creation Wizard (Step 8 of 11). At this point the analysis emphasis is on features that have a strong (abundant) signal to be used in the next step of the analysis - to identify correlation of the variation of feature intensities with our groupings (classifications).

b **Perform an initial differential analysis.** The goal of the initial differential analysis is to find a correlation of the intensity variations among the untargeted features to the sample conditions (classifications). The large number of untargeted features found in the sample data files is reduced to a much smaller number of features that are now considered more important to further analyses.

c **Perform recursive feature finding.** Recursive feature finding improves the confidence measure of your analysis. To perform recursive feature finding you export the features from the initial differential analysis as a targeted list of features. This step improves the quality of finding the features in the original sample files; targeted feature finding focuses on finding a specific set of features with less emphasis on feature strength.

Recursive feature finding involves:

- Export the significant features from your differential analysis.
- Recursively find the significant features from your differential analysis in your sample data files.
- Recreate your experiment and input sample filters.

Performing recursion is always recommended, and is especially recommended when your class prediction model depends on features that show up and down regulation among your classification or the presence and absence of features.

- If your prediction model relies on the regulation of strong features, enhanced finding of weak targeted features may not significantly help you during your initial model creation and recursive feature finding may be optional before creating your initial class prediction model.
- If your prediction model relies on feature presence and absence then it is recommended you perform recursive feature finding before creating your initial class prediction model.
- If your hypothesis does not include a priori knowledge on the regulation or absence of features in your samples then it is recommended you perform recursive feature finding before creating your initial class prediction model.

d **Recreate your initial class prediction model.** During this step you gain additional confidence in your analysis through the use of targeted features that were identified as significant during your initial differential analysis.

e Build your class prediction model.

f ***(Optional)*** **Perform recursive feature finding.** If you did not perform a recursive feature finding after your initial differential analysis, export the prediction model features from your class prediction model for recursive feature finding. Class prediction confidence benefits from targeted feature finding to find the features in your sample files.

- Export the prediction model features from your class prediction model.
- Recursively find the features used by your class prediction model in your sample data files.
- Recreate your experiment, feature filters, differential analysis, and class prediction model.

g Apply your class prediction model.

Recursion using Qualitative Analysis and Profinder

After you have identified the best class prediction algorithm and created a satisfactory class prediction model, you can further improve your class prediction model by recursively finding the model entities in your sample data files. To find your features using recursion, export the entity list that the class prediction model uses and use this entity list as a target list of features to find the features in your existing and new samples using Qualitative Analysis or Profinder.

In **MassHunter Qualitative Analysis** a targeted feature extraction is performed by using Find Compounds by Formula (FbF). For more information and the steps involved in finding features recursively using Find Compounds by Formula within Qualitative Analysis, see the chapter “Recursive find features” in the *Agilent Metabolomics Workflow - Discovery Workflow Guide*.

You can use Batch Targeted Feature Extraction in **MassHunter Profinder** to recursively find features in TOF and Q-TOF data files. For a brief overview of finding features recursively using Batch Recursive Feature Extraction within Profinder, see the chapter “Feature Extraction Workflow Algorithms” in the *Agilent G3835AA MassHunter Profinder Software - Quick Start Guide*.

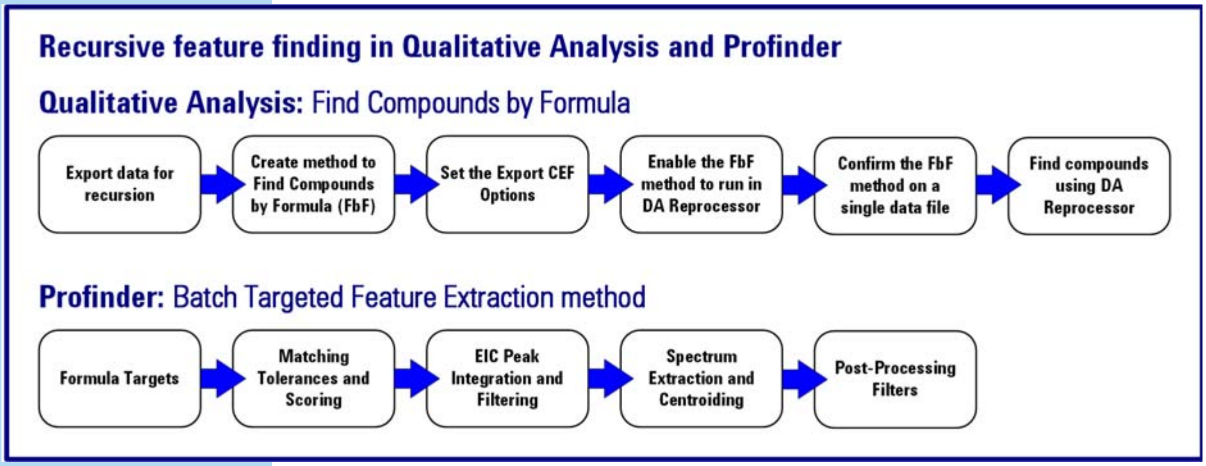

Recursive feature finding in Qualitative Analysis and Profinder
Qualitative Analysis: Find Compounds by Formula
Export data for recursion
Create method to Find Compounds by Formula (FbF)
Set the Export CEF Options
Enable the FbF method to run in DA Reprocessor
Confirm the FbF method on a single data file
Find compounds using DA Reprocessor
Profinder: Batch Targeted Feature Extraction method
Formula Targets
Matching Tolerances and Scoring
EIC Peak Integration and Filtering
Spectrum Extraction and Centroiding
Post-Processing Filters

Figure 15 *Steps to finding targeted features for recursion using Qualitative Analysis and Profinder*

Apply your class prediction model

After you have created and saved your class prediction model, you can run your prediction model on your validation data set and new sample data. When you run your prediction model on your validation data set (samples that were set aside during your model creation that have known class membership but were not part of the training data set), you are performing a final quality inspection of your class prediction model. Validation and new sample prediction are the last two parts of class prediction as shown in Figure 16.

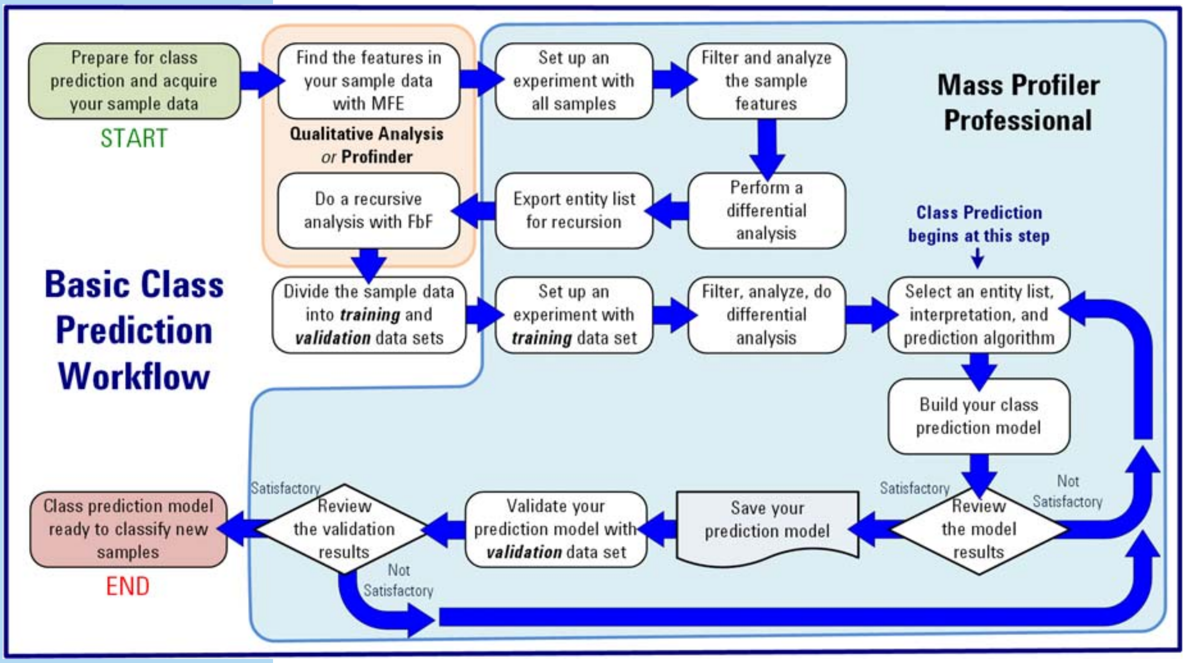

Prepare for class prediction and acquire your sample data
START
Find the features in your sample data with MFE
Qualitative Analysis or Profinder
Do a recursive analysis with FbF
Set up an experiment with all samples
Filter and analyze the sample features
Mass Profiler Professional
Export entity list for recursion
Perform a differential analysis
Class Prediction begins at this step
Basic Class Prediction Workflow
Divide the sample data into training and validation data sets
Set up an experiment with training data set
Filter, analyze, do differential analysis
Select an entity list, interpretation, and prediction algorithm
Build your class prediction model
Satisfactory
Not Satisfactory
Class prediction model ready to classify new samples
END
Satisfactory
Review the validation results
Not Satisfactory
Validate your prediction model with validation data set
Save your prediction model
Review the model results

Figure 16 The basic steps in building your first class prediction model

Final quality inspection

Run your prediction model on your validation data set.

The final quality inspection of your prediction model is made by running the class prediction model on your *validation sample data files* (the sample data files that were not used to build the prediction model).

The final quality inspection can also be performed again after recursively finding the features in your training and validation sample data sets. After you recursively find the features in your data sets, you can rebuild your class prediction model and compare the recursive prediction model's overall accuracy and validation results to the model results obtained using the original, untargeted feature extraction (Find Compounds by Molecular Feature).

Classify new samples

Run your prediction model on new sample data.

Apply your class prediction model

You can manually run your prediction model on feature data files (CEF files) using Mass Profiler Professional or Sample Class Predictor. With Sample Class Predictor you can additionally automate class prediction during data acquisition. During acquisition, your class prediction model is a targeted analysis that is applied to your sample data files.

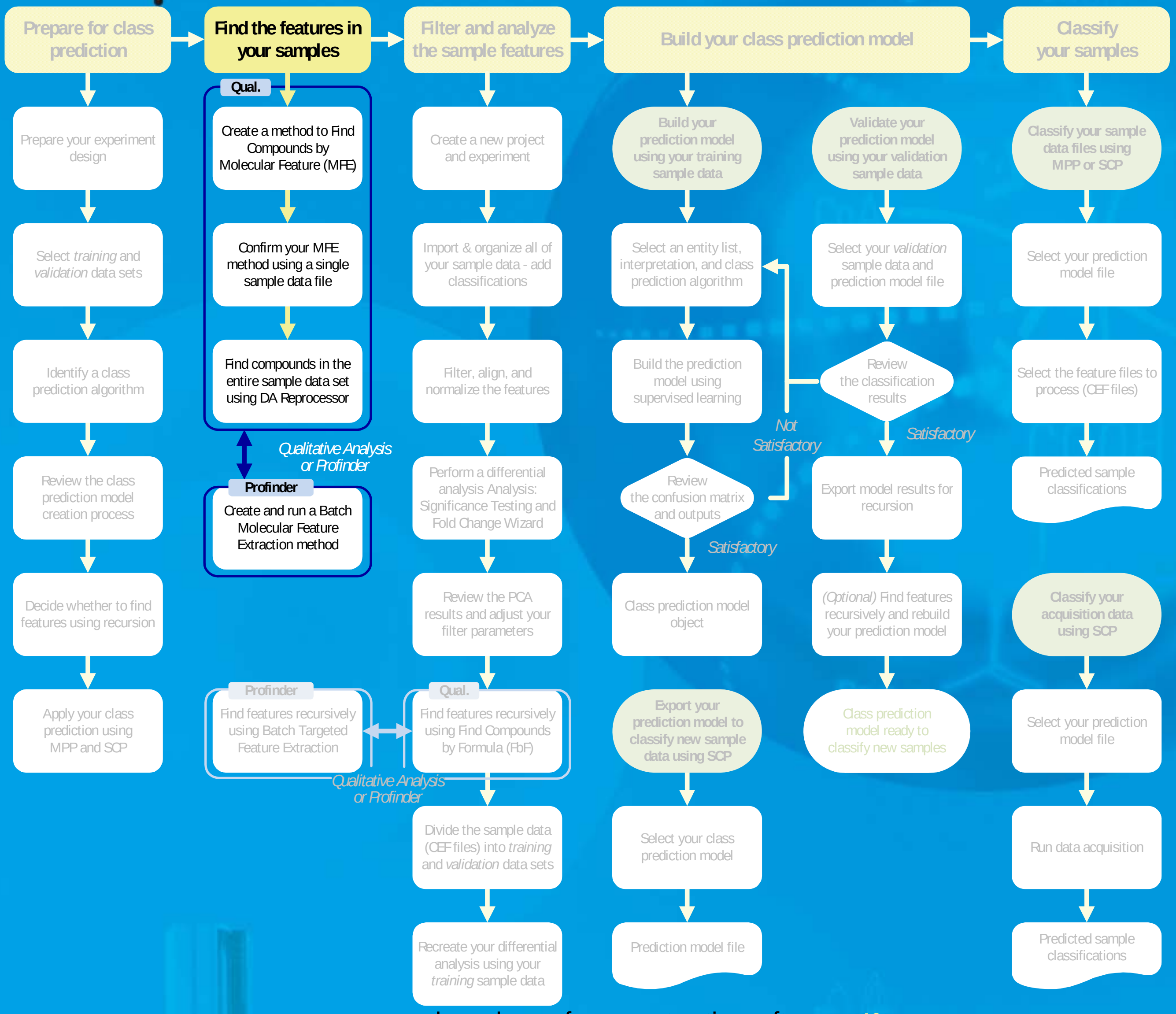
You can use Sample Class Predictor as a standalone tool to classify your sample feature files, similar to **Run Prediction from File** that is available within MPP, and as an automated classification tool that interfaces directly with ChemStation and MassHunter acquisition applications.

Sample Class Predictor can run models on data acquired and processed in Agilent MassHunter or Agilent OpenLAB ChemStation Edition. The supported data file types are AMDIS, ChemStation, MassHunter, GC Scan and Generic. For Prediction on AMDIS data sources, both the FIN and corresponding ELU file should be present in the same working folder.

Samples must be from the same data source as those the model was built on, but they do not need to be from the same technology. When a prediction is run from a data file during acquisition, the targeted feature extraction uses the alignment values included in the file to perform alignments on entities in the new data.

Find the features in your samples

Molecular features are extracted from your sample data files based on mass spectral and chromatographic characteristics. The resulting CEF files are used to build your class prediction model.



Prepare for class prediction

Prepare your experiment design

Select *training* and *validation* data sets

Identify a class prediction algorithm

Review the class prediction model creation process

Decide whether to find features using recursion

Apply your class prediction using MPP and SCP

Find the features in your samples

Qual.

Create a method to Find Compounds by Molecular Feature (MFE)

Confirm your MFE method using a single sample data file

Find compounds in the entire sample data set using DA Reprocessor

Qualitative Analysis or Profinder

Profinder

Create and run a Batch Molecular Feature Extraction method

Profinder

Find features recursively using Batch Targeted Feature Extraction

Qual.

Find features recursively using Find Compounds by Formula (FbF)

Qualitative Analysis or Profinder

Filter and analyze the sample features

Create a new project and experiment

Import & organize all of your sample data - add classifications

Filter, align, and normalize the features

Perform a differential analysis Analysis: Significance Testing and Fold Change Wizard

Review the PCA results and adjust your filter parameters

Divide the sample data (CEF files) into *training* and *validation* data sets

Recreate your differential analysis using your *training* sample data

Build your class prediction model

Build your prediction model using your training sample data

Select an entity list, interpretation, and class prediction algorithm

Build the prediction model using supervised learning

Review the confusion matrix and outputs

Not Satisfactory

Satisfactory

Class prediction model object

Export your prediction model to classify new sample data using SCP

Select your class prediction model

Prediction model file

Validate your prediction model using your validation sample data

Select your *validation* sample data and prediction model file

Review the classification results

Satisfactory

Export model results for recursion

(Optional) Find features recursively and rebuild your prediction model

Class prediction model ready to classify new samples

Classify your samples

Classify your sample data files using MPP or SCP

Select your prediction model file

Select the feature files to process (CEF files)

Predicted sample classifications

Classify your acquisition data using SCP

Select your prediction model file

Run data acquisition

Predicted sample classifications

Agilent Technologies

Launch your feature extraction software

You may use one of two processes to find the features in all of your sample data files: Qualitative Analysis and Profinder (see Figure 17). Both programs export the sample features using a compound exchange format (CEF) file. A single CEF file is generated for each sample data file.

Feature finding using Qualitative Analysis

MassHunter Qualitative Analysis, in conjunction with the batch file processor utility called DA Reprocessor, can be used to find the features in any Agilent mass spectrometry-based sample data file. Qualitative Analysis is used to create your method to find features and can be used to process individual sample data files. **DA Reprocessor**, an efficient batch processor, applies your find feature method to multiple sample data files and is used to extract the features from all of your sample data files. Feature finding using Qualitative Analysis and DA Reprocessor begins at "MassHunter Qualitative Analysis" on page 41.

Feature finding using Profinder

MassHunter Profinder is a stand-alone program that is optimized for batch feature extraction from TOF and Q-TOF based profiling data files. Profinder can be used to process any number of data files. Feature finding using Profinder begins at "MassHunter Profinder" on page 61.

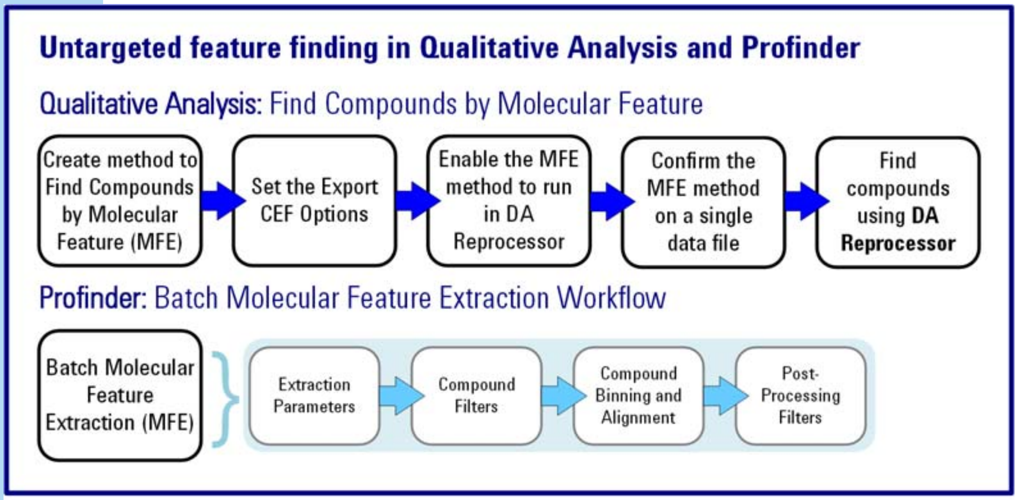

Untargeted feature finding in Qualitative Analysis and Profinder
Qualitative Analysis: Find Compounds by Molecular Feature
Create method to Find Compounds by Molecular Feature (MFE)
Set the Export CEF Options
Enable the MFE method to run in DA Reprocessor
Confirm the MFE method on a single data file
Find compounds using DA Reprocessor
Profinder: Batch Molecular Feature Extraction Workflow
Batch Molecular Feature Extraction (MFE)
Extraction Parameters
Compound Filters
Compound Binning and Alignment
Post-Processing Filters

Figure 17 *Comparison of the process to find untargeted features using Qualitative Analysis and Profinder*

Note: Qualitative Analysis and Profinder can each find the features in your raw sample data files. However, due to differences in how each program interacts with the data files, it is recommended to use one program or the other to review and extract the features from your sample data files. *If you plan to use both Qualitative Analysis and Profinder with your TOF and Q-TOF data files, use Qualitative Analysis before you use Profinder.* You can use Profinder to reprocess raw data files after an analysis with Qualitative Analysis, but you cannot use Qualitative Analysis after using Profinder.

Compounds, referred to as molecular features, are extracted from your data based on mass spectral and chromatographic characteristics. The find feature process is referred to as Molecular Feature Extraction (MFE). Molecular feature extraction quickly and automatically generates a complete, accurate list of your compounds using chromatographic alignment across multiple data files. The extracted features include molecular weight, retention time, *m/z*, and abundance. Both Qualitative Analysis and Profinder minimize the appearance of false positive and false negative features by binning the features in the chromatographic time domain.

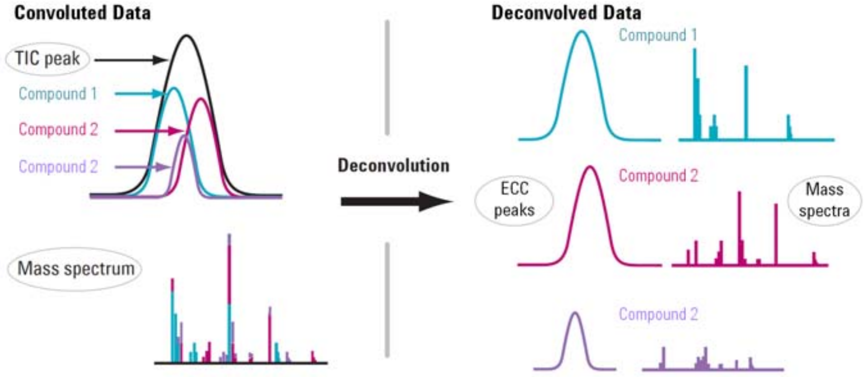


Figure 18 *Deconvolution using MassHunter Qualitative Analysis and Profinder finds compounds that are chromatographically unresolved or poorly resolved. Deconvolution generates an extracted ion chromatogram and a reconstructed, single component spectrum for each compound.*

Molecular feature extraction involves chromatographic deconvolution as illustrated in Figure 18 on page 41. Molecular feature extraction automatically finds related coeluting ions, sums the related ion signals into single values, creates compound spectra, and reports results as a molecular feature, or compound.

After you have extracted the features in all of your sample data files, the next step is to create an initial differential analysis using the sample data. Creating an initial differential analysis provides you with confidence that the sample data can be separated into your classifications.

MassHunter Qualitative Analysis

The following examples use MassHunter Qualitative Analysis B.06.00 running on 64-bit Windows 7 Professional. If you have TOF and/or Q-TOF data you can alternatively find your features using Profinder as described beginning at "MassHunter Profinder" on page 61.

User Interface Note: When you make a change to a parameter in MassHunter Qualitative Analysis, the software automatically places a change icon ▲ (a blue triangle shape) in the Method Editor tab and next to the parameter containing the changed parameter. This icon indicates that you have unsaved changes in your method and helps you remember to save the changes you have made to the method. The original parameter value may be viewed by placing your pointer over the change icon. When you save your method, the change icons disappears.

1. Start MassHunter Qualitative Analysis software.

MassHunter Qualitative Analysis is a software tool used to perform the function of finding molecular features in any sample data file. After the molecular features are found they are imported into Mass Profiler Professional for statistical analysis. Feature finding is an essential prerequisite to using Mass Profiler Professional.

Note: Online help is available within MassHunter Qualitative Analysis by pressing the **F1** key on the keyboard. The information presented specifically relates to the software parameters and options available in the active display.

a Double-click the Qualitative Analysis icon located on the desktop,

or (for Qualitative Analysis version B.06.00 on Windows 7)

Click **Start > All Programs > Agilent > MassHunter Workstation > Qualitative Analysis B.06.00**.

b Click **Cancel** in the **Open Data File** dialog box to start MassHunter Qualitative Analysis without opening any data files. To open data files later, click **File > Open Data File.**

You do not need to open a data file at this time. You are prompted to open a data file in "Confirm the MFE method on a single data file" on page 56.

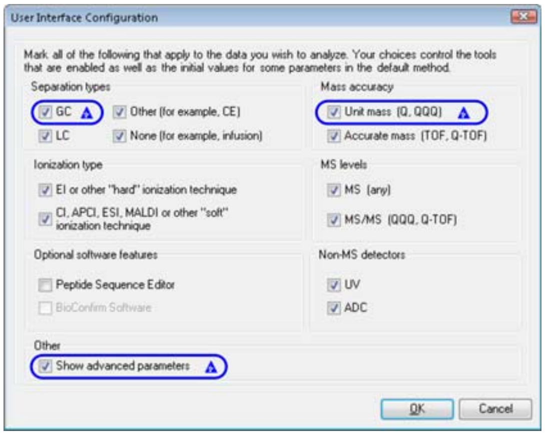
2. Enable advanced parameters in the user interface.

Advanced parameters must be enabled in MassHunter Qualitative Analysis in order to show tabs labeled Advanced in the Method Editor and to enable compound importing for recursive finding of molecular features.

a Click **Configuration > User Interface Configuration**.

b Mark the **Show advanced parameters** check box under the Other group heading. See Figure 19 on page 42.

If the files intended to be processed include GC/MS data, mark the **GC** check box under the Separation types group heading. If your analyzer is a quadrupole, mark the **Unit Mass (Q, QQQ)** check box under the Mass accuracy group heading.



User Interface Configuration

Mark all of the following that apply to the data you wish to analyze. Your choices control the tools that are enabled as well as the initial values for some parameters in the default method.

Separation types
☑ GC
☑ Other (for example, CE)
☑ LC
☑ None (for example, infusion)

Mass accuracy
☑ Unit mass (Q, QQQ)
☑ Accurate mass (TOF, Q-TOF)

Ionization type
☑ EI or other "hard" ionization technique
☑ CI, APCI, ESI, MALDI or other "soft" ionization technique

MS levels
☑ MS (any)
☑ MS/MS (QQQ, Q-TOF)

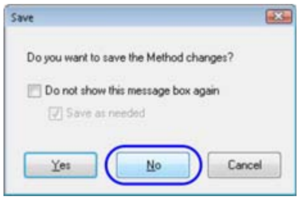
Optional software features
□ Peptide Sequence Editor
□ BioConfirm Software

Non-MS detectors
☑ UV
☑ ADC

Other
☑ Show advanced parameters

Figure 19 User Interface Configuration dialog box

c Click **OK** in the **User Interface Configuration** dialog box.


Do you want to save the Method changes?

☐ Do not show this message box again

☑ Save as needed

Figure 20 *Save dialog box*

d Click **No** in the **Save** dialog box.

At this time you are only making a change to the Qualitative Analysis user interface, not to your method. If you click **Yes** or **Cancel**, the changes to the user interface configuration are not invoked; repeat the steps to change the user interface beginning at "Click Configuration > User Interface Configuration." on page 42.

e Check to make sure that **File > Import Compound** is an available command. See Figure 21. This command is necessary to review CEF files before importing them into Mass Profiler Professional.

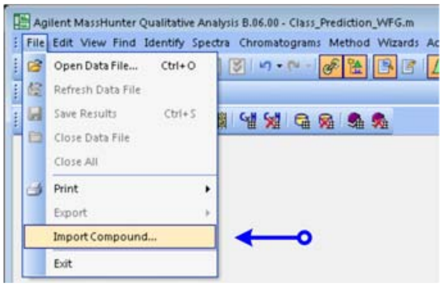


Figure 21 *Confirmation that your changes to the user interface configuration are successful is the appearance of the* ***Import Compound*** *command in the* ***File*** *menu.*

Create a method to Find Compounds by Molecular Feature

Find Compounds by Molecular Feature is commonly referred to as molecular feature extraction (MFE). Molecular feature extraction finds untargeted features (compounds) in you sample data files by using chromatographic deconvolution as illustrated in Figure 18 on page 41. The molecular feature extraction method automatically finds related co-eluting ions, sums the related ion signals into single abundance values, creates compound spectra, and reports results as a molecular feature. An untargeted molecular feature is subsequently uniquely identified by retention time, neutral mass, volume, and composite spectrum. An example of the relationships between some of the molecular features and the TIC is shown in Figure 22.

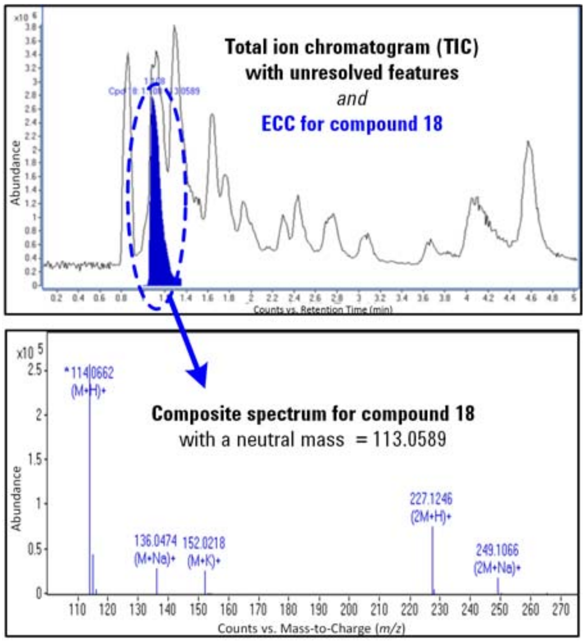


Figure 22 *Example of the relationship of some spectral components to the ECC*

Co-eluting ions: Molecular feature extraction finds co-eluting ions related to the same compound and creates an extracted compound chromatogram (ECC) including isotopes (13C, 15N, 2H, 18O), adducts (most commonly H+, Na+, K+ for positive ions and H- for negative ions), and dimers such as (2M + H)+.

Compound spectra: After creating extracted compound chromatograms, molecular feature extraction generates individual compound spectra for each molecular feature based on the co-eluting ions present.

Volume: The area of the ECC. The ECC is formed from the sum of the individual ion abundances within the compound spectrum at each retention time in the specified time window. The compound volume generated by molecular feature extraction is used by Mass Profiler Professional to make quantitative comparisons.

Composite spectrum: A compound spectrum that contains more than one co-eluting ion, more than just the (M+H) ion, within the molecular feature and is used by Mass Profiler Professional for recursive analysis and by ID Browser for compound identification.

Results: Each molecular feature, or compound, is uniquely identified by retention time, neutral mass, volume, and composite spectrum. An example of the relationships between some of the molecular features and the TIC is shown in Figure 22 on page 44.“Enter parameters for your Find Compounds by Molecular Feature method”

Finding the untargeted features in your sample data files using Qualitative Analysis involves five sequential steps as shown in Figure 23:

- “Enter parameters for your Find Compounds by Molecular Feature method”
- “Set the Export CEF Options” on page 53
- “Enable the method to run in MassHunter DA Reprocessor” on page 54
- “Confirm the MFE method on a single data file” on page 56
- “Find compounds using DA Reprocessor” on page 59

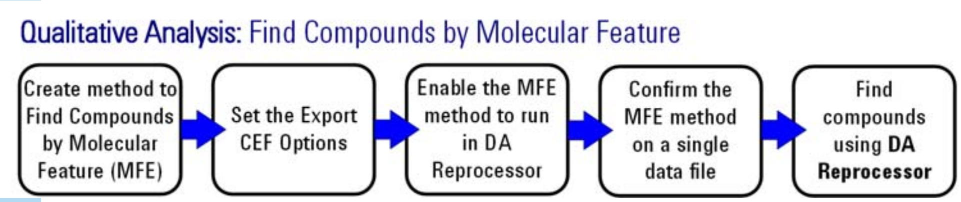

Qualitative Analysis: Find Compounds by Molecular Feature
Create method to Find Compounds by Molecular Feature (MFE)
Set the Export CEF Options
Enable the MFE method to run in DA Reprocessor
Confirm the MFE method on a single data file
Find compounds using DA Reprocessor

Figure 23 *Steps to find untargeted features using Qualitative Analysis*

Enter parameters for your Find Compounds by Molecular Feature method

This section provides a set of parameters that are applicable for finding small organic molecules such as metabolites in your sample data. The suggested parameters work well for the example sample data files in this workflow guide. You may find that slightly different settings work better for your sample data files.

1. Open the Method Editor window for finding compounds by molecular feature.

Open the **Method Editor: Find Compounds by Molecular Feature** from the **Method Explorer** window.

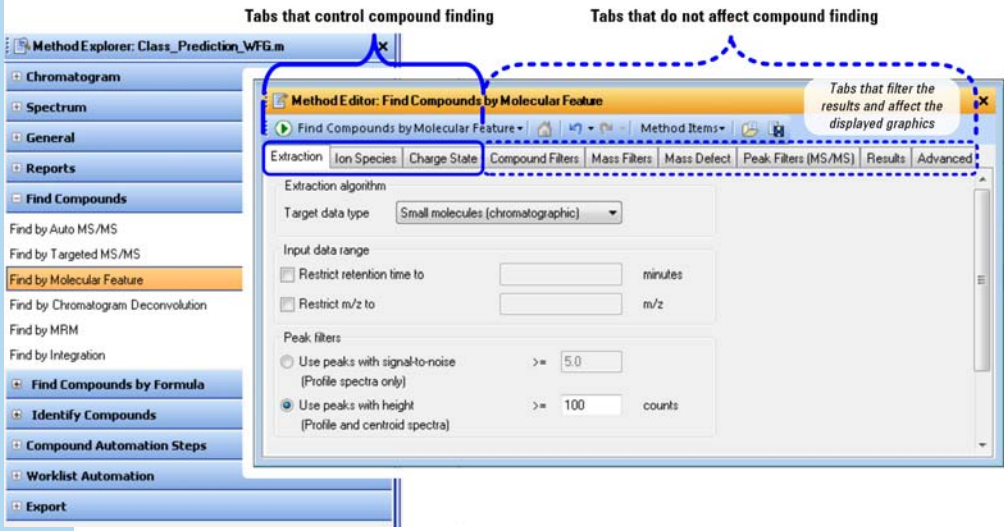
1. Click **Find Compounds** from within the **Method Explorer** window.
2. Click **Find by Molecular Feature**.

Notes regarding finding compounds by molecular feature:

- All of the parameters involved in molecular feature extraction are accessed from the **Method Editor: Find Compounds by Molecular Feature** tabs.
- Use the Extraction, Ion Species, and Charge State tabs to enter parameters that control compound finding. Use the remaining tabs to enter parameters to filter the results and display the graphics.
- Click **Find > Find Compounds by Molecular Feature** or click the Find Compounds by Molecular Feature button Find Compounds by Molecular Feature.
- MFE progress is shown in an **Operation in Progress** status box.
- For more information, see the *Agilent MassHunter Workstation Software Qualitative Analysis - Familiarization Guide* (G3335-90156, Revision A, April 2013).

2. Enter parameters for the tabs that control compound finding.

The parameters you enter in the **Extraction**, **Ion Species**, and **Charge State** tabs affect your untargeted feature finding. After the first time you run molecular feature extraction, any subsequent parameter changes you make within these tabs require you to reprocess your sample data files as described in "Confirm the MFE method on a single data file" on page 56 and "Find compounds using DA Reprocessor" on page 59. Reprocessing takes more time because the features must be re-extracted.


Tabs that control compound finding
Tabs that do not affect compound finding
Tabs that filter the results and affect the displayed graphics
Method Explorer: Class_Prediction_WFG.m
Chromatogram
Spectrum
General
Reports
Find Compounds
Find by Auto MS/MS
Find by Targeted MS/MS
Find by Molecular Feature
Find by Chromatogram Deconvolution
Find by MRM
Find by Integration
Find Compounds by Formula
Identify Compounds
Compound Automation Steps
Worklist Automation
Export
Method Editor: Find Compounds by Molecular Feature
Find Compounds by Molecular Feature
Method Items
Extraction
Ion Species
Charge State
Compound Filters
Mass Filters
Mass Defect
Peak Filters (MS/MS)
Results
Advanced

Extraction algorithm
Target data type: Small molecules (chromatographic)

Input data range
□ Restrict retention time to ______ minutes
□ Restrict m/z to ______ m/z

Peak filters
○ Use peaks with signal-to-noise (Profile spectra only) >= 5.0
● Use peaks with height (Profile and centroid spectra) >= 100 counts

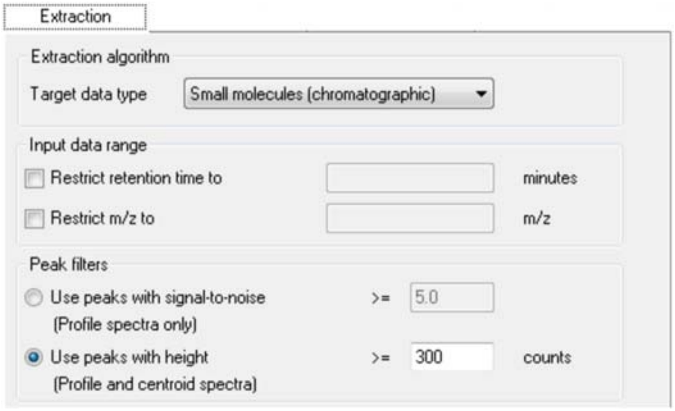
Figure 24 *Overview of the Method Editor tabs associated with Find Compounds by Molecular Feature (MFE)*

Extraction tab

a Edit the parameters on the Extraction tab.

The parameters in this tab let you specify features of the source data that enable the molecular feature extraction algorithm to perform more efficiently.

1. Click the **Extraction** tab.



Extraction algorithm
Target data type: Small molecules (chromatographic)

Input data range
□ Restrict retention time to ______ minutes
□ Restrict m/z to ______ m/z

Peak filters
○ Use peaks with signal-to-noise (Profile spectra only) >= 5.0
● Use peaks with height (Profile and centroid spectra) >= 300 counts

Figure 25 *Parameter values for the MFE Extraction tab*

2. Select **Small molecules (chromatographic)** in the **Target data type** box for working with metabolomic data.

   Note: The data must be collected in profile mode for the **Small molecules (infusion)** target data type. For **Large molecules (proteins, oligos)** the data must be collected in centroid mode or both modes.

Molecular feature extraction starts the data reduction process by creating a copy of the data file using the centroid of all of the ions. If you collect data in profile mode, it saves processing time if you also collect the data with centroid mode turned on. For all other data collection methods, it is recommended to save the data with centroid data.

3. Clear all of the check boxes under the Input data range group heading.

   Note: Marking the options under the Input data range group heading is not necessary. Using **Restrict retention time to** and **Restrict *m/z* to** limits the location where the molecular feature extraction searches for features. The recommended process is to let the molecular feature extraction algorithm find all of the features and then to use the filter parameters available in the "tabs that do not affect compound finding" (Figure 24 on page 46) to remove unwanted features.

4. Click **Use peaks with height** and type `300` for the counts. The counts value you enter represents a signal level at and above which actual ion signals are observed. 300 is typical if the background noise is approximately 100 counts.

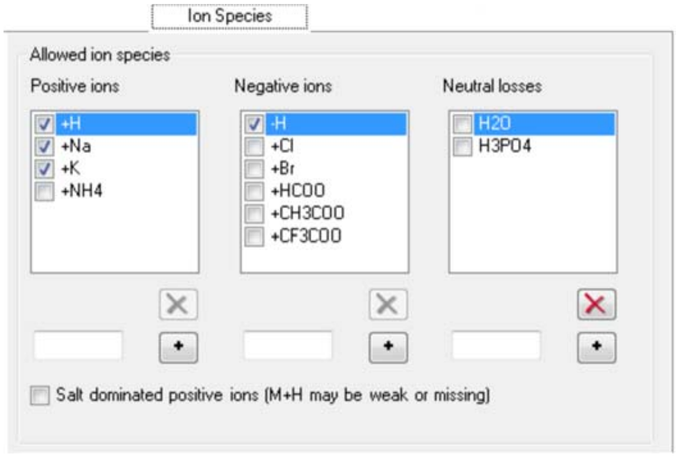
   Note: The target **Use peaks with height** counts is three times (3x) the electronic noise in the mass spectrum, the signal measured by the detector that is not due to actual ions. The electronic noise is found by viewing the background signal level at the higher *m/z* range (around 1,000 *m/z*) of a single mass spectrum in the data set. Do not use an averaged or background subtracted mass spectrum. If the **Use peaks with height** value is set to a value too small, Find Compounds by Molecular Feature takes a very long time to run and finds features that are very small. If the **Use peaks with height** value is set to a value too large, then actual features may not be found.

Ion Species tab

b Edit the parameters on the Ion Species tab.

The parameters in this tab let you specify the ion adducts that the MFE algorithm considers during the process of identifying molecular features.

1. Click the **Ion Species** tab.



Ion Species

Allowed ion species

Positive ions
☑ +H
☑ +Na
☑ +K
☐ +NH4

Negative ions
☑ -H
☐ +Cl
☐ +Br
☐ +HCOO
☐ +CH3COO
☐ +CF3COO

Neutral losses
☐ H2O
☐ H3PO4

☐ Salt dominated positive ions (M+H may be weak or missing)

Figure 26 *Parameter values for the MFE Ion Species tab*

2. Mark **+H**, **+Na**, and **+K** for positive ions and **-H** for negative ions.
3. Mark common **Neutral losses** if your mass spectrometer system is very energetic and thereby induces known neutral losses from the molecular ion.

   Note: Removal of possible ion adducts from the Allowed ion species group heading increases the molecular feature extraction efficiency. Glass bottles leach sodium into the liquid introduction system. Thus, it is recommended to change the solvent and sample delivery bottles to bottles made from PTFE in place of bottles made from glass to reduce the background sodium levels.

   Note: Molecular feature extraction requires that the molecular feature involves at least the addition or loss of a proton unless or salt adduct.

4. Mark the **Salt dominated positive ions (M+H may be weak or missing)** check box to direct the molecular feature extraction algorithm to reduce the emphasis on identifying protonated ions (M+H) because they may be weak or missing in your data. For example, mark this check box when detecting sugars that are sodium adducted.

   Clear this check box to direct the molecular feature extraction algorithm to place a uniform weight across the allowed ion species in calculating the molecular weight for each feature.

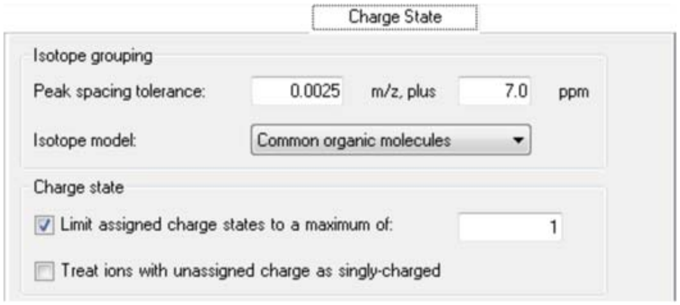
   Note: If your sample contains an ion species that is not an available option in your method, add the ion species in the appropriate charge or neutral column.

Charge State tab

c Edit the parameters on the Charge State tab.

The parameters on this tab let you set limits on the allowable ion charge states, and let you control how isotopes are identified and assigned to groups associated with each feature.

1. Click the **Charge State** tab.



Charge State

Isotope grouping

Peak spacing tolerance: 0.0025 m/z, plus 7.0 ppm

Isotope model: Common organic molecules

Charge state

☑ Limit assigned charge states to a maximum of: 1

□ Treat ions with unassigned charge as singly-charged

Figure 27 *Parameter values for the MFE Charge State tab*

2. Type `0.0025` for *m/z* and `7.0` for ppm into **Peak spacing tolerance**. These values are the tolerance that the molecular feature extraction algorithm uses to find isotope ions associated with each feature.
3. Select **Common organic molecules** for the **Isotope model**. For most metabolomics analyses, the **Common organic molecules** isotope model properly groups ions into the appropriate isotope clusters. Proper isotope clustering leads to an accurate assignment of charge state and mass for the molecular feature.

   Select **Unbiased** if metal containing molecules are expected.

Note: Selecting **Unbiased** slows the molecular feature extraction calculations considerably because all isotope models are considered.

4. Mark the **Limit assigned charge states to a maximum of** check box and type `1`. You only type `2` if you have a very specific reason; otherwise, `1` is used for metabolomics analyses. Increasing the value increases the chance of obtaining an unwanted isotope grouping.
5. Clear the **Treat ions with unassigned charge as singly-charged** check box. When this check box is cleared, ions that cannot be assigned a charge state by the molecular feature extraction algorithm are ignored.

3. Enter parameters for the tabs that filter results or affect the displayed graphics.

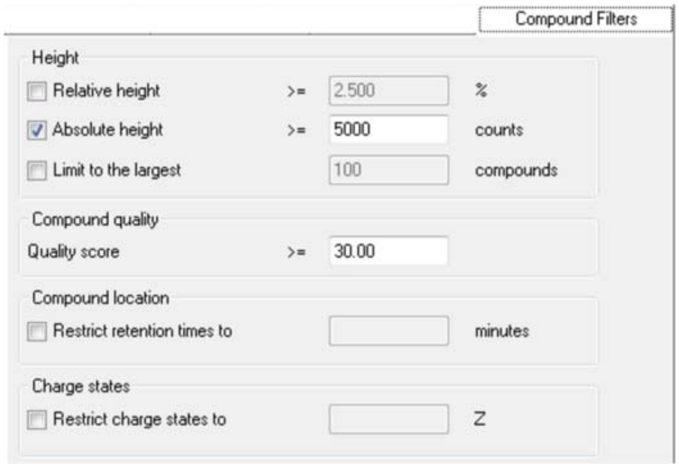
After the first time you run molecular feature extraction, any subsequent parameter changes you make within these tabs require you to reprocess your sample data files as described in "Confirm the MFE method on a single data file" on page 56 and "Find compounds using DA Reprocessor" on page 59. However, changes you make within these tabs reprocess the data much more quickly because the find features algorithm is not repeated. The improved speed for reprocessing the molecular features helps you review several combinations of parameters to find the results that best suit your experiment.

Compound Filters tab

a Edit the parameters on the Compound Filters tab.

The parameters on this tab let you set the filter criteria for retaining features found in your sample data files based on the compound chromatogram. This filter allows you to remove features with low signal and poor quality.

1. Click the **Compound Filters** tab.



Compound Filters

Height

☐ Relative height >= 2.500 %

☑ Absolute height >= 5000 counts

☐ Limit to the largest 100 compounds

Compound quality

Quality score >= 30.00

Compound location

☐ Restrict retention times to minutes

Charge states

☐ Restrict charge states to Z

Figure 28 *Parameter values for the MFE Compound Filters tab*

2. Clear the **Relative height** check box.
3. Mark the **Absolute height** check box and type in a value of `5000` counts.

Note: In the final compound spectrum there must be at least one ion that is greater than or equal to the counts specified. The value typed is determined by empirically reviewing your mass spectral data. The absolute height in counts is different from the volume used to quantify the compound as a feature.

Note: Marking **Relative height** or **Limit to the largest** when performing metabolomics analyses is not recommended.

4. Clear the **Restrict retention times to** check box. Only mark this parameter and type in the time range if you know the chromatographic void volume, solvent peak, or other region containing unwanted peaks in the data set.
5. Clear the **Restrict charge states** check box because the charge state was previously limited to `1` in the Charge State tab. If a larger number of charge states is allowed, then mark this parameter and enter a charge state value to filter the results.
6. Mark the **Absolute height** check box, and type `5000` for the counts.
7. Type `30` for the Quality score.
8. Clear the **Restrict retention times to** check box.
9. Clear the **Restrict charge states to** check box.

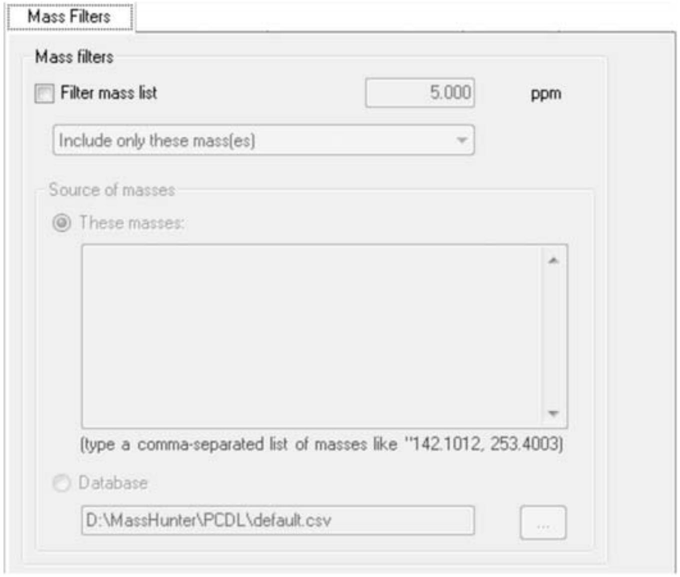
Mass Filters tab

b Edit the parameters on the Mass Filters tab.

The parameters on this tab let you remove noise due to specific ions from the data without regard for retention time. Mass Profiler Professional is the preferred place to perform mass filtering.

This filter feature can be unmarked and performed in Mass Profiler Professional more effectively, and in Mass Profiler Professional you have the ability to include retention time in the filter.

1. Click the **Mass Filters** tab.



Mass Filters

Mass filters

□ Filter mass list 5.000 ppm

Include only these mass(es)

Source of masses

● These masses:

(type a comma-separated list of masses like "142.1012, 253.4003)

○ Database

D:\MassHunter\PCDL\default.csv

Figure 29 *Parameter values for the MFE Mass Filters tab*

2. Clear the **Filter mass list** check box unless a specific list of neutral masses is known to be present in the data set that you wish to remove.
3. Select **Exclude these mass(es)** if you marked the **Filter mass list** check box.

Note: Selecting **Include only theses mass(es)** is not recommended.

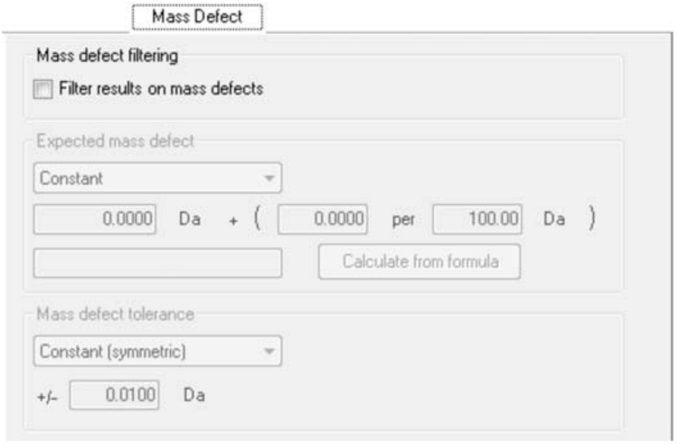
4. If **Filter mass list** is marked, click the appropriate button in the Source of masses group heading indicating your source of the masses for the exclude filter. Two example masses to exclude are `120.0434` and `921.0013`.

Mass Defect tab

c Edit the parameters on the Mass Defect tab.

The parameters on this tab let you supply a range for the mass defect within which the identified mass may still be a metabolite. Since the range necessary for filtering by mass defect must be rather large, it is not recommended to filter metabolomics results by mass defect.

1. Click the **Mass Defect** tab.



Mass Defect

Mass defect filtering

□ Filter results on mass defects

Expected mass defect

Constant

0.0000 Da + (0.0000 per 100.00 Da)

Calculate from formula

Mass defect tolerance

Constant (symmetric)

+/- 0.0100 Da

Figure 30 Parameter values for the MFE Mass Defect tab

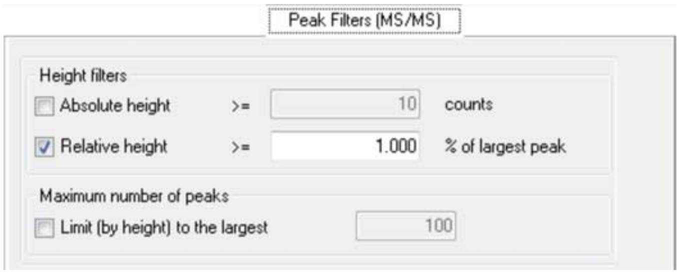
2. Clear the **Filter results on mass defects** check box.
3. If the **Filter results on mass defects** check box is marked, it is recommended to select **Variable** in the Expected mass defect group heading and type in values that work with your data set. If you select **Variable**, then the natural mass defect range increases with increasing mass.

Peak Filters (MS/MS) tab

d Edit the parameters on the Peak Filters (MS/MS) tab.

The parameters on this tab let you filter ions by height and quantity.

1. Click the **Peak Filters (MS/MS)** tab.



Peak Filters (MS/MS)

Height filters

□ Absolute height >= 10 counts

☑ Relative height >= 1.000 % of largest peak

Maximum number of peaks

□ Limit (by height) to the largest 100

Figure 31 Parameter values for the MFE Peak Filters (MS/MS) tab

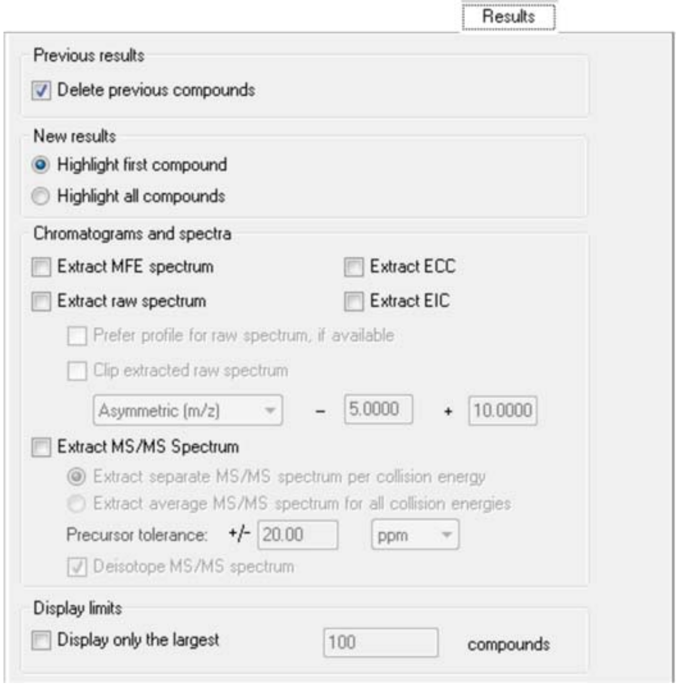
2. Clear the **Absolute height** check box. If this parameter is marked, do not type a counts value that is less than `10` counts. A typical counts value is around `100`. The best value to use is determined empirically.
3. Mark the **Relative height** check box and type `1` for the **% of largest peak**. The best analytical information is found using ions with an intensity at least within a factor of 100 of the base peak.
4. Clear the **Limit (by height) to the largest** check box.

Results tab

e Edit the parameters on the Results tab.

The parameters on this tab let you customize the display of your results. To improve the speed of the extraction, do not draw graphics when running molecular feature extraction. If more information about a feature is desired, it may be selectively obtained later instead of being generated for all of the features.

1. Click the **Results** tab.



Results

Previous results

☑ Delete previous compounds

New results

● Highlight first compound

○ Highlight all compounds

Chromatograms and spectra

□ Extract MFE spectrum □ Extract ECC

□ Extract raw spectrum □ Extract EIC

□ Prefer profile for raw spectrum, if available

□ Clip extracted raw spectrum

Asymmetric (m/z) – 5.0000 + 10.0000

□ Extract MS/MS Spectrum

● Extract separate MS/MS spectrum per collision energy

○ Extract average MS/MS spectrum for all collision energies

Precursor tolerance: +/- 20.00 ppm

☑ Deisotope MS/MS spectrum

Display limits

□ Display only the largest 100 compounds

Figure 32 *Parameter values for the MFE Results tab*

2. Mark the **Delete previous compounds** check box.
3. Click **Highlight first compound**.
4. Clear all of the check boxes in the Chromatograms and spectra group heading.
5. Clear the **Display only the largest** check box if you intend to create a CEF file.

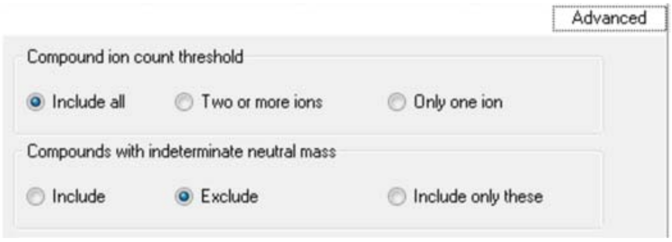
   Note: You must create a CEF file in order to import your molecular features into Mass Profiler Professional for the next step of the class prediction workflow.

Advanced tab

f Edit the parameters on the Advanced tab.

The parameters on this tab let you filter features by ion count and indeterminate neutral mass.

1. Click the **Advanced** tab.



Advanced

Compound ion count threshold

● Include all ○ Two or more ions ○ Only one ion

Compounds with indeterminate neutral mass

○ Include ● Exclude ○ Include only these

Figure 33 *Parameter values for the MFE Advanced tab*

2. Click **Include all** under the Compound ion count threshold group heading. Filtering by two or more ions is a very useful feature, but it can filter out valid ions with small molecular weights.
3. Click **Exclude** under the Compounds with indeterminate neutral mass group heading, especially when you click **Include all** under the *Compound ion count threshold* group heading. **Exclude** disregards features to which molecular feature extraction has not been able to assign a neutral mass.

Save your Find Compounds by Molecular Feature method

After you have edited your method to Find Compounds by Molecular Feature, it is recommended you save the method using a name that is readily distinguished from the name that is used later in this workflow for the method Find Compounds by Formula. By creating two distinct methods you can readily process your data without having to edit the workflow actions every time you switch between running MFE and FbF in the worklist.

Save your method.

a Click **Method** > **Save As**.

b Select the folder and type a method name, such as `Class_Prediction_W-FG_MFE.m`, in the **Save Method** dialog box. Add text, such as, `MFE` within your file name to distinguish it from the file name that is recommended in "Save your Find Compounds by Formula method" on page 114.

c Click **Save**.

Set the Export CEF Options

Export CEF Options specifies where MassHunter DA Reprocessor stores the resulting. CEF feature files and whether the files replace or overwrite any prior files.

1. Open the Method Editor for exporting CEF options.

a Click **Export** from within the **Method Explorer** window.

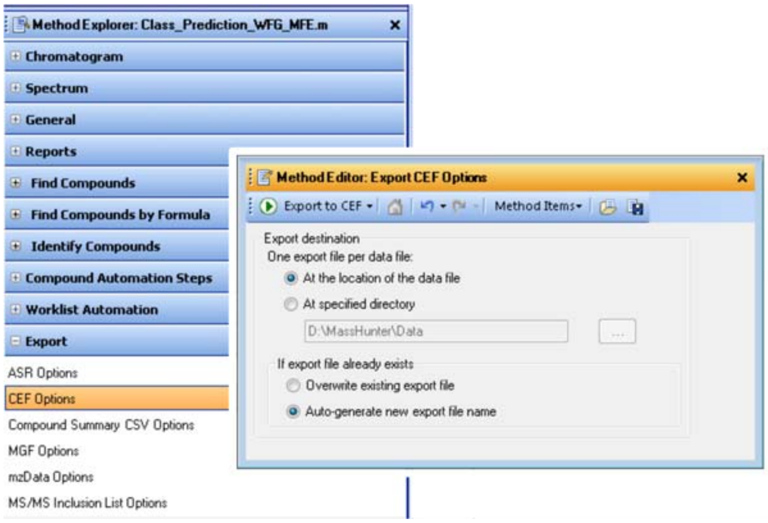
b Click **CEF Options**.

2. Enter the export destination settings for your method.

a Click **At the location of the data file**.

b Click **Auto-generate new export file name**.

c Save your method. Click the save method icon or click **Method > Save**.


Method Explorer: Class_Prediction_WFG_MFE.m
Chromatogram
Spectrum
General
Reports
Find Compounds
Find Compounds by Formula
Identify Compounds
Compound Automation Steps
Worklist Automation
Export
ASR Options
CEF Options
Compound Summary CSV Options
MGF Options
mzData Options
MS/MS Inclusion List Options
Method Editor: Export CEF Options
Export to CEF
Method Items
...

Export destination

One export file per data file:

● At the location of the data file

○ At specified directory

D:\MassHunter\Data

If export file already exists

○ Overwrite existing export file

● Auto-generate new export file name

Figure 34 Export CEF Options for use with DA Reprocessor

Enable the method to run in MassHunter DA Reprocessor

MassHunter software can most efficiently perform computationally intensive tasks, such as finding features, on multiple data files by using MassHunter DA Reprocessor. The following steps enable your method to run using DA Reprocessor.

1. Open the Method Editor to assign actions to run from the worklist.

a Click **Worklist Automation** from the Method Explorer window.

b Click **Worklist Actions**.

2. Remove all actions from the **Actions to be run list**.

a Double-click on any action in the **Actions to be run** list. The action is automatically removed from the **Actions to be run** list. As an alternate to the double-click, you can click on an action in the **Actions to be run** list and then click the delete icon ✕.

b Repeat action removal until the **Actions to be run** list is empty.

c Save your method. Click the save method icon or click **Method > Save**.

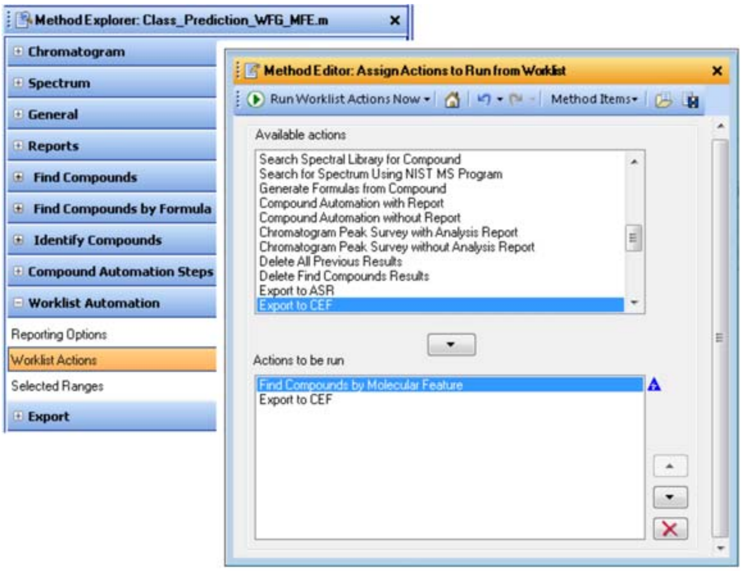
3. Add new actions to the **Actions to be run** list.

a Double-click the **Find Compounds by Molecular Feature** action in the **Available actions** list.

The action is automatically added to the **Actions to be run** list. As an alternate to the double-click, you can click the action and then click the down arrow button ▼ to add the action to the **Actions to be run** list.

b Double-click the **Export to CEF** action in the **Available actions** list. The **Export to CEF** action must be listed after the **Find Compounds by Molecular Feature** action as shown in Figure 35.

c Save your method. Click the save method icon or click **Method > Save**.



Available actions

- Search Spectral Library for Compound
- Search for Spectrum Using NIST MS Program
- Generate Formulas from Compound
- Compound Automation with Report
- Compound Automation without Report
- Chromatogram Peak Survey with Analysis Report
- Chromatogram Peak Survey without Analysis Report
- Delete All Previous Results
- Delete Find Compounds Results
- Export to ASR
- Export to CEF

Actions to be run

- Find Compounds by Molecular Feature
- Export to CEF

Figure 35 *Assign Actions to Run from Worklist for use with DA Reprocessor*

Confirm the MFE method on a single data file

Class prediction involves the analysis of a large number of sample files with each sample containing a large number of compounds. Find Compounds by Molecular Feature is therefore run on the entire sample data set using MassHunter DA Reprocessor. However, before the entire sample set is run in MassHunter DA Reprocessor, you can process a single file within MassHunter Qualitative Analysis to verify the new MFE parameters.

1. Find Compounds by Molecular Feature for a single data file.

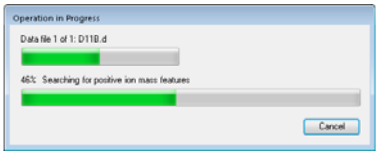
a Click **File > Open Data File**.

b Click on a single data file in the **Open Data File** dialog box.

c Click **Open**.

d Click **Actions > Find Compounds by Molecular Feature**, or click the Find Compounds by Molecular Feature button ▶ Find Compounds by Molecular Feature in the **Method Editor: Find Compounds by Molecular Feature** section in the Method Editor window. Molecular feature extraction begins immediately and the progress is shown in an **Operation in Progress** status box as shown in Figure 36.

If no data file is open, or an inappropriate data file is open, a message box appears as shown in Figure 37. Click **OK** and open a single data file.

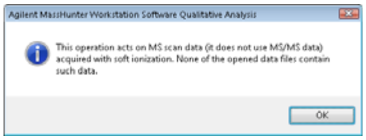


Operation in Progress

Data file 1 of 1: D11B.d

46% Searching for positive ion mass features

Figure 36 *Find Compounds by Molecular Feature progress box*



Agilent MassHunter Workstation Software Qualitative Analysis

This operation acts on MS scan data (it does not use MS/MS data) acquired with soft ionization. None of the opened data files contain such data.

Figure 37 *Message box*

2. Display and review the Compound List.

When molecular feature extraction finishes processing the data file, the results are displayed in several windows within MassHunter Qualitative Analysis. The results may be reviewed and arranged to meet your preferences.

a Set up the recommended columns for viewing your data in the Compound List.

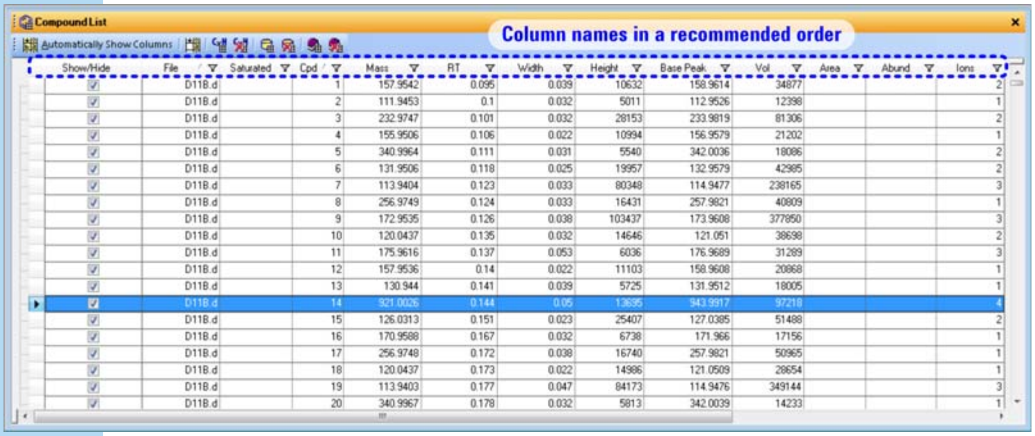
1. Right-click anywhere in the **Compound List** window.
2. Click **Add/Remove Columns** to open the **(Enhanced) Add/Remove Columns** dialog box.
3. Click **Clear All**.
4. Click the **Column Name** column header twice to sort the column names in ascending alphabetical order.
5. Mark the check boxes for at least the following Column Names: **Abund**, **Area**, **Base Peak**, **Cpd**, **File**, **Height**, **Ions**, **Mass**, **RT**, **Saturated**, **Show/Hide**, **Vol**, and **Width**.
   - Each of these columns is documented in the MassHunter Qualitative Analysis Help in Reference > Columns > Compound List Table Columns under the Contents tab.
   - Blank entries for the **Area** and **Abund** columns are normal.

- All of the columns are exported to the CEF file.

6. Click **OK**.

b Arrange the order of the columns and your compound data in the Compound List.

1. Click and drag the column names left or right so that they are arranged in the order you like. A useful order is shown in Figure 38.


Compound List
Automatically Show Columns
Column names in a recommended order

Show/Hide	File	Saturated	Cpd	Mass	RT	Width	Height	Base Peak	Vol	Area	Abund	Ions
☑	D11B.d		1	157.9542	0.095	0.039	10632	158.9614	34877			2
☑	D11B.d		2	111.9453	0.1	0.032	5011	112.9526	12398			1
☑	D11B.d		3	232.9747	0.101	0.032	28153	233.9819	81306			2
☑	D11B.d		4	155.9506	0.106	0.022	10994	156.9579	21202			1
☑	D11B.d		5	340.9964	0.111	0.031	5540	342.0036	18086			2
☑	D11B.d		6	131.9506	0.118	0.025	19957	132.9579	42985			2
☑	D11B.d		7	113.9404	0.123	0.033	80348	114.9477	238165			3
☑	D11B.d		8	256.9749	0.124	0.033	16431	257.9821	40809			1
☑	D11B.d		9	172.9535	0.126	0.038	103437	173.9608	377850			3
☑	D11B.d		10	120.0437	0.135	0.032	14646	121.051	38698			2
☑	D11B.d		11	175.9616	0.137	0.053	6036	176.9689	31289			3
☑	D11B.d		12	157.9536	0.14	0.022	11103	158.9608	20868			1
☑	D11B.d		13	130.944	0.141	0.039	5725	131.9512	18005			1
☑	D11B.d		14	921.0026	0.144	0.05	13695	943.9917	97218			4
☑	D11B.d		15	126.0313	0.151	0.023	25407	127.0385	51488			2
☑	D11B.d		16	170.9588	0.167	0.032	6738	171.966	17156			1
☑	D11B.d		17	256.9748	0.172	0.038	16740	257.9821	50965			1
☑	D11B.d		18	120.0437	0.173	0.022	14986	121.0509	28654			1
☑	D11B.d		19	113.9403	0.177	0.047	84173	114.9476	349144			3
☑	D11B.d		20	340.9967	0.178	0.032	5813	342.0039	14233			1

Figure 38 *Columns arranged in the Compound List window*

2. Click the **Height** column heading to sort the compounds by ascending height (the abundance value of the base peak). The compounds with a lower height value are shown at the top of the list.

3. *Optional* - Extract results.

This step is optional.

a Select the compounds to view and compare chromatogram and MS results.

1. Click and drag across the first few compound rows to select multiple compounds (e.g., select around ten compounds). The selected compounds are highlighted.
2. Right-click the **Compound List** window and click **Extract Complete Result Set**.

The results are displayed in the **Chromatogram Results** and **MS Spectrum Results** windows. These windows are updated when you use the arrow keys to move up and down the Compound List. See Figure 39 on page 58.

- In the **Chromatogram Results** window, the extracted ion chromatogram (EIC) for each compound is compared to the extracted compound chromatogram (ECC) for the ions contained in the molecular features.
- In the **MS Spectrum Results** window, the compound spectrum for each compound is compared to the scan data spectrum.
- A compound may be deleted and removed from the features available for exporting by highlighting the compound and then pressing the delete key.

b If a significant number of compounds are too weak to provide confidence as a molecular feature, adjust the Find by Molecular Feature parameters. Weak compounds have small values for Height and Vol in the **Compound List** (see Figure 38) and have low count values in the **Chromatogram Results** and **MS Spectrum Results** (see Figure 39 on page 58).

1. Adjust the parameters entered in the "Compound Filters tab" on page 49.

2. Re-run **Find Compounds by Molecular Feature**. Molecular feature extraction runs very quickly if no changes are made to the tabs that control compound finding.
3. Review the new results by repeating step 2 - "Display and review the Compound List." on page 56.
4. Repeat these steps until you are satisfied with the molecular feature results.

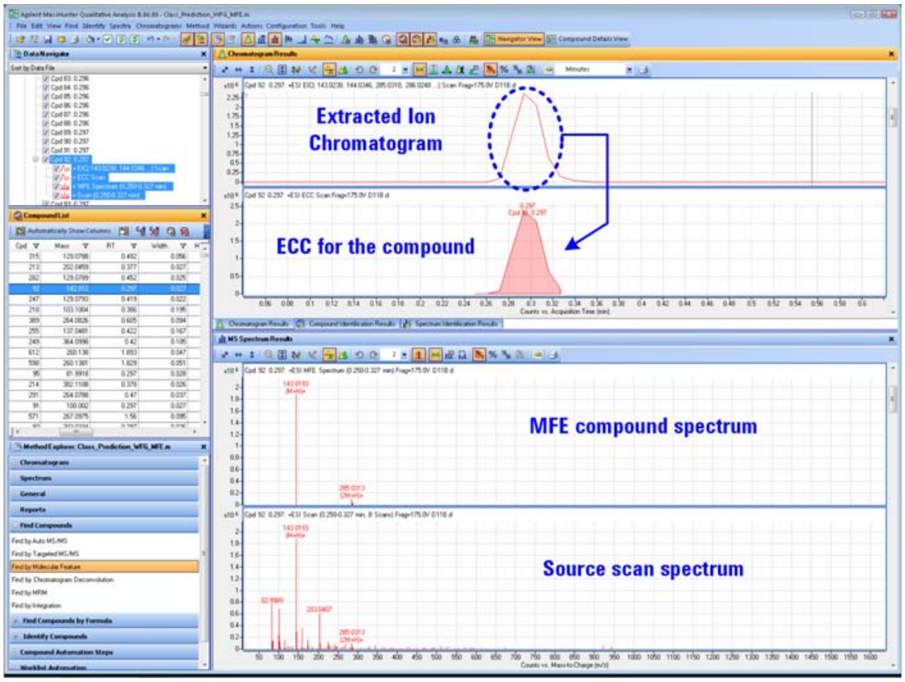


Figure 39 *Display of the Extract Complete Result Set from the Compound List*

4. *Optional* - Export the results for the single sample to a CEF file.

This step is optional. The CEF files for all of the samples are generated in "Find compounds using DA Reprocessor"

a Click **File > Export > as CEF**. The **Export CEF Options** dialog box is opened.

b Select the data files to be exported from the **List of opened data files**. Create a new folder for the exported CEF files to aid documentation of the class prediction workflow and to make it easier to distinguish any new CEF files from previous CEF files.

c Update the other parameters in the **Export CEF Options** dialog box.

d Click **OK**.

You can review the results from this step by importing the CEF back into MassHunter Qualitative Analysis by following step 3 - "Display and review the Compound List after running MassHunter DA Reprocessor." on page 60.

Find compounds using DA Reprocessor

Class prediction involves applying your method to a large number of sample files, each of which may contain a large number of compounds. MassHunter Qualitative Analysis can be used to process all of your data sets. However, MassHunter DA Reprocessor provides a more efficient and automated means to run your MassHunter Qualitative Analysis method on multiple sample files. Therefore your method is run on the entire class prediction sample set using DA Reprocessor.

1. Close your data file.

a Click **File > Close Data File**.

b Click **No**. Do not save the results.

2. Find Compounds by Molecular Feature using MassHunter DA Reprocessor.

a Click the DA Reprocessor icon located on the desktop, or click **Start > All Programs > Agilent > MassHunter Workstation > Acq Tools > DA Reprocessor**.

Press **F1** from within MassHunter DA Reprocessor to start online Help. For example, click **DA Reprocessor > Shortcut Menu for Worklist (from the top left cell)** for instructions on creating a worklist containing multiple samples.

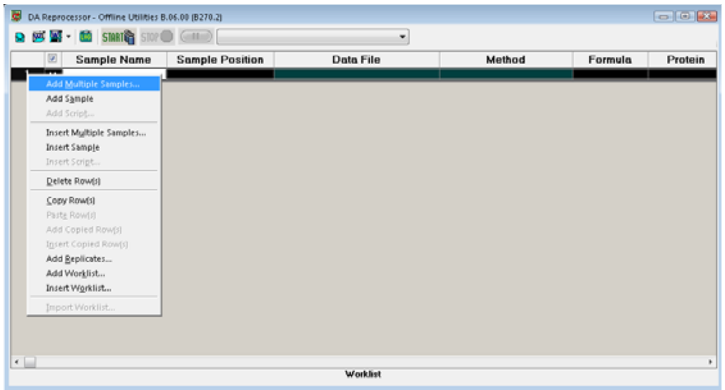


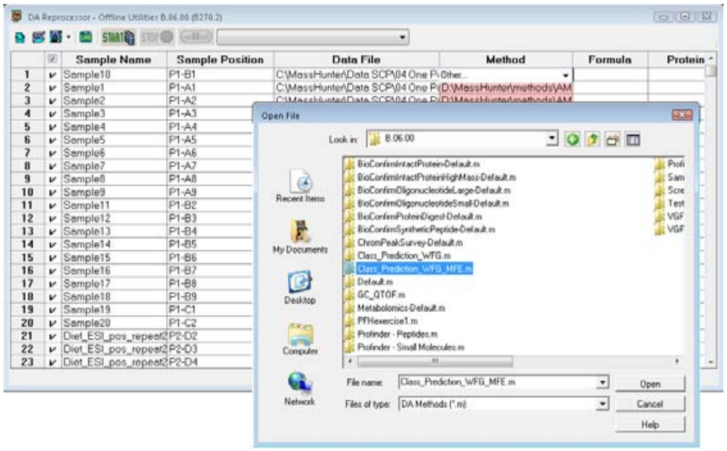
Figure 40 *Adding samples to the MassHunter DA Reprocessor worklist*

b Right-click the top left cell of the worklist and click **Add Multiple Samples** as shown in Figure 40.

c Select the folder and file names that refer to your samples.

d Click **Open**.

e Click the Method name for the first sample in row 1 and select the name of the method saved from MassHunter Qualitative Analysis. If the method you saved is not in the immediate list, select “Other”, and then you can select the folder and method using the **Open File** dialog box as shown in Figure 41 on page 60.



DA Reprocessor - Offline Utilities B.06.00 (B270.2)

	☑	Sample Name	Sample Position	Data File	Method	Formula	Protein
1	✓	Sample10	P1-B1	C:\MassHunter\Data SCP\04 One P	Other...		
2	✓	Sample1	P1-A1	C:\MassHunter\Data SCP\04 One P	D:\MassHunter\methods\AM		
3	✓	Sample2	P1-A2	C:\MassHunter\Data SCP\04 One P	D:\MassHunter\methods\AM		
4	✓	Sample3	P1-A3				
5	✓	Sample4	P1-A4				
6	✓	Sample5	P1-A5				
7	✓	Sample6	P1-A6				
8	✓	Sample7	P1-A7				
9	✓	Sample8	P1-A8				
10	✓	Sample9	P1-A9				
11	✓	Sample11	P1-B2				
12	✓	Sample12	P1-B3				
13	✓	Sample13	P1-B4				
14	✓	Sample14	P1-B5				
15	✓	Sample15	P1-B6				
16	✓	Sample16	P1-B7				
17	✓	Sample17	P1-B8				
18	✓	Sample18	P1-B9				
19	✓	Sample19	P1-C1				
20	✓	Sample20	P1-C2				
21	✓	Diet_ESI_pos_repeat2	P2-D2				
22	✓	Diet_ESI_pos_repeat2	P2-D3				
23	✓	Diet_ESI_pos_repeat2	P2-D4				

Open File

Look in: B.06.00

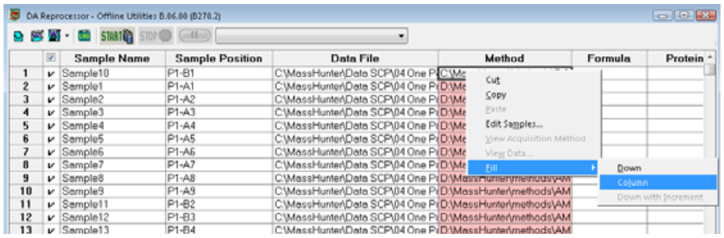
BioConfirmIntactProtein-Default.m
BioConfirmIntactProteinHighMass-Default.m
BioConfirmOligonucleotideLarge-Default.m
BioConfirmOligonucleotideSmall-Default.m
BioConfirmProteinDigest-Default.m
BioConfirmSyntheticPeptide-Default.m
ChromPeakSurvey-Default.m
Class_Prediction_WFG.m
Class_Prediction_WFG_MFE.m
Default.m
GC_QTOF.m
Metabolomics-Default.m
PFHexercise1.m
Profinder - Peptides.m
Profinder - Small Molecules.m
Profi
Sam
Scre
Test
VGF
VGF

File name: Class_Prediction_WFG_MFE.m

Files of type: DA Methods (*.m)

Figure 41 *Selecting the MassHunter Qualitative Analysis method*

f Copy the method from the first sample to each of the samples in the worklist; right-click on the method in row 1, and click **Fill > Column** (see Figure 42).



DA Reprocessor - Offline Utilities B.06.00 (B270.2)

	☑	Sample Name	Sample Position	Data File	Method	Formula	Protein
1	✓	Sample10	P1-B1	C:\MassHunter\Data SCP\04 One P	C:\Ma		
2	✓	Sample1	P1-A1	C:\MassHunter\Data SCP\04 One P	D:\Me		
3	✓	Sample2	P1-A2	C:\MassHunter\Data SCP\04 One P	D:\Me		
4	✓	Sample3	P1-A3	C:\MassHunter\Data SCP\04 One P	D:\Me		
5	✓	Sample4	P1-A4	C:\MassHunter\Data SCP\04 One P	D:\Me		
6	✓	Sample5	P1-A5	C:\MassHunter\Data SCP\04 One P	D:\Me		
7	✓	Sample6	P1-A6	C:\MassHunter\Data SCP\04 One P	D:\Me		
8	✓	Sample7	P1-A7	C:\MassHunter\Data SCP\04 One P	D:\Me		
9	✓	Sample8	P1-A8	C:\MassHunter\Data SCP\04 One P	D:\MassHunter\methods\AM		
10	✓	Sample9	P1-A9	C:\MassHunter\Data SCP\04 One P	D:\MassHunter\methods\AM		
11	✓	Sample11	P1-B2	C:\MassHunter\Data SCP\04 One P	D:\MassHunter\methods\AM		
12	✓	Sample12	P1-B3	C:\MassHunter\Data SCP\04 One P	D:\MassHunter\methods\AM		
13	✓	Sample13	P1-B4	C:\MassHunter\Data SCP\04 One P	D:\MassHunter\methods\AM		

Figure 42 *Copying the data analysis method to each sample in the worklist*

g Click the **Start** icon (START) in the toolbar to run the worklist. The progress is indicated on the worklist sheet as each sample is completed.

The CEF files containing the molecular features from the samples are automatically placed in the folder containing the data files. Each CEF file has the same root name as the sample data file. You import the CEF files into Mass Profiler Professional for feature selection in the next step of the class prediction workflow.

3. Display and review the Compound List after running MassHunter DA Reprocessor.

a Return to MassHunter Qualitative Analysis. If you closed the MassHunter Qualitative Analysis program, do the following:

- Click **Start > All Programs > Agilent > MassHunter Workstation > Qualitative Analysis B.06.00**. The version number may instead be **B.07.00** or later.
- Click **Cancel** when the **Open Data File** dialog box opens.

b Click **File > Close All** to close any open data files. Do not save any results.

c Click **File > Open Data File** to open one of the original sample data files including the chromatographic data and the results of Find Compounds by Molecular Feature. Mark the **Load result data** check box.

or

Click **File > Import Compound** to open one of the CEF files that contains the molecular feature results of Find Compounds by Molecular Feature.

Note: Because of the large number of features in a typical sample file, it is recommended to open only one file at a time to review the results. Close the open file and then open the next file.

d Display and review the Compound List as described previously in step 2 - "Display and review the Compound List." on page 56. The chromatographic results are only visible if the original data file is opened.

4. *Optional* - Extract results.

This step is optional.

Extract your MS results as described previously in step 3 - "Optional - Extract results." on page 57.

MassHunter Profinder

The following examples use MassHunter Profinder B.06.00 running on 64-bit Windows 7 Professional. If you already found your features using Qualitative Analysis skip to "Next step..." on page 66.

Finding the untargeted features in your sample data files using Profinder involves one self-directed wizard that involves four steps as shown in Figure 43.

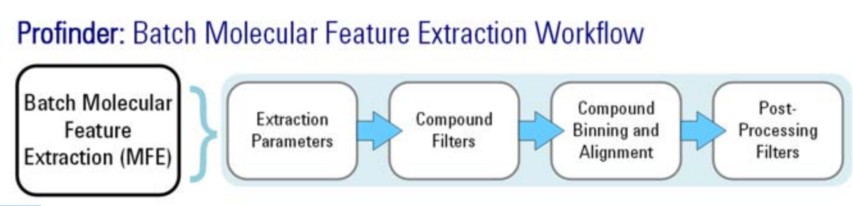

Profinder: Batch Molecular Feature Extraction Workflow
Batch Molecular Feature Extraction (MFE)
Extraction Parameters
Compound Filters
Compound Binning and Alignment
Post-Processing Filters

Figure 43 *Steps to find untargeted features using Profinder*

The parameters recommended in the following steps are part of the Batch Molecular Feature Extraction workflow wizard in Profinder. See *Agilent G3835AA MassHunter Profinder Software - Quick Start Guide* for additional information on using Profinder.

1. Start MassHunter Profinder software.

MassHunter Profinder a software tool used to perform the function of finding molecular features in TOF and Q-TOF data files. After the molecular features are found they are imported into Mass Profiler Professional for statistical analysis. Feature finding is an essential prerequisite to using Mass Profiler Professional.

a Double-click the Profinder icon located on the desktop,

Click **Start > All Programs > Agilent > MassHunter Workstation > Profinder B.06.00**.

b Click the **File > Add/Remove Sample Files** or Add/Remove Sample Files... in the toolbar to begin the workflow.

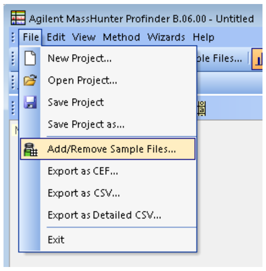

Agilent MassHunter Profinder B.06.00 - Untitled
File Edit View Method Wizards Help
New Project...
Open Project...
Save Project
Save Project as...
Add/Remove Sample Files...
Export as CEF...
Export as CSV...
Export as Detailed CSV...
Exit

Figure 44 Add sample files to begin feature finding in Profinder

2. Add all of the sample data files to Profinder.

a Click **Add file(s)** in the **Add/Remove Sample Files** dialog box.

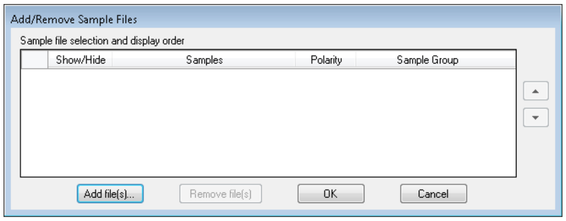
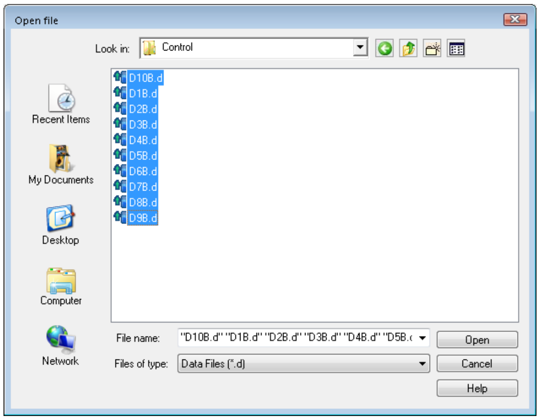


Figure 45 Add/Remove Sample Files dialog box

b Navigate to the folder containing your raw sample data files and in the **Open File** dialog box.

c Select your raw sample data files in the **Open File** dialog box.

d Click **Open**.



Look in: Control

D10B.d
D1B.d
D2B.d
D3B.d
D4B.d
D5B.d
D6B.d
D7B.d
D8B.d
D9B.d

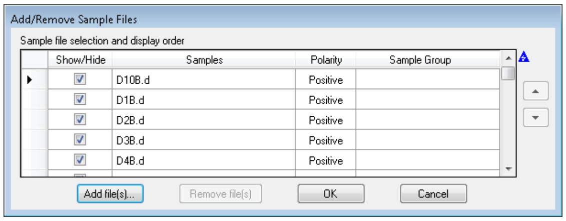
File name: "D10B.d" "D1B.d" "D2B.d" "D3B.d" "D4B.d" "D5B.(

Files of type: Data Files (*.d)

Figure 46 Add/remove Sample Files dialog box

e Repeat steps a through d if your sample data files reside in multiple folders.

f Click **OK** in the **Add/Remove Sample Files** dialog box when all of your sample data files are selected.



Add/Remove Sample Files

Sample file selection and display order

Show/Hide	Samples	Polarity	Sample Group
☑	D10B.d	Positive	
☑	D1B.d	Positive	
☑	D2B.d	Positive	
☑	D3B.d	Positive	
☑	D4B.d	Positive	

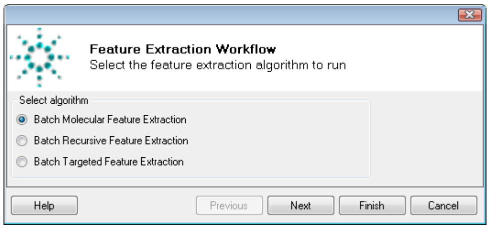
Figure 47 *Samples added to the Add/remove Sample Files dialog box*

3. Select the Batch Molecular Feature Extraction workflow.

Begin the Batch Molecular Feature Extraction (MFE) workflow.

a Click **Batch Molecular Feature Extraction**.

b Click **Next**.



Feature Extraction Workflow

Select the feature extraction algorithm to run

Select algorithm

● Batch Molecular Feature Extraction

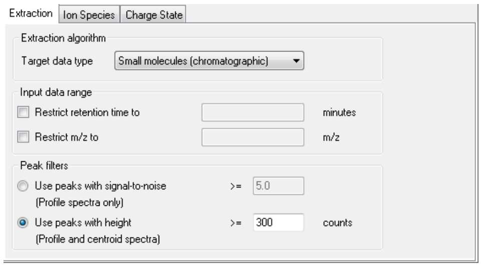
○ Batch Recursive Feature Extraction

○ Batch Targeted Feature Extraction

Figure 48 *Select Batch Molecular Feature Extraction*

4. Enter the extraction parameters in **(Step 1 of 4)** of the MFE workflow.

a Enter the parameters in the **Extraction** tab.



Extraction algorithm

Target data type: Small molecules (chromatographic)

Input data range

□ Restrict retention time to minutes

□ Restrict m/z to m/z

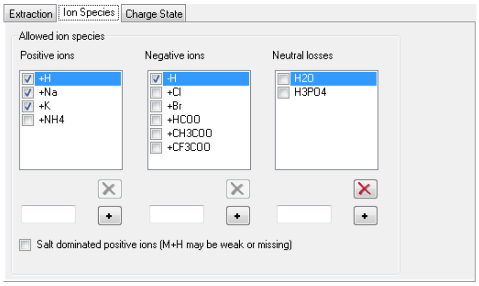
Peak filters

○ Use peaks with signal-to-noise >= 5.0 (Profile spectra only)

● Use peaks with height >= 300 counts (Profile and centroid spectra)

Figure 49 *Extraction Parameters Step 1 of 4 - Extraction tab*

b Enter the parameters in the **Ion Species** tab.



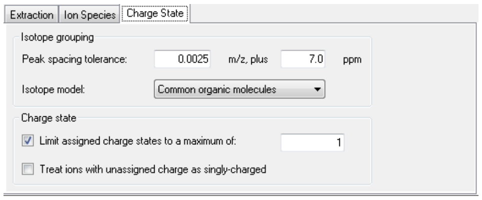
Allowed ion species

Positive ions	Negative ions	Neutral losses
☑ +H	☑ -H	□ H2O
☑ +Na	□ +Cl	□ H3PO4
☑ +K	□ +Br	
□ +NH4	□ +HCOO	
	□ +CH3COO	
	□ +CF3COO	

□ Salt dominated positive ions (M+H may be weak or missing)

Figure 50 *Extraction Parameters Step 1 of 4 - Ion Species tab*

c Enter the parameters in the **Charge State** tab.



Isotope grouping

Peak spacing tolerance: 0.0025 m/z, plus 7.0 ppm

Isotope model: Common organic molecules

Charge state

☑ Limit assigned charge states to a maximum of: 1

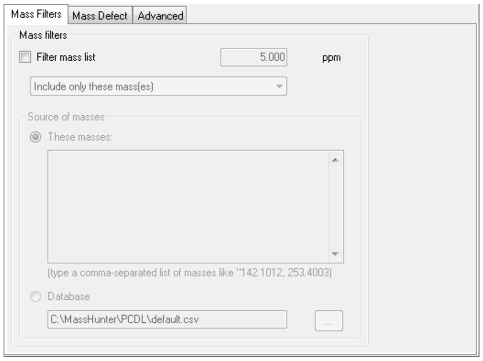
□ Treat ions with unassigned charge as singly-charged

Figure 51 *Extraction Parameters Step 1 of 4 - Charge State tab*

d Click **Next**.

5. Enter the compound filters parameters in **(Step 2 of 4)** of the MFE workflow.

a Enter the parameters in the **Mass Filters** tab.



Mass filters

□ Filter mass list 5.000 ppm

Include only these mass(es)

Source of masses

● These masses:

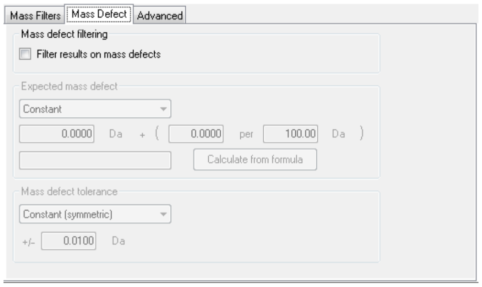
(type a comma-separated list of masses like "142.1012, 253.4003)

○ Database

C:\MassHunter\PCDL\default.csv

Figure 52 *Compound Filters Step 2 of 4 - Mass Filters tab*

b Enter the parameters in the **Ion Species** tab.



Mass defect filtering

□ Filter results on mass defects

Expected mass defect

Constant

0.0000 Da + (0.0000 per 100.00 Da)

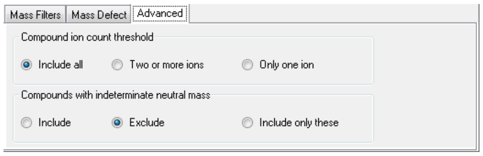
Mass defect tolerance

Constant (symmetric)

+/- 0.0100 Da

Figure 53 *Compound Filters Step 2 of 4 - Mass Defect tab*

c Enter the parameters in the **Charge State** tab.



Compound ion count threshold

● Include all ○ Two or more ions ○ Only one ion

Compounds with indeterminate neutral mass

○ Include ● Exclude ○ Include only these

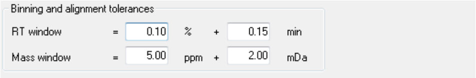
Figure 54 *Compound Filters Step 2 of 4 - Advanced tab*

d Click **Next**.

6. Enter the compound binning and alignment parameters in **(Step 3 of 4)** of the MFE workflow.

a Enter the parameters for compound binning and alignment.

b Click **Next**.



Binning and alignment tolerances

RT window = 0.10 % + 0.15 min

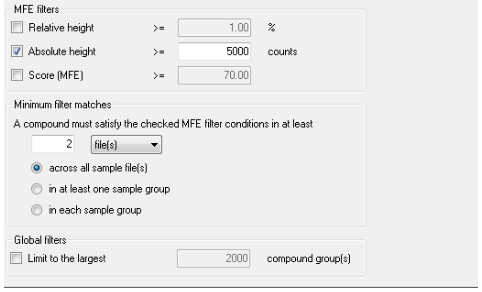
Mass window = 5.00 ppm + 2.00 mDa

Figure 55 *Compound Binning and Alignment Step 3 of 4*

7. Enter the parameters for the post-processing filters in **(Step 4 of 4)** of the MFE workflow.

a Enter the parameters for the post-processing filters.

b Click **Finish**. Feature finding in the sample data files begins immediately.



MFE filters

□ Relative height >= 1.00 %

☑ Absolute height >= 5000 counts

□ Score (MFE) >= 70.00

Minimum filter matches

A compound must satisfy the checked MFE filter conditions in at least

2 file(s)

● across all sample file(s)

○ in at least one sample group

○ in each sample group

Global filters

□ Limit to the largest 2000 compound group(s)

Figure 56 *Post-Processing filters Step 4 of 4*

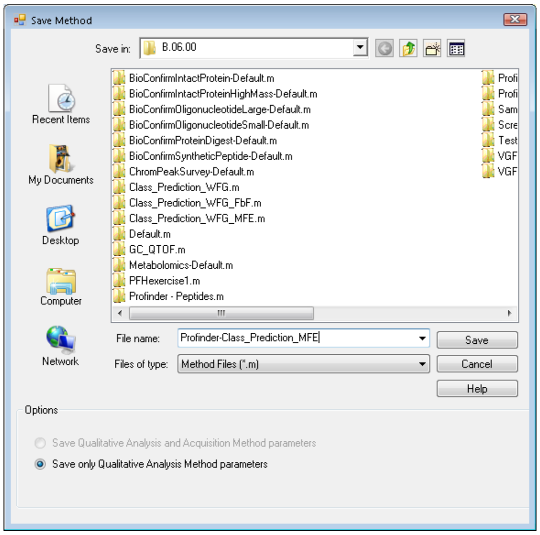
8. Save your method.

a Click **Method > Save As** to save your method.

b Navigate to the appropriate folder.

c Enter your file name.

d Click **Save**.



Save in: B.06.00

BioConfirmIntactProtein-Default.m
BioConfirmIntactProteinHighMass-Default.m
BioConfirmOligonucleotideLarge-Default.m
BioConfirmOligonucleotideSmall-Default.m
BioConfirmProteinDigest-Default.m
BioConfirmSyntheticPeptide-Default.m
ChromPeakSurvey-Default.m
Class_Prediction_WFG.m
Class_Prediction_WFG_FbF.m
Class_Prediction_WFG_MFE.m
Default.m
GC_QTOF.m
Metabolomics-Default.m
PFHexercise1.m
Profinder - Peptides.m
Profi
Profi
Sam
Scre
Test
VGF
VGF

File name: Profinder-Class_Prediction_MFE

Files of type: Method Files (*.m)

Options

○ Save Qualitative Analysis and Acquisition Method parameters

● Save only Qualitative Analysis Method parameters

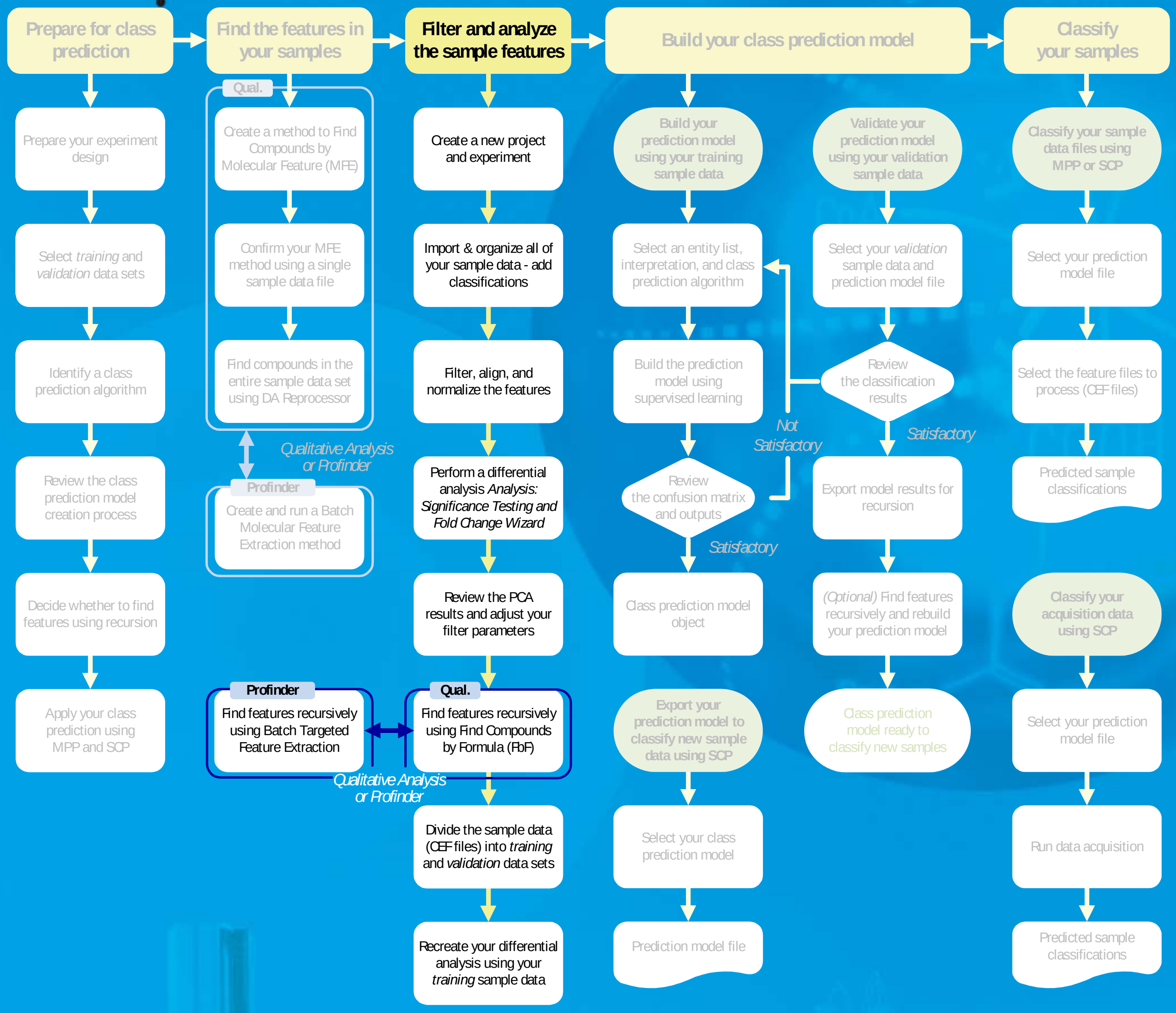
Figure 57 *Saving a Profinder method*

Next step...

You have now completed the find features step of the class prediction. In the next workflow step you import your MFE results into Mass Profiler Professional to filter and analyze the features.

Filter and analyze the sample features

Import the extracted features from your sample data set into MPP, create an initial differential analysis, and review the results for suitability to perform class prediction.



Prepare for class prediction

Prepare your experiment design

Select *training* and *validation* data sets

Identify a class prediction algorithm

Review the class prediction model creation process

Decide whether to find features using recursion

Apply your class prediction using MPP and SCP

Find the features in your samples

Qual.

Create a method to Find Compounds by Molecular Feature (MFE)

Confirm your MFE method using a single sample data file

Find compounds in the entire sample data set using DA Reprocessor

Qualitative Analysis or Profinder

Profinder

Create and run a Batch Molecular Feature Extraction method

Profinder

Find features recursively using Batch Targeted Feature Extraction

Qualitative Analysis or Profinder

Filter and analyze the sample features

Create a new project and experiment

Import & organize all of your sample data - add classifications

Filter, align, and normalize the features

Perform a differential analysis *Analysis: Significance Testing and Fold Change Wizard*

Review the PCA results and adjust your filter parameters

Qual.

Find features recursively using Find Compounds by Formula (FbF)

Divide the sample data (CEF files) into *training* and *validation* data sets

Recreate your differential analysis using your *training* sample data

Build your class prediction model

Build your prediction model using your training sample data

Select an entity list, interpretation, and class prediction algorithm

Build the prediction model using supervised learning

Review the confusion matrix and outputs

Satisfactory

Class prediction model object

Export your prediction model to classify new sample data using SCP

Select your class prediction model

Prediction model file

Not Satisfactory

Validate your prediction model using your validation sample data

Select your *validation* sample data and prediction model file

Review the classification results

Satisfactory

Export model results for recursion

(Optional) Find features recursively and rebuild your prediction model

Class prediction model ready to classify new samples

Classify your samples

Classify your sample data files using MPP or SCP

Select your prediction model file

Select the feature files to process (CEF files)

Predicted sample classifications

Classify your acquisition data using SCP

Select your prediction model file

Run data acquisition

Predicted sample classifications

Agilent Technologies

Create a new project and experiment

During this step of the class prediction workflow you import the CEF files created from MassHunter DA Reprocessor, or Profinder, as a new experiment into Mass Profiler Professional. MPP guides you through an initial differential analysis to help you assess the suitability of your data set to perform class prediction.

Because the advanced operations available in the Workflow Browser do not guide you through the initial steps of data import and differential analysis, it is not recommended to skip this section of the class prediction workflow. All parameters, including the default parameters used during the MS Experiment Creation Wizard, can be edited at the conclusion of the differential analysis by using the operations available in the Workflow Browser (see Figure 93 on page 97).

The following examples use Mass Profiler Professional B.12.61 running on 64-bit, Windows 7 Professional.

Note: To obtain help and detailed information regarding the various fields and statistical treatments press the **F1** key on the keyboard or refer to the *Mass Profiler Professional User Manual*.

Launch Mass Profiler Professional

Double-click the Mass Profiler Professional icon located on the desktop, or click **Start > All Programs > Agilent > MassHunter Workstation > Mass Profiler Professional > Mass Profiler Professional**.

If MPP is already open, close the project. Then, click the **New project** icon or click **Project > New Project**, and begin the workflow at "Enter descriptive information in the Create New Project dialog box." on page 69.

Set up a project and an experiment

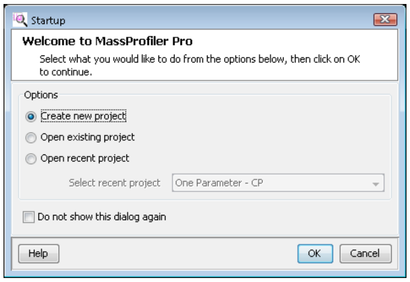
A project is a container for a collection of experiments. A project can have multiple experiments on different sample types and organisms. When you create a new project you are guided through four steps:

- **Startup:** Create your new project.
- **Create New Project:** Add descriptive information about the project.
- **Experiment Selection Dialog:** Create a new experiment to your project.
- **New Experiment:** Add custom information to store with the experiment.

1. Create a new project in the **Startup** dialog box.

a Click **Create new project**.

b Click **OK**.



Startup

Welcome to MassProfiler Pro

Select what you would like to do from the options below, then click on OK to continue.

Options

● Create new project

○ Open existing project

○ Open recent project

Select recent project: One Parameter - CP

□ Do not show this dialog again

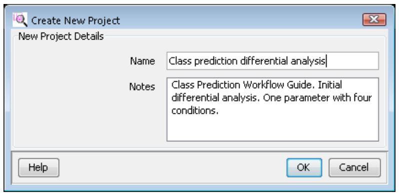
Figure 58 Startup dialog box

2. Enter descriptive information in the **Create New Project** dialog box.

a Type `Class prediction differential analysis` as a descriptive **Name** for the project.

b Type `Class Prediction Workflow Guide. Initial differential analysis. One parameter with four conditions.` as a descriptive **Notes** for the project.

c Click **OK**.



Create New Project

New Project Details

Name: Class prediction differential analysis

Notes: Class Prediction Workflow Guide. Initial differential analysis. One parameter with four conditions.

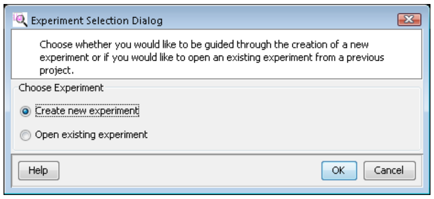
Figure 59 Create New Project dialog box

3. Select your experiment origin in the **Experiment Selection Dialog** dialog box.

Specify whether the wizard guides you through creating a new experiment or whether the wizard opens an existing experiment.

a Click **Create new experiment**.

b Click **OK**. If you clicked the **Open existing experiment** button, you are prompted for the experiment to add to the analysis.



Experiment Selection Dialog

Choose whether you would like to be guided through the creation of a new experiment or if you would like to open an existing experiment from a previous project.

Choose Experiment

● Create new experiment

○ Open existing experiment

Figure 60 Experiment Selection Dialog dialog box

4. Type and select information that guides the experiment creation in the **New Experiment** dialog box.

Available entry options for the **New Experiment** dialog box depend on your experiment type and data sources as outlined by Table 1 and Table 2 in the *Agilent G3835AA MassHunter Mass Profiler Professional - Familiarization Guide* (G3835-90010, Revision A, November 2012).

a Type `Differential Analysis` in **Experiment name**. This entry may be different from the project name previously entered.

b Select **Mass Profiler Professional** for the **Analysis type** to enable class prediction. Only your licensed analysis types are available.

c Select **Unidentified** for the **Experiment type**. Unidentified is the proper selection when the compound features have only been identified by their neutral mass and retention time using molecular feature extraction. The experiment type selection determines how Mass Profiler Professional manages the data. Use Combined (Identified + Unidentified) when you are unsure if the data is identified in full or in part or when MassHunter Qualitative Analysis has been used previously to identify some of the compound features.

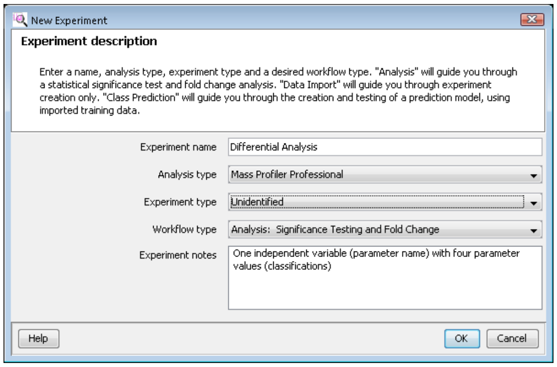
d Select **Analysis: Significance Testing and Fold Change** for the **Workflow type**.

When you select *Analysis: Significance Testing and Fold Change,* the workflow still takes you through the MS Experiment Creation Wizard first.

Regardless of your personal expertise, the *Analysis: Significance Testing and Fold Change* workflow provides you with quality control to your analysis that improves your results. You may customize the entire analysis at the conclusion of the workflow.

e Type `One independent variable (parameter name) with four parameter values (classifications)` in the **Experiment notes**.

f Click **OK**.



Experiment description

Enter a name, analysis type, experiment type and a desired workflow type. "Analysis" will guide you through a statistical significance test and fold change analysis. "Data Import" will guide you through experiment creation only. "Class Prediction" will guide you through the creation and testing of a prediction model, using imported training data.

Experiment name: Differential Analysis

Analysis type: Mass Profiler Professional

Experiment type: Unidentified

Workflow type: Analysis: Significance Testing and Fold Change

Experiment notes: One independent variable (parameter name) with four parameter values (classifications)

Figure 61 *New Experiment dialog box*

Import and organize your sample data

Importing and organizing your sample data consists of sequential steps that defines the experiment containing your samples (data files), interpretations, and associated entity lists. Up to eleven steps are involved in the MS Experiment Creation Wizard. The steps you use with your experiment depend on your description and data source.

Importing your data and creating your experiment from the features found in this example involves only the steps presented below:

Import your CEF files

Step 1. Select Data Source: Select the data source that generated the molecular features you are using for your experiment.

Step 2. Select Data to Import: Select the molecular feature sample files.

Organize your files

Step 5. Sample Reordering: Organize your samples by selecting and deselecting individual samples and reordering the selection to group the samples based on the independent variables.

Step 6. Experiment Grouping: Define the sample grouping with respect to your independent variables, including the replicate structure of your experiment.

Filter and align features

Step 7. Filtering: Filter the molecular features by abundance, mass range, number of ions per feature, and charge state.

Step 8. Alignment: Align the features across the samples based on tolerances established by retention time and mass. This step is omitted when the experiment type is “identified” because identified compounds are treated as aligned by identification.

Review features and samples

Step 9. Sample Summary: Display a mass versus retention time plot, spreadsheet, and compound frequency for the distribution of aligned and unaligned entities in the samples. Compound Frequency charts provide a quick view into the effectiveness of the alignment of unidentified experiment types. The **Back** and **Next** buttons in the wizard let you easily review the effects of different alignment and filter options.

Normalize features

Step 10. Normalization Criteria: Scale the signal intensity of sample features to a value calculated by the specified algorithm or an external scalar.

Step 11. Baselining Options: Compare the signal intensity of each sample to a representative value calculated across all of the samples or the control samples.

Import the data files into the experiment

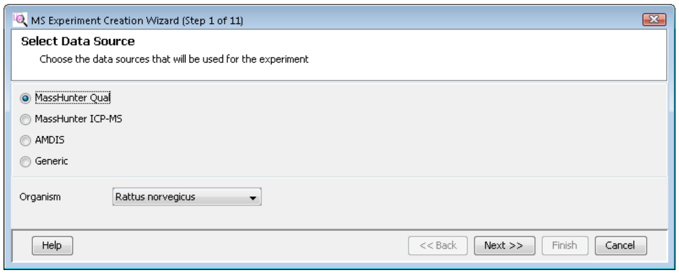
Your data files are imported in to Mass Profiler Professional during Step 1 and Step 2 of the MS Experiment Creation Wizard.

1. Select the data source in the **MS Experiment Creation Wizard (Step 1 of 11)**.

a Click **MassHunter Qual**.

b Select the **Organism** represented by your samples. Selection of an organism is important if you plan to use pathways.

c Click **Next**.



MS Experiment Creation Wizard (Step 1 of 11)

Select Data Source

Choose the data sources that will be used for the experiment

● MassHunter Qual
○ MassHunter ICP-MS
○ AMDIS
○ Generic

Organism: Rattus norvegicus

Figure 62 *MS Experiment Creation Wizard Step 1*

2. Select the sample data to import in the **MS Experiment Creation Wizard (Step 2 of 11)**.

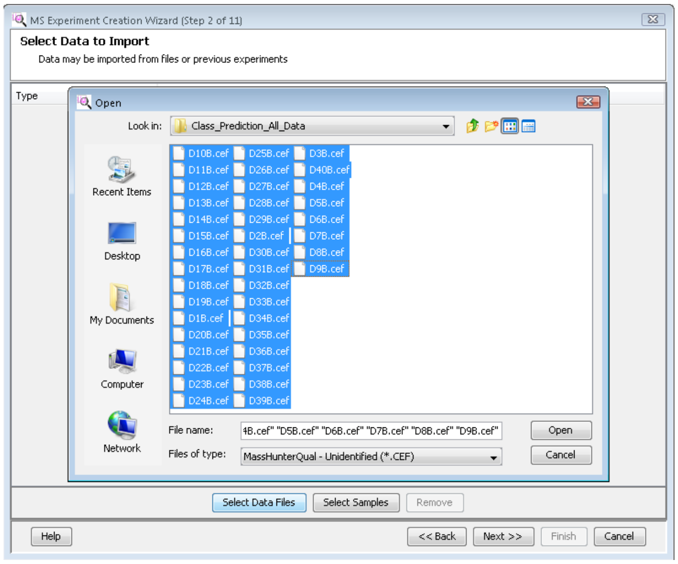
a Click **Select Data Files**.

b Select all of the class prediction data files.

Note: Orderly naming of the data files with respect to the parameters related to the independent variables helps you make sure that all of the data is selected.

c Click **Open**.

d Click **Next**.



MS Experiment Creation Wizard (Step 2 of 11)

Select Data to Import

Data may be imported from files or previous experiments

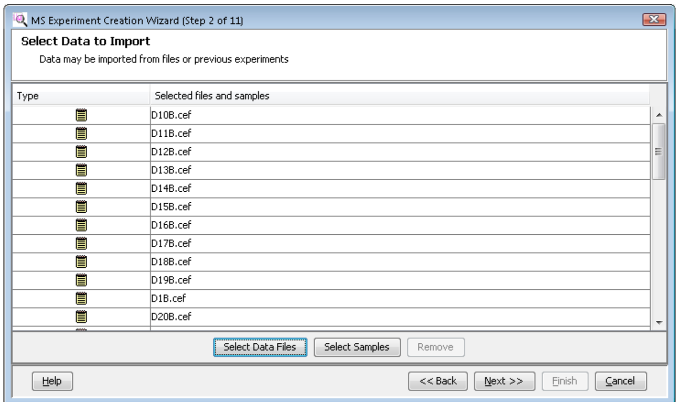
Open

Look in: Class_Prediction_All_Data

File name: 4B.cef" "D5B.cef" "D6B.cef" "D7B.cef" "D8B.cef" "D9B.cef"

Files of type: MassHunterQual - Unidentified (*.CEF)

Figure 63 *Selection of the sample CEF files from MassHunter Qualitative Analysis*



MS Experiment Creation Wizard (Step 2 of 11)

Select Data to Import

Data may be imported from files or previous experiments

Type	Selected files and samples
▤	D10B.cef
▤	D11B.cef
▤	D12B.cef
▤	D13B.cef
▤	D14B.cef
▤	D15B.cef
▤	D16B.cef
▤	D17B.cef
▤	D18B.cef
▤	D19B.cef
▤	D1B.cef
▤	D20B.cef

Figure 64 *Selected sample files in the MS Experiment Creation Wizard Step 2 are arranged in alphabetical order. This order is not necessarily the experimental order.*

Order and group the sample data files

When the data files are imported into your experiment, they are ordered alphabetically. Depending on your sample naming, the files may or may not be ordered in a fashion that simplifies assigning your experimental grouping in the next step of the MS Experiment Creation Wizard. Your sample data files are ordered in this step of the MS Experiment Creation Wizard.

1. Review and order the selected files that are imported in the **MS Experiment Creation Wizard (Step 5 of 11)**.

a Click one or more samples that you want to reorder. Selected sample rows are highlighted as shown in Figure 65 on page 74.

In this example reordering the sample files to restore the numerical order from that shown in Figure 64 to the order shown in Figure 65 on page 74 reduces the potential for make a grouping error in the next step. The numerical order in this example correlates to the parameter values shown in Figure 7 on page 21.

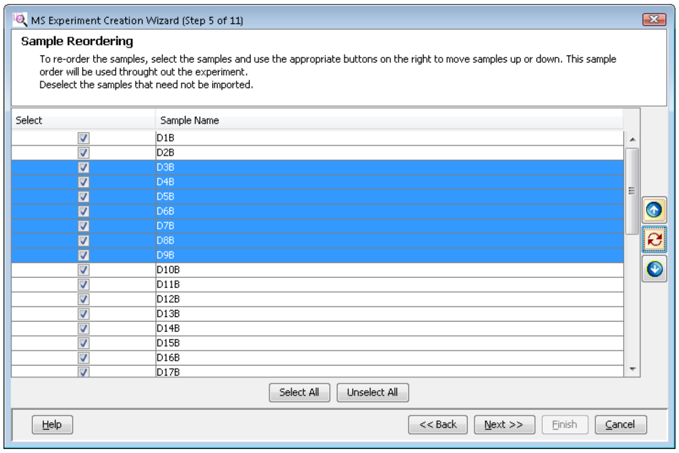
b Click the **Up** ⬆ or **Down** ⬇ buttons to reorder the selected sample or samples.

c Click the **Restore** 🔄 button at any time to return the sample order to your starting point when this step was begun.

d Repeat the reordering steps as often as necessary to obtain your order.

e Mark the **Select** check box in the same row as the **Sample Name** for the samples to import for your analysis, or click **Select All**.

f Click **Next**.



MS Experiment Creation Wizard (Step 5 of 11)

Sample Reordering

To re-order the samples, select the samples and use the appropriate buttons on the right to move samples up or down. This sample order will be used throught out the experiment.
Deselect the samples that need not be imported.

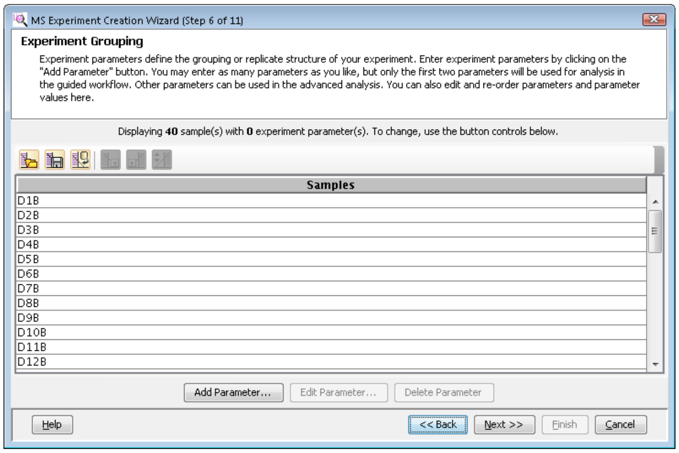
Select	Sample Name
☑	D1B
☑	D2B
☑	D3B
☑	D4B
☑	D5B
☑	D6B
☑	D7B
☑	D8B
☑	D9B
☑	D10B
☑	D11B
☑	D12B
☑	D13B
☑	D14B
☑	D15B
☑	D16B
☑	D17B

__Figure 65__ Selection and reordering of the sample files in the MS Experiment Creation Wizard Step 5

2. Group samples based on the independent variables and replicate structure of your experiment in the **MS Experiment Creation Wizard (Step 6 of 11)**.

Your sample grouping is determined by your experiment definition. An independent variable is referred to as a **parameter name**. The attribute values, or conditions, within an independent variable are referred to as **parameter values**. Samples with the same parameter values within a parameter name are treated as replicates. In order to proceed, at least one parameter with two parameter values must be assigned.

Only the first two parameter names (independent variables) in your experiment are presented in the summary at the conclusion of the MS Experiment Creation Wizard. All parameter names and values entered at this time can be edited during the *Analysis: Significance Testing and Fold Change* workflow and at the conclusion of the workflow by using operations available in the Workflow Browser.



MS Experiment Creation Wizard (Step 6 of 11)

Experiment Grouping

Experiment parameters define the grouping or replicate structure of your experiment. Enter experiment parameters by clicking on the "Add Parameter" button. You may enter as many parameters as you like, but only the first two parameters will be used for analysis in the guided workflow. Other parameters can be used in the advanced analysis. You can also edit and re-order parameters and parameter values here.

Displaying **40** sample(s) with **0** experiment parameter(s). To change, use the button controls below.

Samples
D1B
D2B
D3B
D4B
D5B
D6B
D7B
D8B
D9B
D10B
D11B
D12B

__Figure 66__ Before grouping samples in the MS Experiment Creation Wizard Step 6

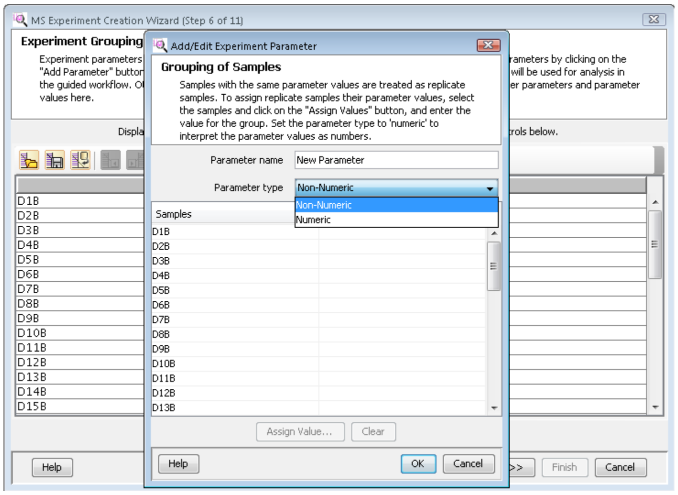
Figure 66 the wizard correctly indicates that 40 samples are displayed in the experiment; ten replicate samples from each of the four conditions.

Assign parameter name and values for the first, or only, independent variable

Note: When entering **Parameter Names** and parameter **Assign Values**, it is very important that the entries use identical letters, numbers, punctuation, and case in order for the Experiment Grouping to function properly. Click **Back** or **Experiment Setup > Experiment Grouping** to return to experiment grouping if an error is identified later in the *Analysis: Significance Testing and Fold Change* workflow or when performing operations available in the Workflow Browser, respectively.

To apply previously saved experiment parameters and parameter values saved in a tab separated value (.tsv) file, click the **Load experiment parameters** button, or the **Import parameters from samples** button, and skip most of the following steps.

a Click **Add Parameter**. The **Add/Edit Experiment parameters** dialog box is opened.



Experiment parameters... "Add Parameter" button... the guided workflow. O... values here.

...rameters by clicking on the ... will be used for analysis in ...er parameters and parameter

Samples with the same parameter values are treated as replicate samples. To assign replicate samples their parameter values, select the samples and click on the "Assign Values" button, and enter the value for the group. Set the parameter type to 'numeric' to interpret the parameter values as numbers.

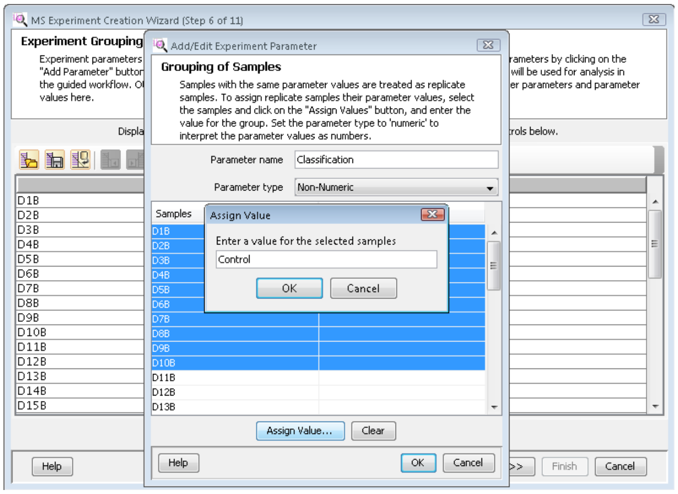
Parameter name: New Parameter

Parameter type: Non-Numeric

Figure 67 *Grouping of samples*

b Type a brief, descriptive name for the independent variable into the **Parameter name**. Type `Classification` for the example experiment.

c Select **Non-Numeric** for the **Parameter type** for your grouping when the grouping is not a quantitative value.



Experiment parameters... "Add Parameter" button... the guided workflow. O... values here.

...rameters by clicking on the ... will be used for analysis in ...er parameters and parameter

Samples with the same parameter values are treated as replicate samples. To assign replicate samples their parameter values, select the samples and click on the "Assign Values" button, and enter the value for the group. Set the parameter type to 'numeric' to interpret the parameter values as numbers.

Parameter name: Classification

Parameter type: Non-Numeric

Enter a value for the selected samples: Control

Figure 68 *Entering a non-numeric parameter value during Experiment Grouping*

d Click the sample rows, while pressing the **Shift** or **Ctrl** key as necessary, to select the samples that are part of the first parameter value of the parameter name.

e Click **Assign Value** after the sample rows have been selected.

f Type `Control` in the **Assign Value** dialog box.

g Click **OK**.

Assign additional parameter values for the first parameter name

a Click the sample rows, while pressing the **Shift** or **Ctrl** key as necessary, to select the samples that are part of the next attribute value within the parameter name.

b Click **Assign Value** after the rows have been selected.

c Type an appropriate description for the parameter value in the **Assign Value** dialog box. The remaining parameters values are `Var AW`, `Var BR`, and `Var CN`.

d Click **OK**.

e Repeat the row selection and assign value process as necessary to complete the assignment of the samples to each of the classifications for the independent variable.

f Click **OK** when all of the attribute values (Assign Value) are assigned to the current independent variable (Parameter Name).

Assign parameter name and values for the second independent variable

The example class prediction sample data set does not have a second independent variable. At this time you can use an additional parameter to divide your entire data set into training and validation data sets for class prediction (as shown in Figure 69 on page 77). Note that when you import the validation data set with your training data set, the features of the validation data set are aligned and normalized with the training data set.

If your experiment and sample data contain a second independent variable, follow the steps below, otherwise skip ahead to “Assign parameter values for the remaining independent variables” on page 77.

a Click **Add Parameter** to begin parameter assignment for the next independent variable.

b Type a brief, descriptive name for the second independent variable into the **Parameter Name**. Type `Class Prediction` for the example experiment to separate the training and validation data sets.

c Select the **Parameter type**. Use **Numeric** when the values are quantitative or reflect a degree of proportionality among the samples with respect to the independent variable.

d Click the sample rows, while pressing the **Shift** or **Ctrl** key as necessary, to select the first attribute of the typed parameter name.

e Click **Assign Value** after the rows have been selected. The recommended parameters values are `Training` and `Validation`.

f Type a descriptive name or value in the **Assign Value** dialog box.

g Click **OK**.

Assign additional parameter values for the second independent variable

a Click the sample rows, while pressing the **Shift** or **Ctrl** key as necessary, to select the next attribute value within the parameter name.

b Click **Assign Value** after the rows have been selected.

c Type a descriptive name or value in the **Assign Value** dialog box.

d Click **OK**.

e Repeat the row selection and assign value process as necessary to complete the assignment of the samples to each of the attribute values for the independent variable.

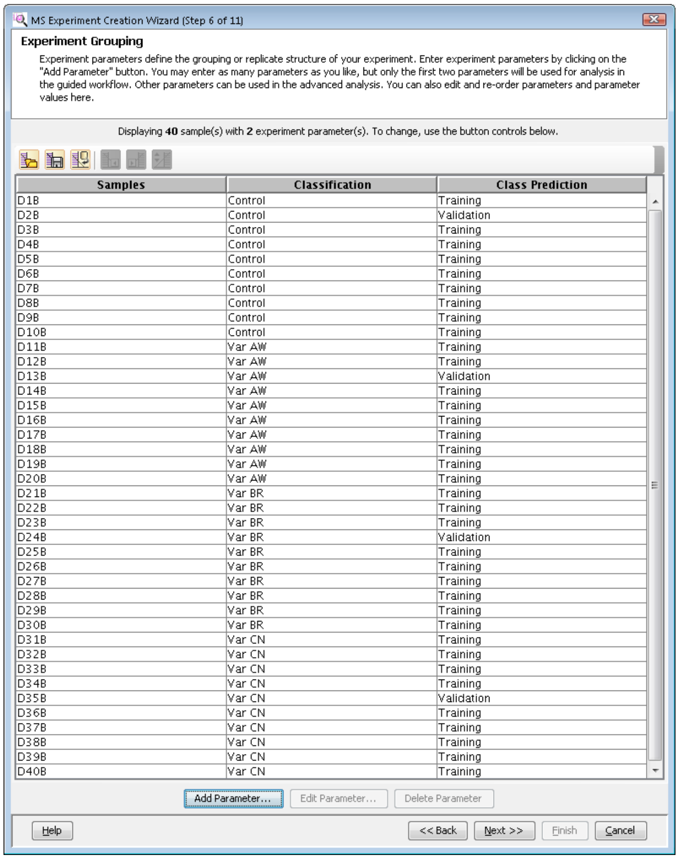
f Click **OK** when all of the attribute values (Assign Value) are assigned to the current independent variable (Parameter Name).

Assign parameter values for the remaining independent variables

a Repeat "Assign parameter name and values for the second independent variable" through "Assign additional parameter values for the second independent variable" as often as necessary to assign all of the independent variable parameter names and to assign their concomitant attribute values.

b To save your experiment parameters and parameter values to a .tsv file, click the **Save experiment parameters** button.

c Click **Next**.



Experiment Grouping

Experiment parameters define the grouping or replicate structure of your experiment. Enter experiment parameters by clicking on the "Add Parameter" button. You may enter as many parameters as you like, but only the first two parameters will be used for analysis in the guided workflow. Other parameters can be used in the advanced analysis. You can also edit and re-order parameters and parameter values here.

Displaying **40** sample(s) with **2** experiment parameter(s). To change, use the button controls below.

Samples	Classification	Class Prediction
D1B	Control	Training
D2B	Control	Validation
D3B	Control	Training
D4B	Control	Training
D5B	Control	Training
D6B	Control	Training
D7B	Control	Training
D8B	Control	Training
D9B	Control	Training
D10B	Control	Training
D11B	Var AW	Training
D12B	Var AW	Training
D13B	Var AW	Validation
D14B	Var AW	Training
D15B	Var AW	Training
D16B	Var AW	Training
D17B	Var AW	Training
D18B	Var AW	Training
D19B	Var AW	Training
D20B	Var AW	Training
D21B	Var BR	Training
D22B	Var BR	Training
D23B	Var BR	Training
D24B	Var BR	Validation
D25B	Var BR	Training
D26B	Var BR	Training
D27B	Var BR	Training
D28B	Var BR	Training
D29B	Var BR	Training
D30B	Var BR	Training
D31B	Var CN	Training
D32B	Var CN	Training
D33B	Var CN	Training
D34B	Var CN	Training
D35B	Var CN	Validation
D36B	Var CN	Training
D37B	Var CN	Training
D38B	Var CN	Training
D39B	Var CN	Training
D40B	Var CN	Training

Figure 69 *Assigned parameter values for the entire sample data set after experiment grouping is competed*

Filter, align, and normalize the sample data

You filter, align, and normalize your sample data in Steps 7 through 11 of the MS Experiment Creation Wizard. At each step of the process, you can view your progress and return to prior steps to adjust your results.

1. Select and enter the data filter parameters in the **MS Experiment Creation Wizard (Step 7 of 11)**.

Filtering during the data import process may be used to reject low-intensity data or restrict the range of data. After data is imported, several filtering options may be applied: Abundance, Retention Time, Mass, Flags, Number of ions, Mass and Minimum Quality Score. Since filtering works with both GC/MS and LC/MS data, the term **Abundance** actually refers to volume for MFE generated CEF files, and the term **Abundance** actually refers to area for FbF generated CEF files. The parameters may be cleared to preserve prior filtering that was used to generate the CEF file.

a Mark the **Minimum absolute abundance** check box and type a value of `5000` counts.

b Clear the **Limit to the largest** check box. An arbitrary limit for metabolomics analyses is not recommended.

c Clear the **Minimum relative abundance** check box under the Abundance filtering group heading.

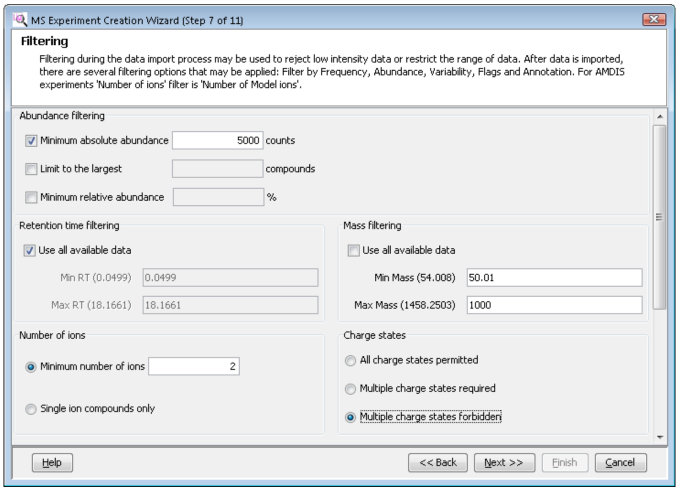
d Mark the **Use all available data** check box.

e Clear the **Use all available data** check box and type `50.01` for the **Min Mass** and `1000` for the **Max Mass**. Filtering by maximum mass improves the statistical analysis by rejecting masses that are not significant to the experiment.

f Click **Minimum number of ions** and type `2`. The mass filter does not need to include reference ions.

g Click **Multiple charge states forbidden**. The example data are from a metabolomics analyses and involve singly charged ions.

h Click **Next**.



MS Experiment Creation Wizard (Step 7 of 11)

Filtering

Filtering during the data import process may be used to reject low intensity data or restrict the range of data. After data is imported, there are several filtering options that may be applied: Filter by Frequency, Abundance, Variability, Flags and Annotation. For AMDIS experiments 'Number of ions' filter is 'Number of Model ions'.

Abundance filtering

☑ Minimum absolute abundance 5000 counts

□ Limit to the largest compounds

□ Minimum relative abundance %

Retention time filtering

☑ Use all available data

Min RT (0.0499) 0.0499

Max RT (18.1661) 18.1661

Mass filtering

□ Use all available data

Min Mass (54.008) 50.01

Max Mass (1458.2503) 1000

Number of ions

● Minimum number of ions 2

○ Single ion compounds only

Charge states

○ All charge states permitted

○ Multiple charge states required

● Multiple charge states forbidden

Figure 70 Recommended filtering parameters in the MS Experiment Creation Wizard Step 7

2. Select and enter the retention time and mass alignment parameters in the **MS Experiment Creation Wizard (Step 8 of 11)**.

Compounds from different samples are aligned or grouped together if their retention times are within the specified tolerance window, and the mass spectral similarity as determined by a simple dot product calculation is above the specified level. Retention alignment rewrites the retention times in the data file so that your input or algorithmically selected features are used to correct the retention times.

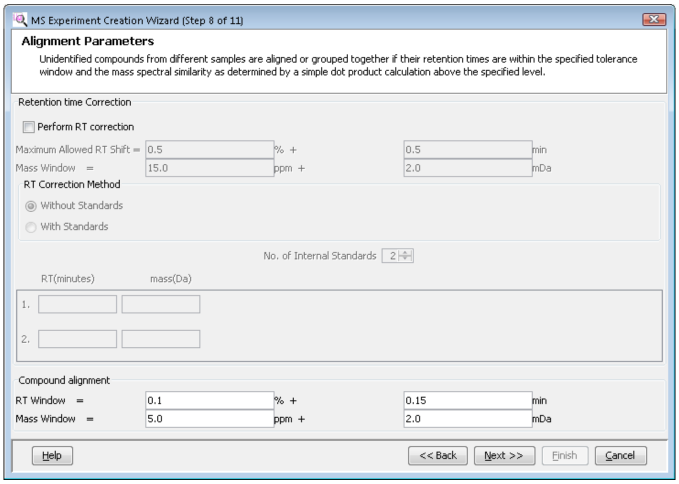
a Clear the **Perform RT correction** check box. A larger retention time shift may be used to compensate for less than ideal chromatography.

If retention time correction is used, it is recommended to perform retention time correction with at least two widely spaced standards, and the standards must be present in every sample. With standards the correction is based on a piecewise linear fit.

b Type `0.1` % and `0.15` min for **RT Window** under Compound alignment. Smaller values result in reduced compound grouping among the samples leading to a larger list of unique compounds in the experiment.

c Type `5.0` ppm and `2.0` mDa for **Mass Window**. A mass window less than 2.0 mDa for higher masses is not recommended.

d Click **Next**.



MS Experiment Creation Wizard (Step 8 of 11)

Alignment Parameters

Unidentified compounds from different samples are aligned or grouped together if their retention times are within the specified tolerance window and the mass spectral similarity as determined by a simple dot product calculation above the specified level.

Retention time Correction

□ Perform RT correction

Maximum Allowed RT Shift = 0.5 % + 0.5 min

Mass Window = 15.0 ppm + 2.0 mDa

RT Correction Method

● Without Standards

○ With Standards

No. of Internal Standards 2

RT(minutes)	mass(Da)
1.	
2.	

Compound alignment

RT Window = 0.1 % + 0.15 min

Mass Window = 5.0 ppm + 2.0 mDa

Figure 71 *Recommended alignment parameters in the MS Experiment Creation Wizard Step 8*

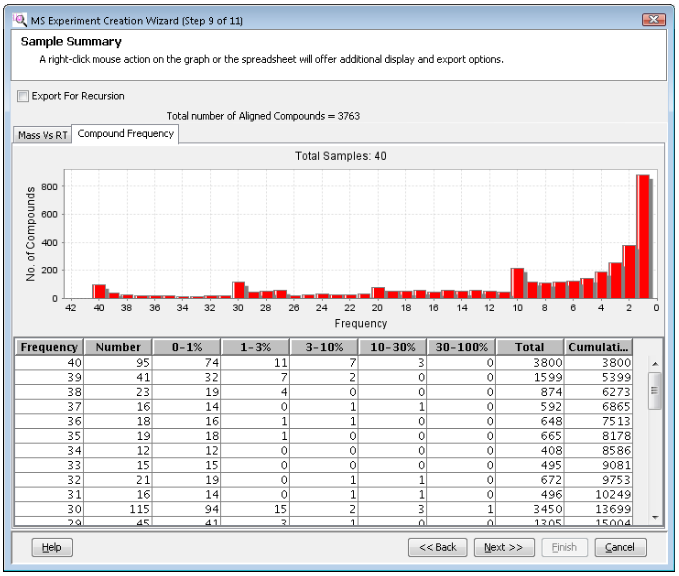
3. View and review the compounds present and absent in each sample in the **MS Experiment Creation Wizard (Step 9 of 11)**.

This step presents a summary of the compounds present and absent in each of the samples based on the experiment parameters including the application of the filter and alignment parameters.

Note: Click **Back** to make changes in the **Filtering (Step 7 of 11)** page and the **Alignment Parameters (Step 8 of 11)** page parameters, and then return to this **Sample Summary (Step 9 of 11)** page several times to develop a feel for how each of the parameters affects the compound summary.

a Clear the **Export for Recursion** check box. Exporting the compounds for recursion at this step in the class prediction workflow is not recommended. Better results are obtained after the data has been filtered for significance in the following steps.

b Click **Next**.



Sample Summary

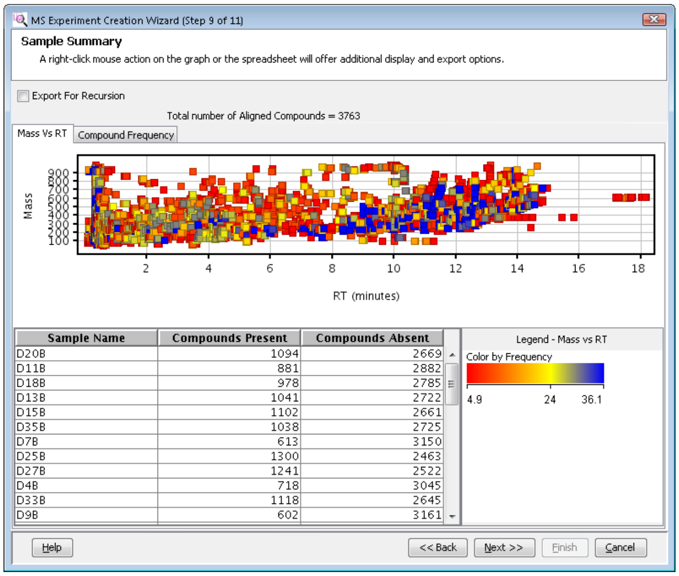
A right-click mouse action on the graph or the spreadsheet will offer additional display and export options.

□ Export For Recursion

Total number of Aligned Compounds = 3763

Frequency	Number	0-1%	1-3%	3-10%	10-30%	30-100%	Total	Cumulati...
40	95	74	11	7	3	0	3800	3800
39	41	32	7	2	0	0	1599	5399
38	23	19	4	0	0	0	874	6273
37	16	14	0	1	1	0	592	6865
36	18	16	1	1	0	0	648	7513
35	19	18	1	0	0	0	665	8178
34	12	12	0	0	0	0	408	8586
33	15	15	0	0	0	0	495	9081
32	21	19	0	1	1	0	672	9753
31	16	14	0	1	1	0	496	10249
30	115	94	15	2	3	1	3450	13699
29	45	41	3	1	0	0	1305	15004

Figure 72 *Compound frequency view of the sample summary page in the MS Experiment Creation Wizard Step 9 - Untargeted feature finding*



Sample Summary

A right-click mouse action on the graph or the spreadsheet will offer additional display and export options.

□ Export For Recursion

Total number of Aligned Compounds = 3763

Sample Name	Compounds Present	Compounds Absent
D20B	1094	2669
D11B	881	2882
D18B	978	2785
D13B	1041	2722
D15B	1102	2661
D35B	1038	2725
D7B	613	3150
D25B	1300	2463
D27B	1241	2522
D4B	718	3045
D33B	1118	2645
D9B	602	3161

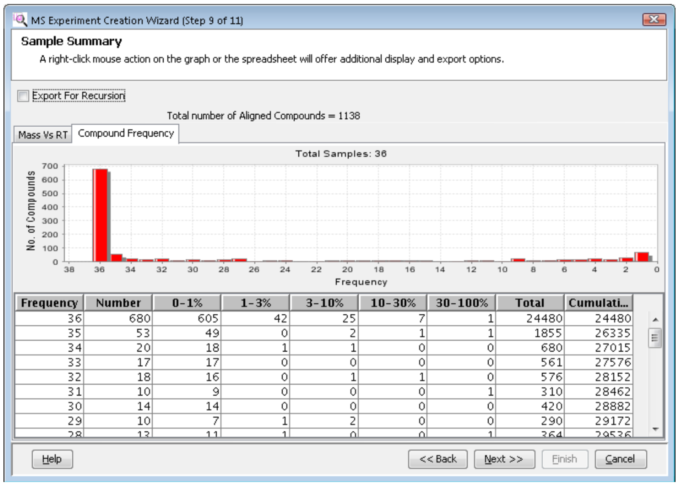
Figure 73 *Mass versus retention time view of the sample summary page in the MS Experiment Creation Wizard Step 9 - Untargeted feature finding*

Replicates should have similar numbers of compounds present and absent. You can see this easily if the files have a systematic naming system that sorts the replicates together by file name. In this example the sample files names are mixed but the similarity in compounds present and absent is still visible.

Sample summary with recursive features

When the experiment is created using recursive features (targeted feature finding) the compound frequency (Figure 74 on page 81) shows a high frequency of compounds present in all of the samples compared to when the features were untargeted (Figure 72 on page 80).

With a recursive, targeted feature finding, data set the compound frequency and mass versus retention time views of the 36 training sample data files (Figure 74 and Figure 75) show that all of the training samples have a similar number of compounds present, compared to before recursion when just replicates within a classification showed similarities (Figure 73 and Figure 74 on page 81).



Sample Summary

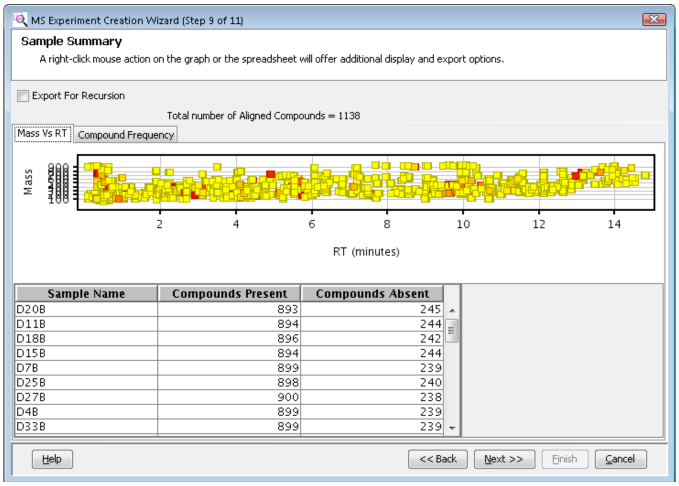
A right-click mouse action on the graph or the spreadsheet will offer additional display and export options.

□ Export For Recursion

Total number of Aligned Compounds = 1138

Frequency	Number	0-1%	1-3%	3-10%	10-30%	30-100%	Total	Cumulati...
36	680	605	42	25	7	1	24480	24480
35	53	49	0	2	1	1	1855	26335
34	20	18	1	1	0	0	680	27015
33	17	17	0	0	0	0	561	27576
32	18	16	0	1	1	0	576	28152
31	10	9	0	0	0	1	310	28462
30	14	14	0	0	0	0	420	28882
29	10	7	1	2	0	0	290	29172
28	13	11	1	0	0	1	364	29536

Figure 74 *Compound frequency view of the sample summary page in the MS Experiment Creation Wizard Step 9 - Targeted feature finding (recursive)*



Sample Summary

A right-click mouse action on the graph or the spreadsheet will offer additional display and export options.

□ Export For Recursion

Total number of Aligned Compounds = 1138

Sample Name	Compounds Present	Compounds Absent
D20B	893	245
D11B	894	244
D18B	896	242
D15B	894	244
D7B	899	239
D25B	898	240
D27B	900	238
D4B	899	239
D33B	899	239

Figure 75 *Mass versus retention time view of the sample summary page in the MS Experiment Creation Wizard Step 9 - Targeted feature finding (recursive)*

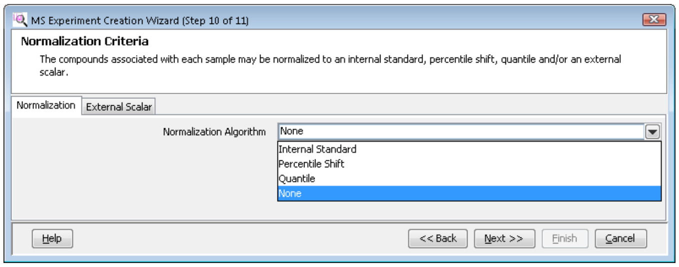
4. Select whether to normalize the data in the **MS Experiment Creation Wizard (Step 10 of 11)**.

Normalizing the data reduces the variability caused by sample preparation and instrument response. From the list of compounds present in all of the samples you may pick one as an internal standard. No internal or external standard is selected at this time.

a Select **None** for the **Normalization Algorithm** in the Normalization tab.

b Clear the **Use External Scalar** check box on the External Scalar tab.

c Click **Next**.



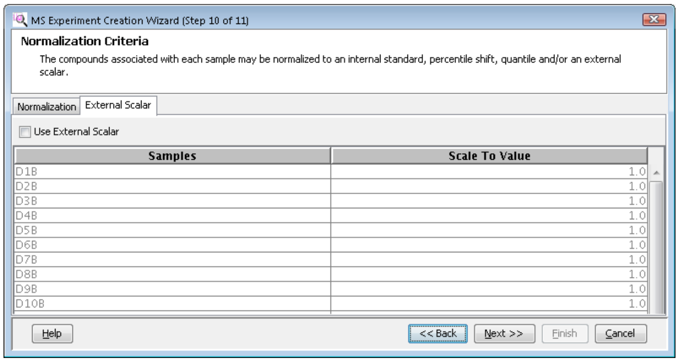
Normalization Criteria

The compounds associated with each sample may be normalized to an internal standard, percentile shift, quantile and/or an external scalar.

Normalization Algorithm: None

- Internal Standard
- Percentile Shift
- Quantile
- None

Figure 76 *Normalization tab in the MS Experiment Creation Wizard Step 10*



Normalization Criteria

The compounds associated with each sample may be normalized to an internal standard, percentile shift, quantile and/or an external scalar.

□ Use External Scalar

Samples	Scale To Value
D1B	1.0
D2B	1.0
D3B	1.0
D4B	1.0
D5B	1.0
D6B	1.0
D7B	1.0
D8B	1.0
D9B	1.0
D10B	1.0

Figure 77 *External Scalar tab in the MS Experiment Creation Wizard Step 10*

5. Select whether to compare features in each sample to the response of the features across multiple samples in the **MS Experiment Creation Wizard (Step 11 of 11)**.

Baselining is a technique used to view and compare data. It involves converting the original data values to values that are expressed as changes in the data values relative to a calculated statistical value derived from the data. The calculated statistical value is referred to as the baseline.

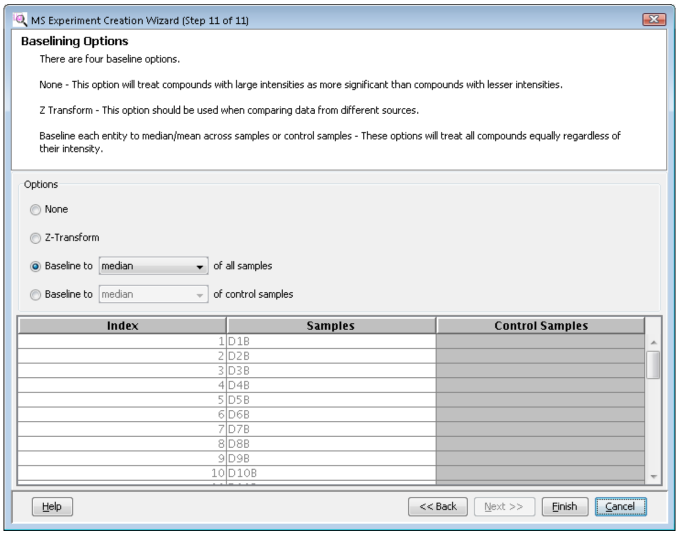
Four baselining options are available:

1. **None**: Recommended if only a few features in the samples exist.
2. **Z-Transform**: Recommended if the data sets are very dense, data where very few instances of compounds are absent from any sample, such as a quantitation data set from recursion.
3. **Baseline to ____ of all samples**: The abundance for each compound is normalized to its selected statistical abundance (median or mean) across all of the samples. This has the effect of reducing the weight of very large and very small compound features on later statistical analyses.
4. **Baseline to ____ of control samples**: The abundance for each compound is normalized to its selected statistical abundance (median or mean) across just the samples selected as the control samples. This has the effect of weighting the compound features to a known value that is considered to be normal in the population while reducing the effect of large and small compound features.

a Click **Baseline to ____ of all samples**.

b Select **median** for the **Baseline to ____of all samples**.

c Click **Finish**.



Baselining Options

There are four baseline options.

None - This option will treat compounds with large intensities as more significant than compounds with lesser intensities.

Z Transform - This option should be used when comparing data from different sources.

Baseline each entity to median/mean across samples or control samples - These options will treat all compounds equally regardless of their intensity.

Options

○ None

○ Z-Transform

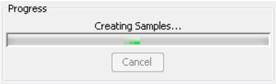
● Baseline to median of all samples

○ Baseline to median of control samples

Index	Samples	Control Samples
1	D1B	
2	D2B	
3	D3B	
4	D4B	
5	D5B	
6	D6B	
7	D7B	
8	D8B	
9	D9B	
10	D10B	

Figure 78 *Selecting baselining options in the MS Experiment Creation Wizard Step 11*

A **Progress** dialog box (Figure 79) is displayed while your samples are processed and prepared for performing your initial differential analysis.



Creating Samples...

Figure 79 *Selecting baselining options in the MS Experiment Creation Wizard Step 11*

Note: The *Analysis: Significance Testing and Fold Change Wizard* immediately starts if you selected **Analysis: Significance Testing and Fold Change** for the **Workflow type** in the **New Experiment** dialog box.

Perform a differential analysis

The *Analysis: Significance Testing and Fold Change* workflow helps you create an initial differential expression from your data and identify the most significant features from among all of the features previously found using molecular feature extraction. The steps necessary to create your initial differential expression are predetermined and based on the experiment type, experiment grouping, and conditions you entered when creating your project and setting up your experiment. The Significance Testing and Fold Change workflow does not start if **Data Import Wizard** was selected as the **Workflow type** in the **New Experiment** dialog box (Figure 61 on page 70).

Differential analysis steps

The workflow displays the sequence of steps on the left-hand side navigator with the current step highlighted (see Figure 80 on page 85). Some steps may be automatically skipped for your experiment. All of the parameters can be edited at the conclusion of the *Analysis: Significance Testing and Fold Change* workflow by using the operations available in the Workflow Browser (see Figure 93 on page 97).

Step 1. Summary Report: Displays a summary view of your experiment based on the parameters you provided in the Import Data wizard. A profile plot with the samples on the x-axis and the log normalized abundance values on the y-axis is displayed. If the number of samples is more than 30, the data is represented by a spreadsheet view instead of a profile plot.

Step 2. Experiment Grouping: Independent variables and the attribute values of the independent variables must be specified to define grouping of the samples. An independent variable is referred to as a parameter name. The attribute values within an independent variable are referred to as parameter values. Samples with the same parameter values within a parameter name are treated as replicates.

Step 3. Filter Flags: The compounds created during the experiment creation are now referred to as entities. The entities are filtered (removed) from further analysis based on their presence across samples and parameter values (now referred to as a condition).

Step 4. Filter by Frequency: Entities are further filtered based on their frequency of presence in specified samples and conditions. This filter removes irreproducible entities.

Step 5. Quality Control on Samples: The samples are presented by grouping and the current Principal Component Analysis (PCA). PCA calculates all the possible principal components and visually represents them in a 3D scatter plot. The scores shown by the axes scales are used to check data quality. The scatter plot shows one point per sample colored-coded by the experiment grouping. Replicates within a group should cluster together and be separated from samples in other groups

Step 6. Significance Analysis: The entities are filtered based on their p-values calculated from a statistical analysis. The statistical analysis performed depends on the samples and experiment grouping.

Step 7. Fold Change: Compounds are further filtered based on their abundance ratios or differences between a treatment and a control that are greater than a specified cut-off or threshold value.

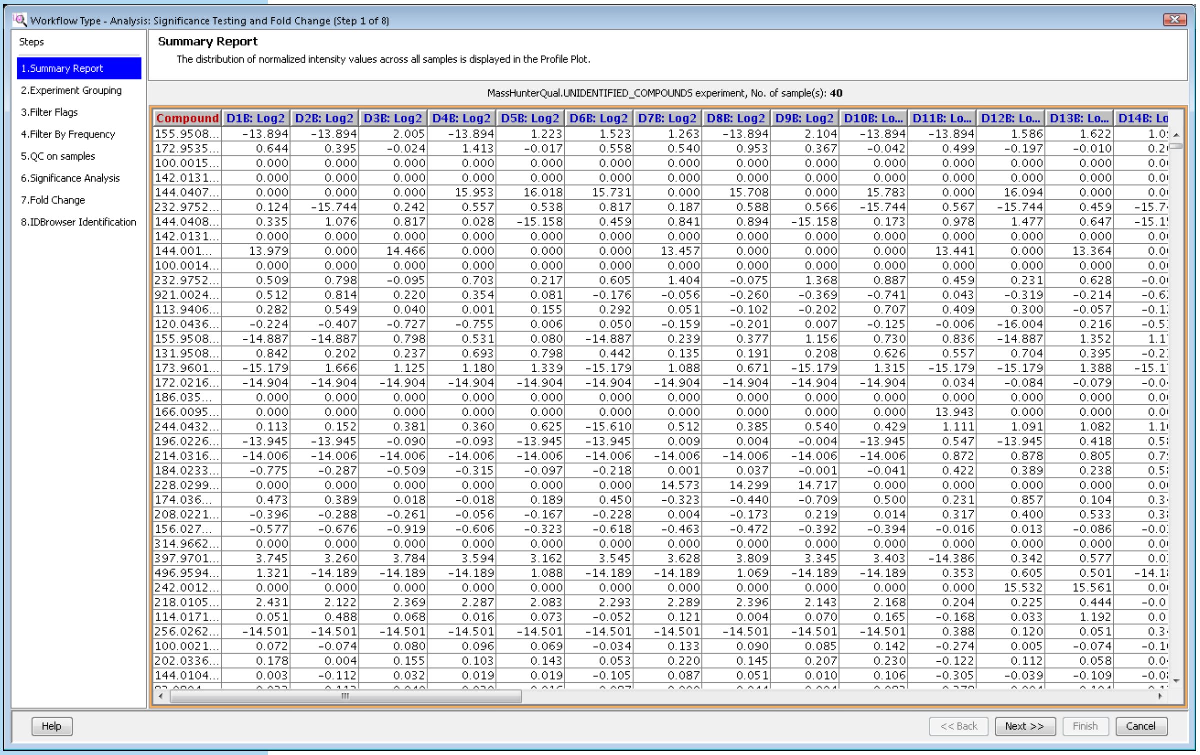
Step 8. ID Browser Identification: The final entity list is directly imported into ID Browser for identification and returned to Mass Profiler Professional.

Feature selection for recursion

If your main objective for this initial differential analysis is to export the significant features identified in your data so that they can be used as targeted features to improve your feature finding, it is recommended to process the features through at least "Enter the parameters for Filter Flags in the Analysis: Significance Testing and Fold Change (Step 3 of 8) workflow." on page 87. The Filter Flags step is used to require that a feature must be present in at least two samples, which removes "one-hit wonder" features and helps recursive finding in MassHunter Qualitative Analysis run efficiently. A "one-hit wonder" is an entity that appears in only one sample, is absent from the replicate samples, and does not provide any utility for statistical analysis.

1. Inspect the entities in the **Analysis: Significance Testing and Fold Change (Step 1 of 8)** workflow.

Double-click and right-click to enable the actions available on the spreadsheet, or profile plot, to inspect an entity, to change the plot view, to export selected data, or to export the plot to a file. For the example data set, the summary report is displayed in a spreadsheet as shown in Figure 80.



Workflow Type - Analysis: Significance Testing and Fold Change (Step 1 of 8)

Steps
1. Summary Report
2. Experiment Grouping
3. Filter Flags
4. Filter By Frequency
5. QC on samples
6. Significance Analysis
7. Fold Change
8. IDBrowser Identification

Summary Report

The distribution of normalized intensity values across all samples is displayed in the Profile Plot.

MassHunterQual.UNIDENTIFIED_COMPOUNDS experiment, No. of sample(s): **40**

Compound	D1B: Log2	D2B: Log2	D3B: Log2	D4B: Log2	D5B: Log2	D6B: Log2	D7B: Log2	D8B: Log2	D9B: Log2	D10B: Lo...	D11B: Lo...	D12B: Lo...	D13B: Lo...	D14B: Lo
155.9508...	-13.894	-13.894	2.005	-13.894	1.223	1.523	1.263	-13.894	2.104	-13.894	-13.894	1.586	1.622	1.0
172.9535...	0.644	0.395	-0.024	1.413	-0.017	0.558	0.540	0.953	0.367	-0.042	0.499	-0.197	-0.010	0.2
100.0015...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
142.0131...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
144.0407...	0.000	0.000	0.000	15.953	16.018	15.731	0.000	15.708	0.000	15.783	0.000	16.094	0.000	0.0
232.9752...	0.124	-15.744	0.242	0.557	0.538	0.817	0.187	0.588	0.566	-15.744	0.567	-15.744	0.459	-15.7
144.0408...	0.335	1.076	0.817	0.028	-15.158	0.459	0.841	0.894	-15.158	0.173	0.978	1.477	0.647	-15.1
142.0131...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
144.001...	13.979	0.000	14.466	0.000	0.000	0.000	13.457	0.000	0.000	0.000	13.441	0.000	13.364	0.0
100.0014...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
232.9752...	0.509	0.798	-0.095	0.703	0.217	0.605	1.404	-0.075	1.368	0.887	0.459	0.231	0.628	-0.0
921.0024...	0.512	0.814	0.220	0.354	0.081	-0.176	-0.056	-0.260	-0.369	-0.741	0.043	-0.319	-0.214	-0.6
113.9406...	0.282	0.549	0.040	0.001	0.155	0.292	0.051	-0.102	-0.202	0.707	0.409	0.300	-0.057	-0.1
120.0436...	-0.224	-0.407	-0.727	-0.755	0.006	0.050	-0.159	-0.201	0.007	-0.125	-0.006	-16.004	0.216	-0.5
155.9508...	-14.887	-14.887	0.798	0.531	0.080	-14.887	0.239	0.377	1.156	0.730	0.836	-14.887	1.352	1.1
131.9508...	0.842	0.202	0.237	0.693	0.798	0.442	0.135	0.191	0.208	0.626	0.557	0.704	0.395	-0.2
173.9601...	-15.179	1.666	1.125	1.180	1.339	-15.179	1.088	0.671	-15.179	1.315	-15.179	-15.179	1.388	-15.1
172.0216...	-14.904	-14.904	-14.904	-14.904	-14.904	-14.904	-14.904	-14.904	-14.904	-14.904	0.034	-0.084	-0.079	-0.0
186.035...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
166.0095...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	13.943	0.000	0.000	0.0
244.0432...	0.113	0.152	0.381	0.360	0.625	-15.610	0.512	0.385	0.540	0.429	1.111	1.091	1.082	1.1
196.0226...	-13.945	-13.945	-0.090	-0.093	-13.945	-13.945	0.009	0.004	-0.004	-13.945	0.547	-13.945	0.418	0.5
214.0316...	-14.006	-14.006	-14.006	-14.006	-14.006	-14.006	-14.006	-14.006	-14.006	-14.006	0.872	0.878	0.805	0.7
184.0233...	-0.775	-0.287	-0.509	-0.315	-0.097	-0.218	0.001	0.037	-0.001	-0.041	0.422	0.389	0.238	0.5
228.0299...	0.000	0.000	0.000	0.000	0.000	0.000	14.573	14.299	14.717	0.000	0.000	0.000	0.000	0.0
174.036...	0.473	0.389	0.018	-0.018	0.189	0.450	-0.323	-0.440	-0.709	0.500	0.231	0.857	0.104	0.3
208.0221...	-0.396	-0.288	-0.261	-0.056	-0.167	-0.228	0.004	-0.173	0.219	0.014	0.317	0.400	0.533	0.3
156.027...	-0.577	-0.676	-0.919	-0.606	-0.323	-0.618	-0.463	-0.472	-0.392	-0.394	-0.016	0.013	-0.086	-0.0
314.9662...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
397.9701...	3.745	3.260	3.784	3.594	3.162	3.545	3.628	3.809	3.345	3.403	-14.386	0.342	0.577	0.0
496.9594...	1.321	-14.189	-14.189	-14.189	1.088	-14.189	-14.189	1.069	-14.189	-14.189	0.353	0.605	0.501	-14.1
242.0012...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	15.532	15.561	0.0
218.0105...	2.431	2.122	2.369	2.287	2.083	2.293	2.289	2.396	2.143	2.168	0.204	0.225	0.444	-0.0
114.0171...	0.051	0.488	0.068	0.016	0.073	-0.052	0.121	0.004	0.070	0.165	-0.168	0.033	1.192	0.0
256.0262...	-14.501	-14.501	-14.501	-14.501	-14.501	-14.501	-14.501	-14.501	-14.501	-14.501	0.388	0.120	0.051	0.3
100.0021...	0.072	-0.074	0.080	0.096	0.069	-0.034	0.133	0.090	0.085	0.142	-0.274	0.005	-0.074	-0.1
202.0336...	0.178	0.004	0.155	0.103	0.143	0.053	0.220	0.145	0.207	0.230	-0.122	0.112	0.058	0.0
144.0104...	0.003	-0.112	0.032	0.019	0.019	-0.105	0.087	0.051	0.010	0.106	-0.305	-0.039	-0.109	-0.0

Figure 80 *Summary Report spreadsheet the initial features found in the example experiment for 40 sample data files*

d Click **Next**.

The *Analysis: Significance Testing and Fold Change* workflow helps you proceed through each step using the **Next** button. A summary of your analysis is presented in each subsequent step. After review of your analysis progress you may return to any previous step and make changes by using the **Back** button. To become more familiar with the analysis parameters and how the parameters affect your data, it is recommended that you frequently use the **Back** and **Next** buttons. To exit the wizard and skip the later steps in the wizard, click **Finish** at any step. When you click **Finish**, the entity list is saved, and you may commence analysis using the advanced operations available in the **Workflow Browser**.

2. Review the experiment grouping parameters associated with the independent variables and their classification in the **Analysis: Significance Testing and Fold Change (Step 2 of 8)** workflow.

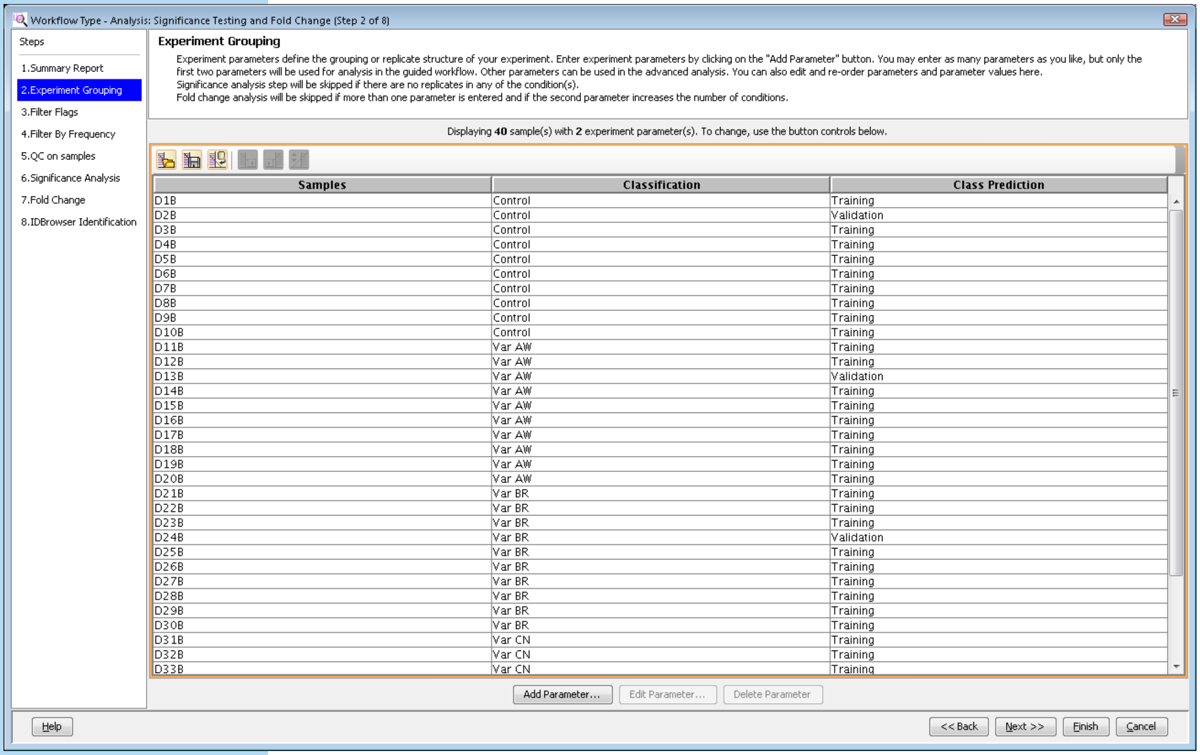
In this step you have an opportunity to edit or change your experiment grouping. An independent variable is referred to as a parameter name. The attribute values, or classifications, within an independent variable are referred to as parameter values. Samples with the same parameter value within a parameter name are treated as replicates.

Only the first two parameter names (independent variables) are used for analysis in the *Analysis: Significance Testing and Fold Change* workflow. All of the parameters are available in the Workflow Browser at the completion of the *Analysis: Significance Testing and Fold Change* workflow.

Note: In order to proceed, at least one parameter name with two parameters values must be assigned.

a Click **Add Parameter**. The **Grouping of Samples** dialog box is opened.

b Edit or change your experiment grouping by following the procedure presented in section "Import and organize your sample data" on page 71 step "Group samples based on the independent variables and replicate structure of your experiment in the MS Experiment Creation Wizard (Step 6 of 11)." on page 74.



Experiment Grouping

Experiment parameters define the grouping or replicate structure of your experiment. Enter experiment parameters by clicking on the "Add Parameter" button. You may enter as many parameters as you like, but only the first two parameters will be used for analysis in the guided workflow. Other parameters can be used in the advanced analysis. You can also edit and re-order parameters and parameter values here.
Significance analysis step will be skipped if there are no replicates in any of the condition(s).
Fold change analysis will be skipped if more than one parameter is entered and if the second parameter increases the number of conditions.

Displaying **40** sample(s) with **2** experiment parameter(s). To change, use the button controls below.

Samples	Classification	Class Prediction
D1B	Control	Training
D2B	Control	Validation
D3B	Control	Training
D4B	Control	Training
D5B	Control	Training
D6B	Control	Training
D7B	Control	Training
D8B	Control	Training
D9B	Control	Training
D10B	Control	Training
D11B	Var AW	Training
D12B	Var AW	Training
D13B	Var AW	Validation
D14B	Var AW	Training
D15B	Var AW	Training
D16B	Var AW	Training
D17B	Var AW	Training
D18B	Var AW	Training
D19B	Var AW	Training
D20B	Var AW	Training
D21B	Var BR	Training
D22B	Var BR	Training
D23B	Var BR	Training
D24B	Var BR	Validation
D25B	Var BR	Training
D26B	Var BR	Training
D27B	Var BR	Training
D28B	Var BR	Training
D29B	Var BR	Training
D30B	Var BR	Training
D31B	Var CN	Training
D32B	Var CN	Training
D33B	Var CN	Training

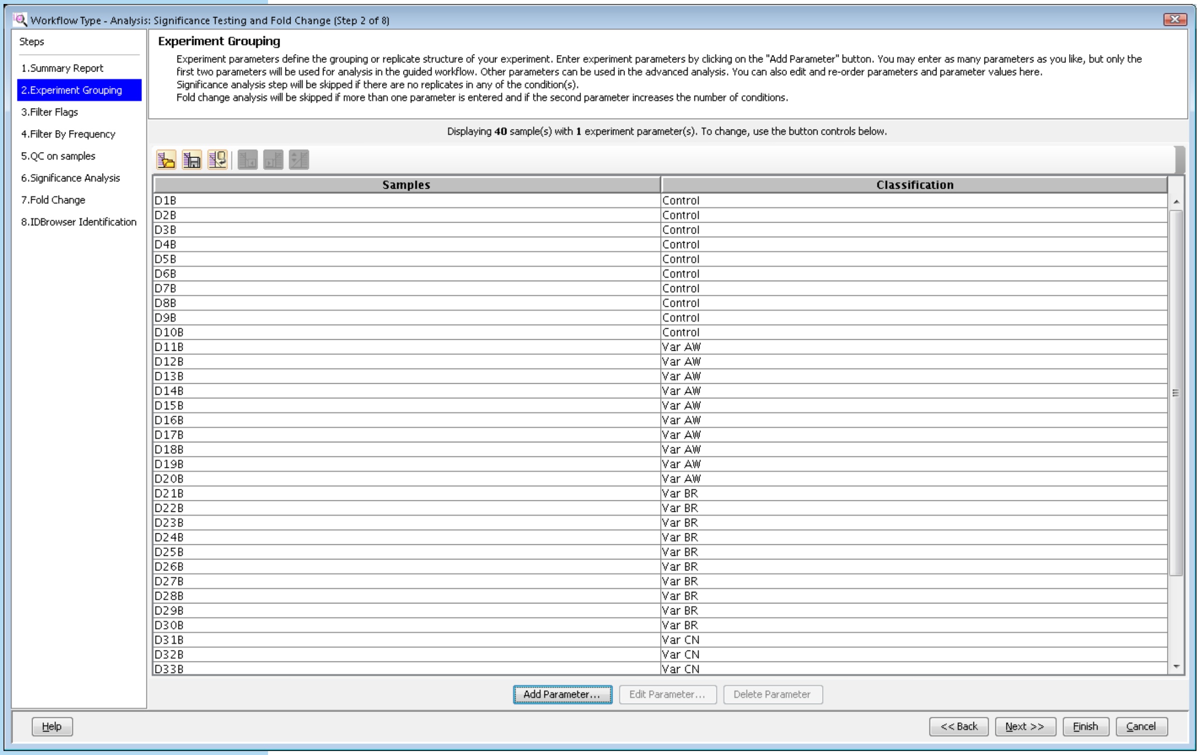
Figure 81 *Experiment grouping for the example experiment with a parameter name used to identify training and validation sample data files*

c Click on any entry in the **Class Prediction** column to select the entire column.

d Click **Delete Parameter** to remove the **Class Prediction** grouping.

The small number of sample files associated with the validation data set in this example prevents the Filter by Frequency filter from operating as desired among the sample classifications; when a grouping has a single sample data file the number of entities that pass the Filter by Frequency criteria is larger than expected. In this example 2030 entities pass when the Class Prediction grouping is retained compared to 907 entities without the Class Prediction grouping.

e Click **Yes** in the **Confirm Delete** dialog box.



Experiment Grouping

Experiment parameters define the grouping or replicate structure of your experiment. Enter experiment parameters by clicking on the "Add Parameter" button. You may enter as many parameters as you like, but only the first two parameters will be used for analysis in the guided workflow. Other parameters can be used in the advanced analysis. You can also edit and re-order parameters and parameter values here.
Significance analysis step will be skipped if there are no replicates in any of the condition(s).
Fold change analysis will be skipped if more than one parameter is entered and if the second parameter increases the number of conditions.

Displaying **40** sample(s) with **1** experiment parameter(s). To change, use the button controls below.

Samples	Classification
D1B	Control
D2B	Control
D3B	Control
D4B	Control
D5B	Control
D6B	Control
D7B	Control
D8B	Control
D9B	Control
D10B	Control
D11B	Var AW
D12B	Var AW
D13B	Var AW
D14B	Var AW
D15B	Var AW
D16B	Var AW
D17B	Var AW
D18B	Var AW
D19B	Var AW
D20B	Var AW
D21B	Var BR
D22B	Var BR
D23B	Var BR
D24B	Var BR
D25B	Var BR
D26B	Var BR
D27B	Var BR
D28B	Var BR
D29B	Var BR
D30B	Var BR
D31B	Var CN
D32B	Var CN
D33B	Var CN

Figure 82 *Experiment grouping for the example experiment after deleting the Class Prediction parameter name shown in Figure 81 on page 86*

f Click **Next**.

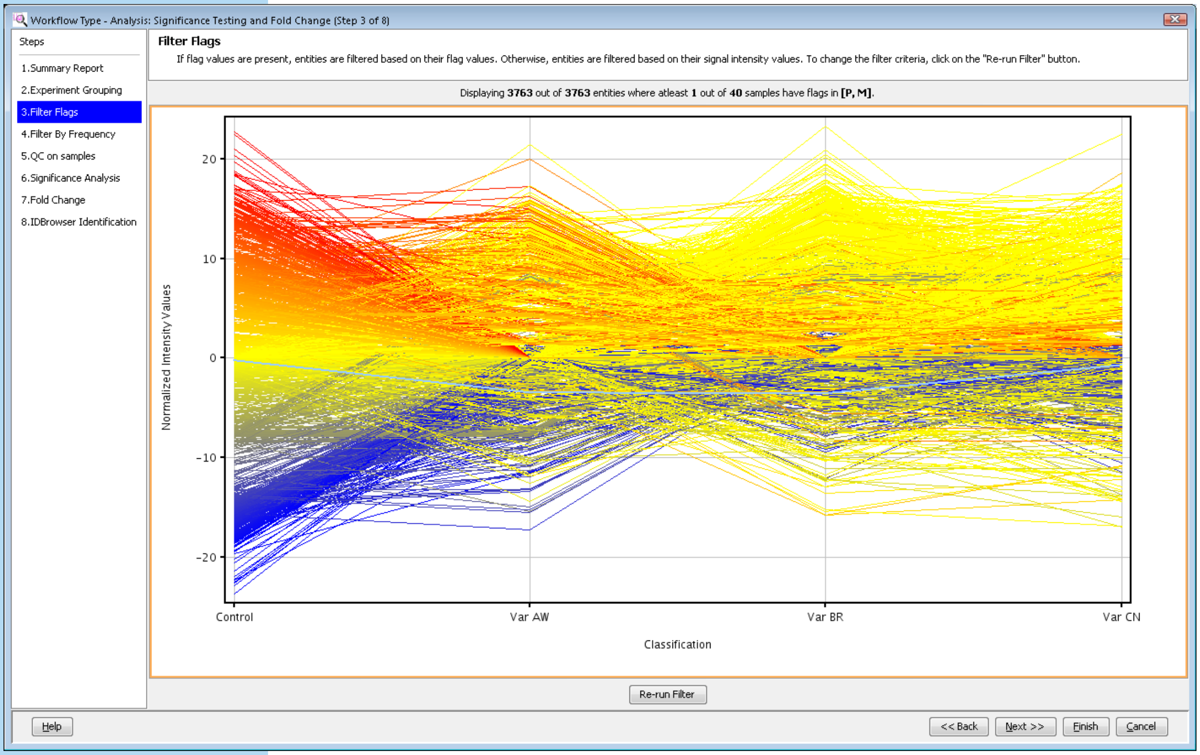
3. Enter the parameters for Filter Flags in the **Analysis: Significance Testing and Fold Change (Step 3 of 8)** workflow.

The entities may now be filtered (removed) from further analysis based on their presence or absence across the samples and classification (now referred to as a condition). A flag is a term used to denote the quality of an entity within a sample. A flag indicates if the entity was detected in each sample as follows:

- **Present** means the entity was detected
- **Absent** means the entity was not detected
- **Marginal** means the signal for the entity was present but saturated

See "Definitions" on page 160 for more definitions and relationships of the terms used by the class prediction workflow.

Note: Before using the **Re-run Filter** button, you can review the entities, change the plot view, export selected data, or export the plot to a file using click, double-click and right-click features available on the plot.



Workflow Type - Analysis: Significance Testing and Fold Change (Step 3 of 8)

Filter Flags

If flag values are present, entities are filtered based on their flag values. Otherwise, entities are filtered based on their signal intensity values. To change the filter criteria, click on the "Re-run Filter" button.

Displaying **3763** out of **3763** entities where atleast **1** out of **40** samples have flags in **[P, M]**.

Figure 83 *Profile plot of the entities showing the four classifications of the example experiment. The number of entities meeting, or passing, the filter parameters is shown along the top of the profile plot - 3763 entities.*

A major objective of Filter Flags is to remove "one-hit wonders" from further consideration. A "one-hit wonder" is an entity that appears in only one sample, is absent from the replicate samples, and does not provide any utility for statistical analysis.

a Click **Re-run Filter**.

b Mark the **Present** check box.

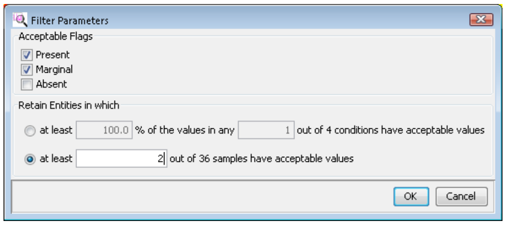
c Mark the **Marginal** check box.

d Clear the **Absent** check box. This flag is useful when you want to identify entities that are missing in the samples. You can use this flag in conjunction with the **Next** and **Back** buttons to review the entities that are missing in some samples.

e Click **at least ___ out of X samples have acceptable values**. The value "X" is replaced in your display with the total number of samples in your data set.

f Type 2 in the entry box. By setting this parameter to a value of two or more, one-hit wonders are filtered.

g Click **OK**.



Filter Parameters

Acceptable Flags

☑ Present

☑ Marginal

□ Absent

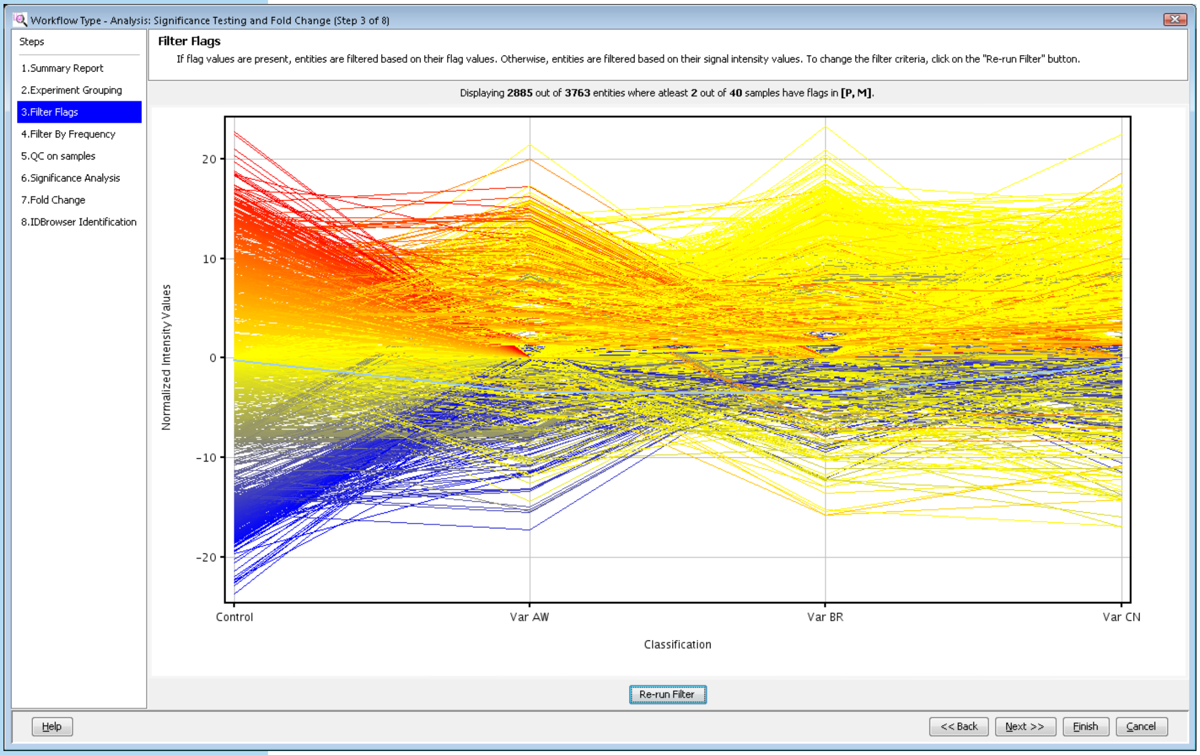
Retain Entities in which

○ at least 100.0 % of the values in any 1 out of 4 conditions have acceptable values

● at least 2 out of 36 samples have acceptable values

Figure 84 *Recommended filter parameters for Filter Flags*

Note: With the example experiment, the number of displayed entities declined from 3763 to 2885 entities when one-hit wonders are removed.



Workflow Type - Analysis: Significance Testing and Fold Change (Step 3 of 8)

Steps

1.Summary Report

2.Experiment Grouping

3.Filter Flags

4.Filter By Frequency

5.QC on samples

6.Significance Analysis

7.Fold Change

8.IDBrowser Identification

Filter Flags

If flag values are present, entities are filtered based on their flag values. Otherwise, entities are filtered based on their signal intensity values. To change the filter criteria, click on the "Re-run Filter" button.

Displaying **2885** out of **3763** entities where atleast **2** out of **40** samples have flags in **[P, M]**.

Figure 85 *Profile plot after removing one-hit wonders. The number of entities is reduced from 3763 to 2885 entities.*

h *(optional)* To re-adjust the filter parameters again, click **Re-run Filter** until the results displayed in the Profile Plot are satisfactory. Re-run the filter several times with differing parameters to develop an understanding of how each parameter affects the results.

i Click **Next**.

---

4. Enter the parameters for Filter By Frequency in the **Analysis: Significance Testing and Fold Change (Step 4 of 8)** workflow.

The entities are filtered from further analysis based on their frequency of occurrence among the samples and conditions. See "Definitions" on page 160 for definitions and relationships of the terms used by the class prediction workflow.

Filter by Frequency defines the filter by the minimum percentage of samples an entity must be present in to pass the filter. The filter is specified by typing the minimum percentage and selecting the applicable condition of the samples for which each entity must be present, i.e., Retain entities that appear in at least ________ %:

- of all the samples (classifications, or conditions, are not evaluated)
- of samples in only one condition (one and only one classification, or condition)
- of samples in at least one condition (one or more classification, or condition)
- of samples within each condition (all classifications, or conditions)

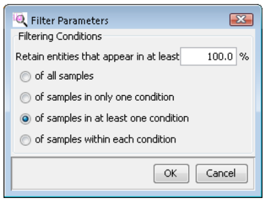
Filter by Frequency is set by default to retain the entities that appear in at least 100% of all the samples in at least one condition. This is the recommended percentage for experiments that contain five or fewer replicates. A larger percentage removes more entities from further statistical consideration. For experiments with a larger number of replicates the filter frequency percentage may be lowered to reflect the required occurrence.

a Click **Re-run Filter**.

b Type `100` in the **Retain entities that appear in at least** box.

c Click **of samples in at least one condition**.

d Click **OK**.



Filter Parameters

Filtering Conditions

Retain entities that appear in at least 100.0 %

○ of all samples

○ of samples in only one condition

● of samples in at least one condition

○ of samples within each condition

Figure 86 *Recommended Filter by Frequency parameters*

Note: With the example experiment, the number of entities changes to display 907 out of 2885 entities, reflecting the successful additional filtering by frequency (see Figure 87 on page 91).

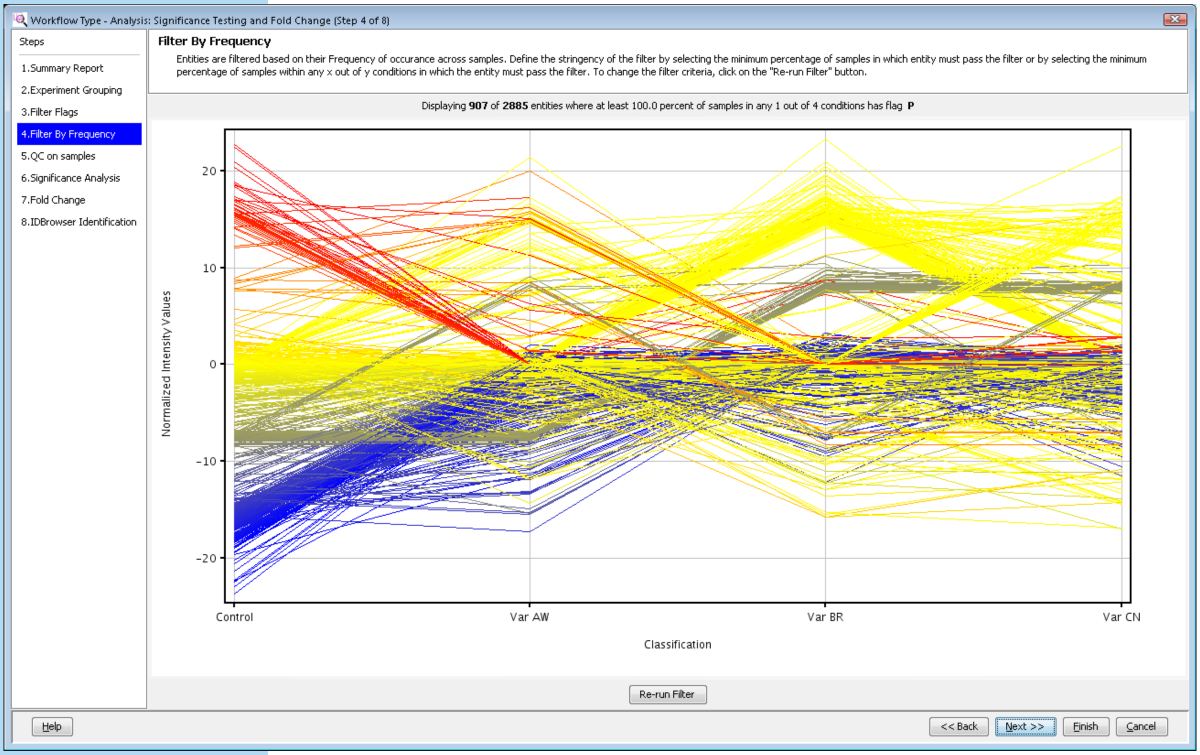
Bias from a condition with a single data file

Effect of a condition containing a single sample data file: If you retained the Class Prediction grouping, see Figure 81 on page 86, the number of entities passing this filter is larger (2030 out of 2885 entities) that expected. The "Class Prediction" grouping in this experiment added four conditions that contain a single data file, each of the four "Classification" parameter values with a "Class Prediction" parameter value of Validation. All of the entities in these four data files therefore meet the filter parameters. This bias in the number of entities that pass the Filter by Frequency filter from a condition containing a single sample data file is why the "Class Prediction" grouping was removed in "Review the experiment grouping parameters associated with the independent variables and their classification in the Analysis: Significance Testing and Fold Change (Step 2 of 8) workflow." on page 86.

e Click **Re-run Filter** to re-adjust the filter parameters until the results displayed in the Profile Plot are satisfactory. Re-run the filter several times with different parameters to develop an understanding of how each parameter affects the results.

You can review the data, change the plot view, export selected data, or export the plot to a file by using left-click and right-click features available on the plot in the same manner as presented in "Inspect the entities in the Analysis: Significance Testing and Fold Change (Step 1 of 8) workflow." on page 85.

f Click **Next**.



Workflow Type - Analysis: Significance Testing and Fold Change (Step 4 of 8)

Steps

1.Summary Report
2.Experiment Grouping
3.Filter Flags
4.Filter By Frequency
5.QC on samples
6.Significance Analysis
7.Fold Change
8.IDBrowser Identification

Filter By Frequency

Entities are filtered based on their Frequency of occurance across samples. Define the stringency of the filter by selecting the minimum percentage of samples in which entity must pass the filter or by selecting the minimum percentage of samples within any x out of y conditions in which the entity must pass the filter. To change the filter criteria, click on the "Re-run Filter" button.

Displaying **907** of **2885** entities where at least 100.0 percent of samples in any 1 out of 4 conditions has flag **P**

Figure 87 Profile plot after removing entities that do not appear in at least 100% of all the samples in at least one of four conditions. The number of entities is reduced from 2885 to 907 entities.

5. Review the sample quality in QC on samples in the **Analysis: Significance Testing and Fold Change (Step 5 of 8)** workflow.

A clear differentiation of your classifications in the 3D PCA Scores scatter plot is very important to y our analysis.

This step provides the first view of the data using a Principal Component Analysis (PCA). PCA helps you review the data by displaying a 3D scatter plot of the calculated principal components. The PCA scores are shown in each of the selection boxes located along the bottom of the 3D PCA Scores window. A higher score indicates that the principal component contains more of the variability of the data. The components generated in the 3D PCA Scores graph are represented in the X, Y, and Z axes and are numbered 1, 2, 3 ... in order of their decreasing significance.

Principal component analysis: The mathematical process by which data containing a number of potentially correlated variables is transformed into a data set in relation to a smaller number of variables called principal components that account for the most variability in the data. The result of the data transformation leads to the identification of the best explanation of the variance in the data, e.g. identification of the components in the data that contain the meaningful information providing differentiation.

Principal component: Transformed data into axes, principal components, so that the patterns between the axes most closely describe the relationships between the data. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The principal components are viewed and interpreted in 3D graphical axes with additional dimensions represented by different colors and/or shapes representing the parameter names.

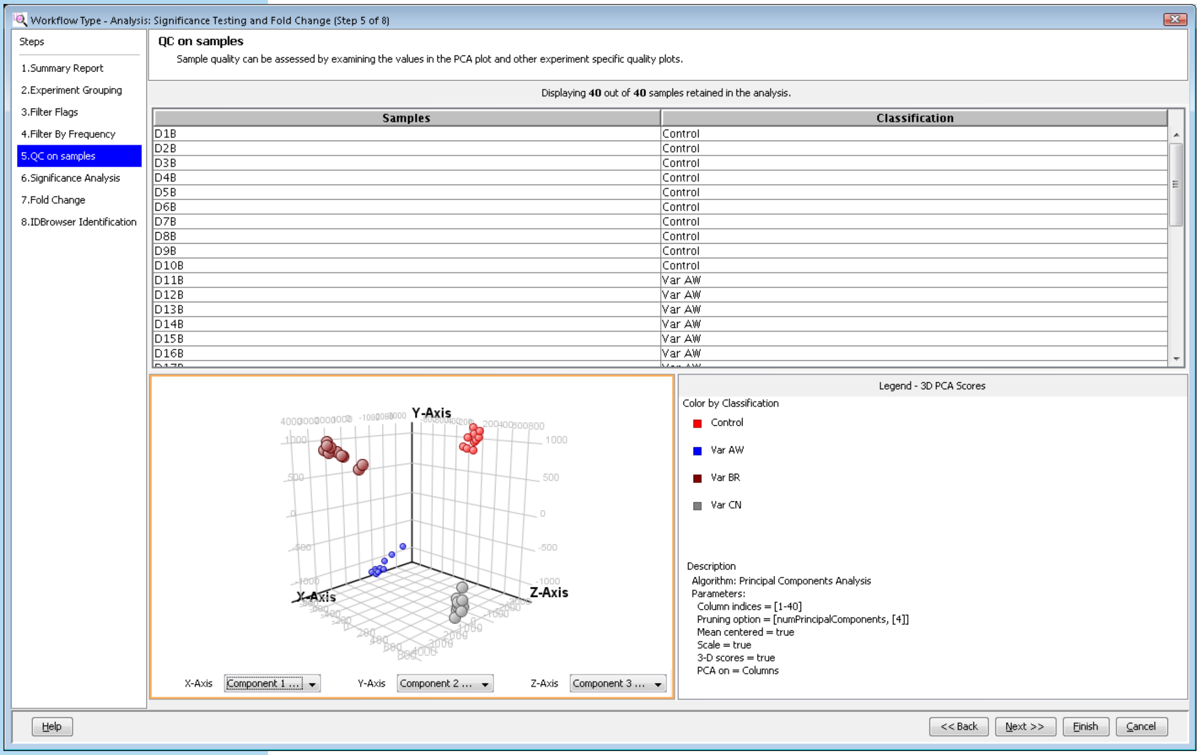
QC on samples display is divided into three viewing areas

a Review the **Experiment Grouping view**. This table displays each of the samples within a parameter (now referred to as a group). Ideally, replicates within a group should cluster together and be separated from samples in other groups.

b Review the **3D PCA Scores scatter plot view**. You may change the plot view or export the plot to a file by using the left-click and right-click actions. Additional controls available are:

- To customize the 3D PCA scores plot, right-click and then click **Properties**.
- To zoom into the 3D scatter plot, press the **Shift** key and simultaneously click the mouse button and drag the mouse upwards.
- To zoom out, press the **Shift** key and simultaneously click the mouse button and drag the mouse downwards.
- To rotate, press the **Ctrl** key and simultaneously click the mouse button and drag the mouse around the plot.

c Review the **Legend - 3D PCA Scores view**. This window shows the legend of the scatter plot.



Workflow Type - Analysis: Significance Testing and Fold Change (Step 5 of 8)

QC on samples

Sample quality can be assessed by examining the values in the PCA plot and other experiment specific quality plots.

Displaying **40** out of **40** samples retained in the analysis.

Samples	Classification
D1B	Control
D2B	Control
D3B	Control
D4B	Control
D5B	Control
D6B	Control
D7B	Control
D8B	Control
D9B	Control
D10B	Control
D11B	Var AW
D12B	Var AW
D13B	Var AW
D14B	Var AW
D15B	Var AW
D16B	Var AW

Legend - 3D PCA Scores

Color by Classification

🟥 Control

🟦 Var AW

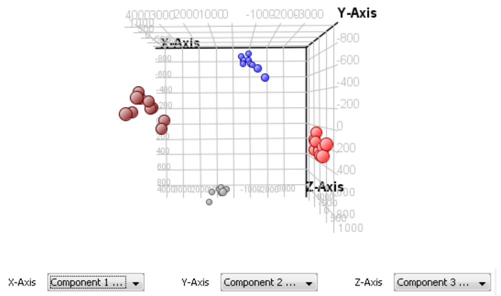
🟫 Var BR

⬜ Var CN

Description
Algorithm: Principal Components Analysis
Parameters:
Column indices = [1-40]
Pruning option = [numPrincipalComponents, [4]]
Mean centered = true
Scale = true
3-D scores = true
PCA on = Columns

Figure 88 *QC on samples PCA Score showing the initial separation of the experiment classifications. The 907 entities that met the filter flags and filter by frequency parameters are evaluated by this quality control step.*

Note: Click **Back** to make changes in the parameters for **Filter Flags** on page 87 and **Filter By Frequency** on page 89. Return to the **QC on samples** step several times to understand how each of the parameters affects your compound summary.



X-Axis Component 1 ... Y-Axis Component 2 ... Z-Axis Component 3 ...

Figure 89 *Rotated view of the 3D PCA Scores scatter plot*

d Click **Next**.

6. Assess the differential Significance Analysis in the **Analysis: Significance Testing and Fold Change (Step 6 of 8)** workflow.

The entities among your samples are expected to show significant differentiation among the classifications as shown in the Venn diagram or the Volcano Plot shown in th is step. Data with a single independent variable only a spreadsheet of your data is shown in this step, therefore the 3D PCA Scores on the prior step is used to view your sample differentiation.

The entities are filtered based on their p-values calculated from a statistical analysis that is selected based on the samples and experiment grouping.

The statistical analysis is either a T-test or an Analysis of Variance (ANOVA) based on the samples and experiment grouping.

a Review the Significance Analysis display. The display is divided into four viewing areas:

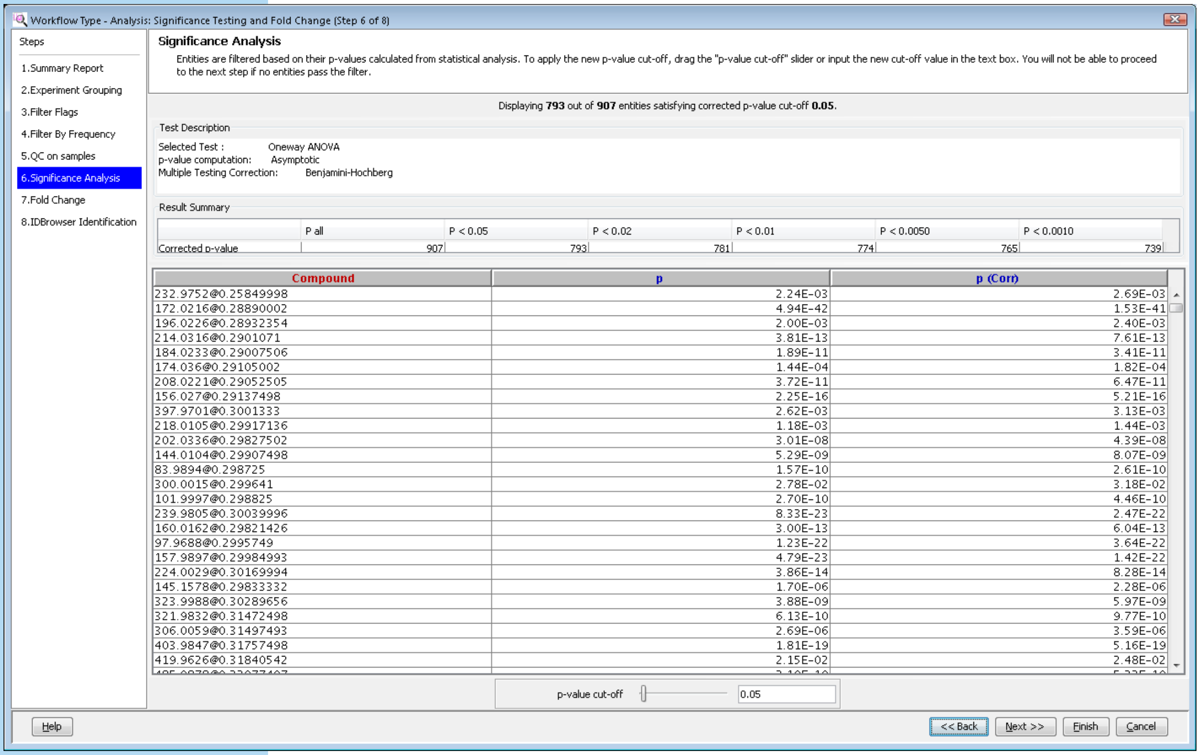
Test Description view: The statistical test applied to the samples is described.

Result Summary view: A summary table that organizes the results by p-value. A p-value of 0.05 is similar to stating that if the mean values for each parameter value (a condition of an independent variable) are identical, then a 5% chance or less exists of observing a difference in the mean of the parameter values as large as you observed. In other words, statistical treatment of random sampling from identical populations with a p-value set at 0.05 leads to a difference smaller than you observed in 95% of the experiments and larger than you observed in 5% of the experiments.

The last row of data in the Result Summary (see Figure 90 on page 94) shows the number of entities that would be expected to meet the significance analysis by random chance based on the p-value specified in each column heading. If the number of entities expected by chance is much smaller than the number expected based on the corrected p-value you have realized a selection of entities that show significance in the difference of the mean values of the parameter values.

Compounds p-Values Table view: Each entity that survived the filters is now presented by compound along with the p-values expected and corrected for each of the interpretation sets. Each entity is uniquely identified by its average neutral mass and retention time from across the data sets.

Venn Diagram view: Display of the Venn Diagram, or other plot, depends on the samples and experiment grouping for the analysis (see Figure 90 on page 94). The entities that make up each selected section of the Venn diagram are highlighted in the p-values spreadsheet. The Venn diagram is a graphical view of the most significant entities in each of the samples. Where entities in common to the analyses exist, they are depicted as overlapping sections of the circles. Fewer entities in the regions of overlap are an indication that the samples support the hypothesis that a difference exists in the samples based on the experimental parameters.



Significance Analysis

Entities are filtered based on their p-values calculated from statistical analysis. To apply the new p-value cut-off, drag the "p-value cut-off" slider or input the new cut-off value in the text box. You will not be able to proceed to the next step if no entities pass the filter.

Displaying **793** out of **907** entities satisfying corrected p-value cut-off **0.05**.

Test Description

Selected Test : Oneway ANOVA
p-value computation: Asymptotic
Multiple Testing Correction: Benjamini-Hochberg

Result Summary

	P all	P < 0.05	P < 0.02	P < 0.01	P < 0.0050	P < 0.0010
Corrected p-value	907	793	781	774	765	739

Compound	p	p (Corr)
232.9752@0.25849998	2.24E-03	2.69E-03
172.0216@0.28890002	4.94E-42	1.53E-41
196.0226@0.28932354	2.00E-03	2.40E-03
214.0316@0.2901071	3.81E-13	7.61E-13
184.0233@0.29007506	1.89E-11	3.41E-11
174.036@0.29105002	1.44E-04	1.82E-04
208.0221@0.29052505	3.72E-11	6.47E-11
156.027@0.29137498	2.25E-16	5.21E-16
397.9701@0.3001333	2.62E-03	3.13E-03
218.0105@0.29917136	1.18E-03	1.44E-03
202.0336@0.29827502	3.01E-08	4.39E-08
144.0104@0.29907498	5.29E-09	8.07E-09
83.9894@0.298725	1.57E-10	2.61E-10
300.0015@0.299641	2.78E-02	3.18E-02
101.9997@0.298825	2.70E-10	4.46E-10
239.9805@0.30039996	8.33E-23	2.47E-22
160.0162@0.29821426	3.00E-13	6.04E-13
97.9688@0.2995749	1.23E-22	3.64E-22
157.9897@0.29984993	4.79E-23	1.42E-22
224.0029@0.30169994	3.86E-14	8.28E-14
145.1578@0.29833332	1.70E-06	2.28E-06
323.9988@0.30289656	3.88E-09	5.97E-09
321.9832@0.31472498	6.13E-10	9.77E-10
306.0059@0.31497493	2.69E-06	3.59E-06
403.9847@0.31757498	1.81E-19	5.16E-19
419.9626@0.31840542	2.15E-02	2.48E-02

p-value cut-off 0.05

__Figure 90__ Significance analysis based on a one-way ANOVA using the example experiment. The results show a spreadsheet view of the data instead of a Venn diagram. A one-way ANOVA significance analysis does not present a graphical representation of the entities relationships. The number of entities is reduced from 907 to 793 entities.

b Click and move the **p-value cut-off** slider or type in the p-value cut-off value and press the **Enter** key. The default value is 0.05. The results in the display window are automatically updated.

c Re-adjust the p-value cut-off until the results displayed are satisfactory. The analysis can be re-run several times to develop an understanding of how the p-value cut-off affects the results. A larger p-value passes a larger number of entities.

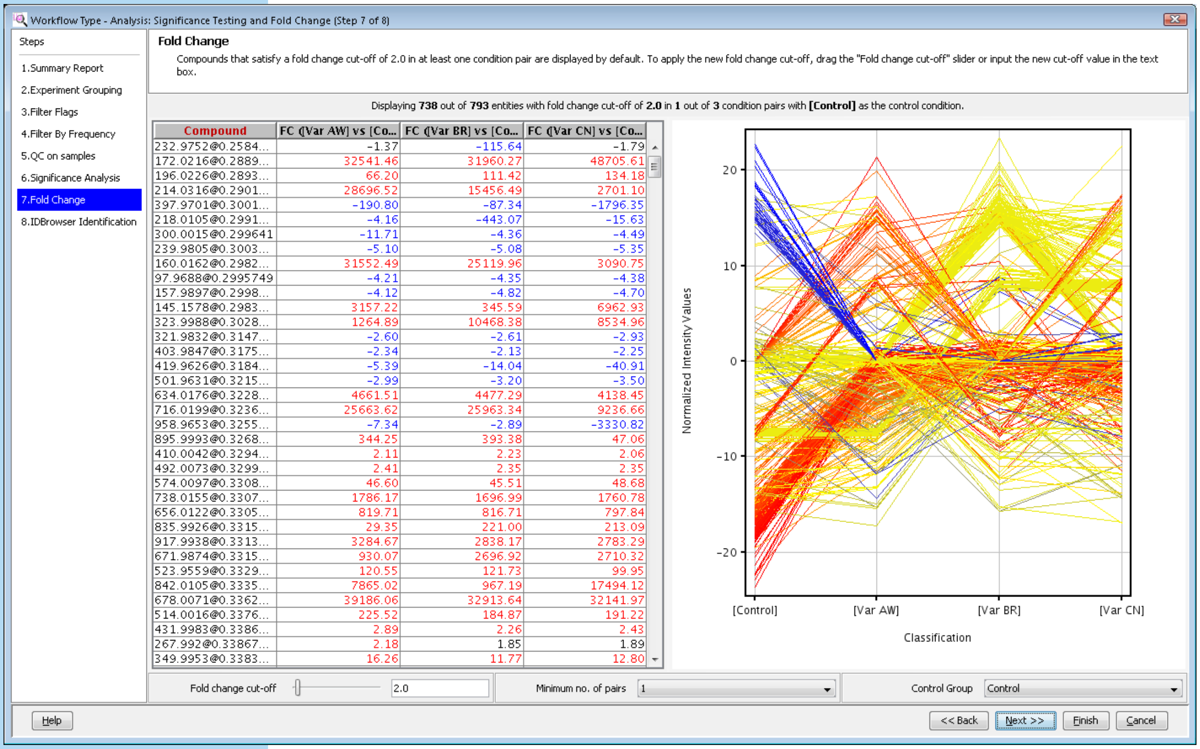
d Click **Next**.

7. Adjust the Fold Change in the **Analysis: Significance Testing and Fold Change (Step 7 of 8)** workflow.

Fold Change is used to identify entities with abundance ratios or differences between a pair of conditions, or classifications, that are greater than a set cut-off or threshold. Fold change is calculated between the conditions where Condition 1 and Condition 2 are treated as an aggregate. Fold change calculates the ratio between Condition 1 and Condition 2 as an absolute ration (Fold change= |Condition1 / Condition2|).

a Review the results of the default fold change parameters.

b Adjust the **Fold change cut-off** to obtain the best results. A Fold Change value of 2.0 is illustrated in Figure 91.



Fold Change

Compounds that satisfy a fold change cut-off of 2.0 in at least one condition pair are displayed by default. To apply the new fold change cut-off, drag the "Fold change cut-off" slider or input the new cut-off value in the text box.

Displaying **738** out of **793** entities with fold change cut-off of **2.0** in **1** out of **3** condition pairs with **[Control]** as the control condition.

Compound	FC ([Var AW] vs [Co...	FC ([Var BR] vs [Co...	FC ([Var CN] vs [Co...
232.9752@0.2584...	-1.37	-115.64	-1.79
172.0216@0.2889...	32541.46	31960.27	48705.61
196.0226@0.2893...	66.20	111.42	134.18
214.0316@0.2901...	28696.52	15456.49	2701.10
397.9701@0.3001...	-190.80	-87.34	-1796.35
218.0105@0.2991...	-4.16	-443.07	-15.63
300.0015@0.299641	-11.71	-4.36	-4.49
239.9805@0.3003...	-5.10	-5.08	-5.35
160.0162@0.2982...	31552.49	25119.96	3090.75
97.9688@0.2995749	-4.21	-4.35	-4.38
157.9897@0.2998...	-4.12	-4.82	-4.70
145.1578@0.2983...	3157.22	345.59	6962.93
323.9988@0.3028...	1264.89	10468.38	8534.96
321.9832@0.3147...	-2.60	-2.61	-2.93
403.9847@0.3175...	-2.34	-2.13	-2.25
419.9626@0.3184...	-5.39	-14.04	-40.91
501.9631@0.3215...	-2.99	-3.20	-3.50
634.0176@0.3228...	4661.51	4477.29	4138.45
716.0199@0.3236...	25663.62	25963.34	9236.66
958.9653@0.3255...	-7.34	-2.89	-3330.82
895.9993@0.3268...	344.25	393.38	47.06
410.0042@0.3294...	2.11	2.23	2.06
492.0073@0.3299...	2.41	2.35	2.35
574.0097@0.3308...	46.60	45.51	48.68
738.0155@0.3307...	1786.17	1696.99	1760.78
656.0122@0.3305...	819.71	816.71	797.84
835.9926@0.3315...	29.35	221.00	213.09
917.9938@0.3313...	3284.67	2838.17	2783.29
671.9874@0.3315...	930.07	2696.92	2710.32
523.9559@0.3329...	120.55	121.73	99.95
842.0105@0.3335...	7865.02	967.19	17494.12
678.0071@0.3362...	39186.06	32913.64	32141.97
514.0016@0.3376...	225.52	184.87	191.22
431.9983@0.3386...	2.89	2.26	2.43
267.992@0.33867...	2.18	1.85	1.89
349.9953@0.3383...	16.26	11.77	12.80

Fold change cut-off 2.0

Minimum no. of pairs 1

Control Group Control

Figure 91 *Fold change using the example experiment. The number of entities is reduced from 793 to 738 entities.*

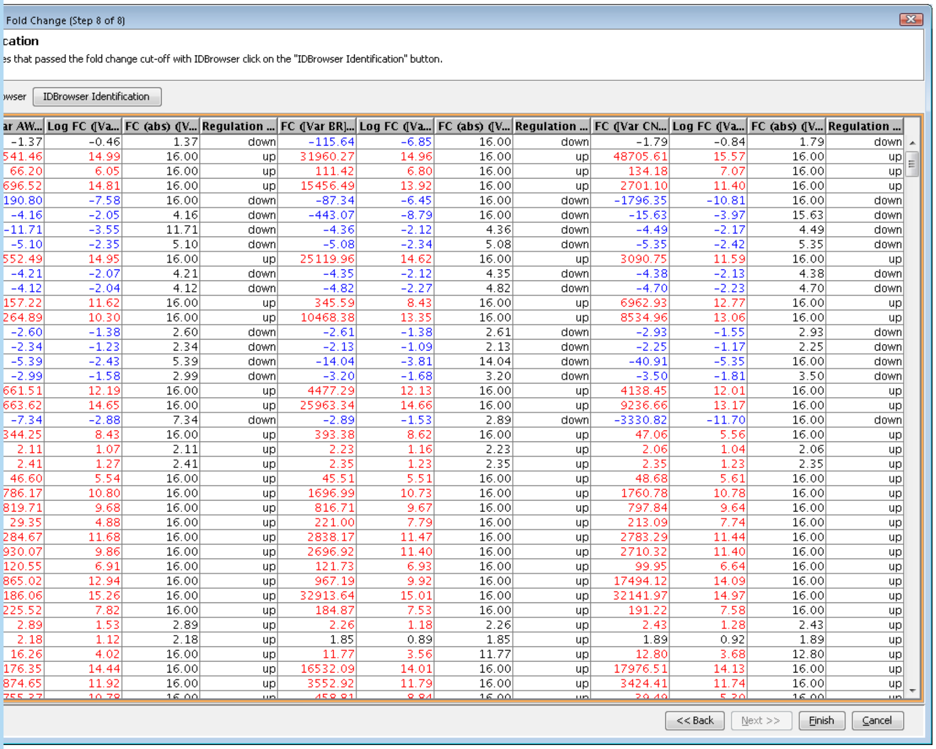
c Click **Next**.

8. Skip ID Browser Identification in the **Analysis: Significance Testing and Fold Change (Step 8 of 8)** workflow.

Feature identification is not necessary at this time because the object of this experiment is to determine suitability of the features for class prediction, and optionally saving the features for recursive finding.

Additional information regarding the use of ID Browser is covered in the *Integrated Biology with Agilent Mass Profiler Professional - Workflow Guide* (5991-1909EN, Revision A, June 2013).

Click **Finish**.



Fold Change (Step 8 of 8)

cation

es that passed the fold change cut-off with IDBrowser click on the "IDBrowser Identification" button.

ar AW...	Log FC ([Va...	FC (abs) ([V...	Regulation ...	FC ([Var BR]...	Log FC ([Va...	FC (abs) ([V...	Regulation ...	FC ([Var CN...	Log FC ([Va...	FC (abs) ([V...	Regulation ...
-1.37	-0.46	1.37	down	-115.64	-6.85	16.00	down	-1.79	-0.84	1.79	down
541.46	14.99	16.00	up	31960.27	14.96	16.00	up	48705.61	15.57	16.00	up
66.20	6.05	16.00	up	111.42	6.80	16.00	up	134.18	7.07	16.00	up
696.52	14.81	16.00	up	15456.49	13.92	16.00	up	2701.10	11.40	16.00	up
190.80	-7.58	16.00	down	-87.34	-6.45	16.00	down	-1796.35	-10.81	16.00	down
-4.16	-2.05	4.16	down	-443.07	-8.79	16.00	down	-15.63	-3.97	15.63	down
-11.71	-3.55	11.71	down	-4.36	-2.12	4.36	down	-4.49	-2.17	4.49	down
-5.10	-2.35	5.10	down	-5.08	-2.34	5.08	down	-5.35	-2.42	5.35	down
552.49	14.95	16.00	up	25119.96	14.62	16.00	up	3090.75	11.59	16.00	up
-4.21	-2.07	4.21	down	-4.35	-2.12	4.35	down	-4.38	-2.13	4.38	down
-4.12	-2.04	4.12	down	-4.82	-2.27	4.82	down	-4.70	-2.23	4.70	down
157.22	11.62	16.00	up	345.59	8.43	16.00	up	6962.93	12.77	16.00	up
264.89	10.30	16.00	up	10468.38	13.35	16.00	up	8534.96	13.06	16.00	up
-2.60	-1.38	2.60	down	-2.61	-1.38	2.61	down	-2.93	-1.55	2.93	down
-2.34	-1.23	2.34	down	-2.13	-1.09	2.13	down	-2.25	-1.17	2.25	down
-5.39	-2.43	5.39	down	-14.04	-3.81	14.04	down	-40.91	-5.35	16.00	down
-2.99	-1.58	2.99	down	-3.20	-1.68	3.20	down	-3.50	-1.81	3.50	down
661.51	12.19	16.00	up	4477.29	12.13	16.00	up	4138.45	12.01	16.00	up
663.62	14.65	16.00	up	25963.34	14.66	16.00	up	9236.66	13.17	16.00	up
-7.34	-2.88	7.34	down	-2.89	-1.53	2.89	down	-3330.82	-11.70	16.00	down
344.25	8.43	16.00	up	393.38	8.62	16.00	up	47.06	5.56	16.00	up
2.11	1.07	2.11	up	2.23	1.16	2.23	up	2.06	1.04	2.06	up
2.41	1.27	2.41	up	2.35	1.23	2.35	up	2.35	1.23	2.35	up
46.60	5.54	16.00	up	45.51	5.51	16.00	up	48.68	5.61	16.00	up
786.17	10.80	16.00	up	1696.99	10.73	16.00	up	1760.78	10.78	16.00	up
819.71	9.68	16.00	up	816.71	9.67	16.00	up	797.84	9.64	16.00	up
29.35	4.88	16.00	up	221.00	7.79	16.00	up	213.09	7.74	16.00	up
284.67	11.68	16.00	up	2838.17	11.47	16.00	up	2783.29	11.44	16.00	up
930.07	9.86	16.00	up	2696.92	11.40	16.00	up	2710.32	11.40	16.00	up
120.55	6.91	16.00	up	121.73	6.93	16.00	up	99.95	6.64	16.00	up
865.02	12.94	16.00	up	967.19	9.92	16.00	up	17494.12	14.09	16.00	up
186.06	15.26	16.00	up	32913.64	15.01	16.00	up	32141.97	14.97	16.00	up
225.52	7.82	16.00	up	184.87	7.53	16.00	up	191.22	7.58	16.00	up
2.89	1.53	2.89	up	2.26	1.18	2.26	up	2.43	1.28	2.43	up
2.18	1.12	2.18	up	1.85	0.89	1.85	up	1.89	0.92	1.89	up
16.26	4.02	16.00	up	11.77	3.56	11.77	up	12.80	3.68	12.80	up
176.35	14.44	16.00	up	16532.09	14.01	16.00	up	17976.51	14.13	16.00	up
874.65	11.92	16.00	up	3552.92	11.79	16.00	up	3424.41	11.74	16.00	up
755.37	10.78	16.00	up	458.81	8.84	16.00	up	39.49	5.30	16.00	up

Figure 92 *ID Browser using the example experiment.*

Layout of the Mass Profiler Professional screen

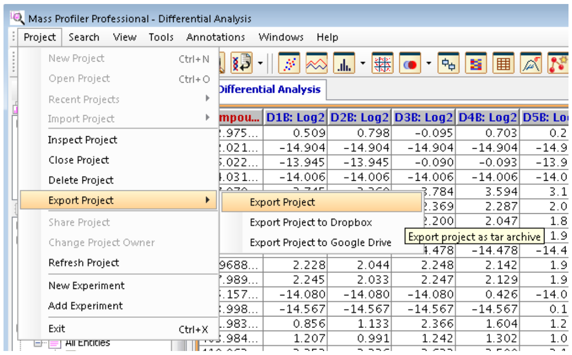
At the completion of your initial differential analysis, you are now in the advanced workflow mode and have access to all features available in Mass Profiler Professional through the Workflow Browser. Figure 93 on page 97 shows the layout of Mass Profiler Professional.

Figure 93 The main functional areas of Mass Profiler Professional after creating your project and experiment

Save the project containing your differential analysis

Save your current analysis as a TAR file for archiving, restoration of any future analysis to the current results, sharing the data with a collaborator, or sharing the data with Agilent customer support.

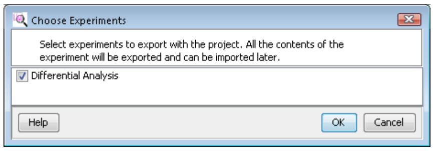
a Click **Project > Export Project > Export Project**.



mpou...	D1B: Log2	D2B: Log2	D3B: Log2	D4B: Log2	D5B: Lo
.975...	0.509	0.798	-0.095	0.703	0.2
.021...	-14.904	-14.904	-14.904	-14.904	-14.9
.022...	-13.945	-13.945	-0.090	-0.093	-13.9
.031...	-14.006	-14.006	-14.006	-14.006	-14.0
			.784	3.594	3.1
			.369	2.287	2.0
			.200	2.047	1.8
					1.9
			.478	-14.478	-14.4
9688...	2.228	2.044	2.248	2.142	1.9
.989...	2.245	2.033	2.247	2.129	1.9
.157...	-14.080	-14.080	-14.080	0.426	-14.0
.998...	-14.567	-14.567	-14.567	-14.567	0.1
.983...	0.856	1.133	2.366	1.604	1.2
.984...	1.207	0.991	1.242	1.302	1.0

Figure 94 Menu selection to export your current analysis

b Mark the check box next to the experiment you wish to save.



Choose Experiments

Select experiments to export with the project. All the contents of the experiment will be exported and can be imported later.

☑ Differential Analysis

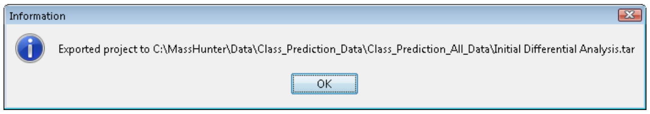
Figure 95 *Choose Experiments dialog box for saving your project*

c Click **OK**.

d Select or create the file folder.

e Type the **File name**.

f Click **Save**.



Information

Exported project to C:\MassHunter\Data\Class_Prediction_Data\Class_Prediction_All_Data\Initial Differential Analysis.tar

Figure 96 *Information dialog box confirming your saved project*

g Click **OK** in the **Information** dialog box that confirms the saved project.

Perform a recursive feature finding

After confirming that you can separate the sample data files by the classifications, use the filtered features from your initial differential analysis to perform a recursive find features in your sample data files. From the total number of untargeted features originally found in your sample data files, the initial differential analysis identifies a subset of features that contribute to the best separation of your experimental conditions (classifications). By using the entity list containing these significant features to perform a targeted feature finding, you improve the statistical accuracy of your differential analysis and can improve the accuracy of your subsequent class prediction model development.

Combined with collecting replicate samples in your experiment, recursive feature finding improves the statistical accuracy (confidence) of your analysis and reduces the potential for obtaining a false positive or a false negative answer to your hypothesis and sample classification. For an overview of finding features recursively, see “Review recursive feature finding” on page 34.

If you are unsure whether to perform recursive feature finding, continue with “Decide whether or not to perform recursive feature finding” below. If your analysis does not involve recursive feature finding, you may skip this step in the workflow and continue with your analysis at “Build your class prediction model” on page 127.

Decide whether or not to perform recursive feature finding

Performing a recursive feature finding is beneficial to your analysis; it provides an improved measure of confidence of your initial differential analysis and subsequent class prediction model. The improvement gained in your class prediction workflow depends partially on whether your model is based on features that show up and down regulation among your classification or the presence and absence of features.

- If your prediction model relies on the regulation of strong features, enhanced finding of weak targeted features may not significantly help you during your initial model creation; recursive feature finding is optional before creating your initial class prediction model.

- If your prediction model relies on feature presence and absence, then it is recommended to perform a recursive feature find before creating your initial class prediction model.

- If your hypothesis does not include a priori knowledge on the regulation or absence of features in your samples, then it is recommended to perform recursive feature finding before creating your initial class prediction model.

Overview of the steps for performing recursive feature finding

During recursive (targeted) feature finding you export the more important features from your initial differential analysis as a targeted list of features for finding in your original sample data files. Recursive feature finding improves the quality of finding the features in the original sample files; targeted feature finding focuses on finding a specific set of features with less emphasis on feature strength. The steps followed to perform a recursive feature finding are similar to the untargeted feature finding you performed in the Chapter “Find the features in your samples” on page 39.

You may use one of two processes to perform a recursive feature finding of the features in all of your sample data files: Qualitative Analysis and Profinder (see Figure 97). Both programs export the sample features using a compound exchange format (CEF) file. A single CEF file is generated for each sample data file.

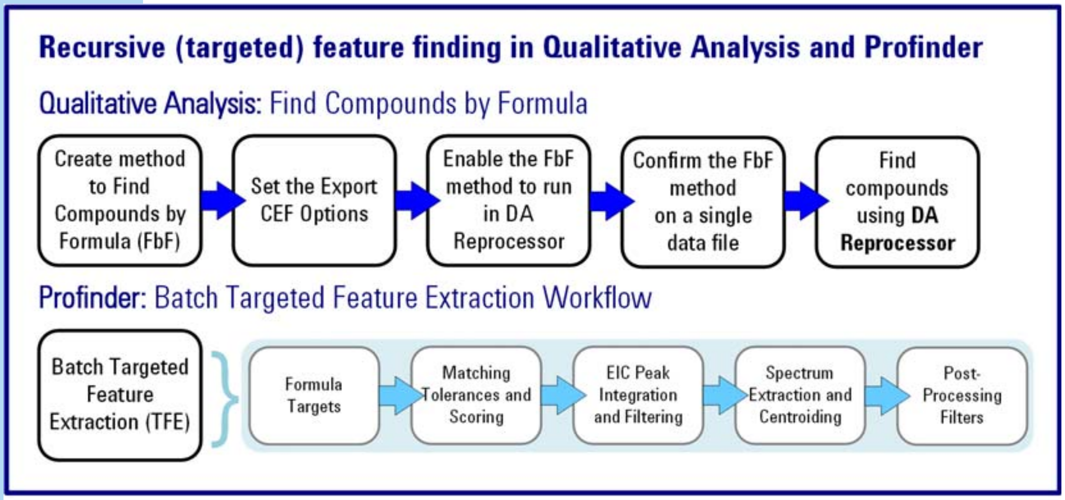

Recursive (targeted) feature finding in Qualitative Analysis and Profinder
Qualitative Analysis: Find Compounds by Formula
Create method to Find Compounds by Formula (FbF)
Set the Export CEF Options
Enable the FbF method to run in DA Reprocessor
Confirm the FbF method on a single data file
Find compounds using DA Reprocessor
Profinder: Batch Targeted Feature Extraction Workflow
Batch Targeted Feature Extraction (TFE)
Formula Targets
Matching Tolerances and Scoring
EIC Peak Integration and Filtering
Spectrum Extraction and Centroiding
Post-Processing Filters

Figure 97 *Comparison of the process to find targeted features using Qualitative Analysis and Profinder*

Export the significant features from your differential analysis.

1. "Export the entity list for recursion" on page 101

Recursive feature finding using Qualitative Analysis

Create a method to Find Compounds by Formula (FbF) in Qualitative Analysis.

2. "Recursive feature finding using Qualitative Analysis" on page 102
3. "Create a method to Find Compounds by Formula" on page 102
4. "Save your Find Compounds by Formula method" on page 114
5. "Set the Export CEF Options" on page 114
6. "Enable the method to run in MassHunter DA Reprocessor" on page 115

Confirm your FbF method using a single sample data file.

7. "Confirm the FbF method on a single data file" on page 116

Perform recursive feature finding in the entire sample data set using DA Reprocessor.

8. "Find compounds using DA Reprocessor" on page 117

Recreate your project experiment, filters, and differential analysis in MPP.

9. "Import and organize your recursive data" on page 125
10. "Recreate your differential analysis using the recursive features" on page 125
11. "Save the project containing your recursive analysis" on page 126

Recursive feature finding using Profinder

Finding the features in your sample data files using Profinder involves one self-directed wizard that involves five steps,"Recursive feature finding using MassHunter Profinder" on page 117

As you follow the steps involved in recursive feature finding, keep in mind that the features in a sample may be individually referred to as a **compound**, **descriptor**, **element**, **entity**, **feature**, or **metabolite** during the various steps of the class prediction workflow.

Export the entity list for recursion

Export the subset of features that contribute to the best separation of your experimental conditions (classifications). These significant features from your differential analysis are used to perform a targeted feature find from the original data files.

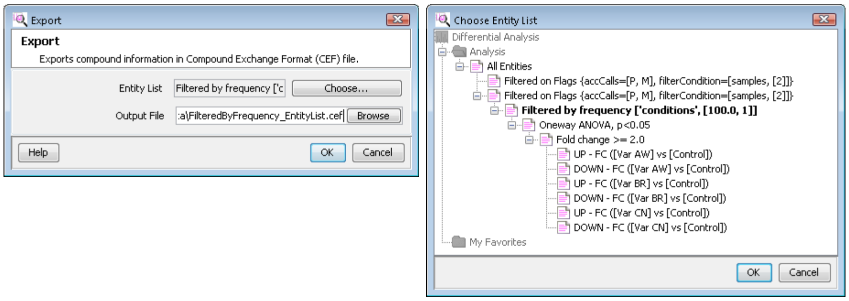
a Click **Export for Recursion** in the Workflow Browser under the Results Interpretations group heading. This displays the **Export** dialog box.

b Click **Choose** to select the Entity List for exporting.

c Click the **Filtered by frequency** from the entity lists in the **Choose Entity List** dialog box.

Note: If a fold change entity list is available in your experiment, you can use the fold change entity list for recursive feature finding. Not all experiments create a fold change entity list; since a filter by frequency entity list is always created, it is used as the example in this workflow guide.

The **Filtered by frequency** entity list is the list of features used to show differentiation in the PCA of the 40 samples among the four classifications in "Review the sample quality in QC on samples in the Analysis: Significance Testing and Fold Change (Step 5 of 8) workflow." on page 91. For optimal significance in recursive feature finding, select an entity list for recursion that list that has at least been **Filtered on Flags** to remove one-hit wonders, and preferably use a **Fold change** entity list.



Export

Exports compound information in Compound Exchange Format (CEF) file.

Entity List: Filtered by frequency ['c

Output File: :a\FilteredByFrequency_EntityList.cef

```
Differential Analysis
  Analysis
    All Entities
      Filtered on Flags {accCalls=[P, M], filterCondition=[samples, [2]]}
      Filtered on Flags {accCalls=[P, M], filterCondition=[samples, [2]]}
        Filtered by frequency ['conditions', [100.0, 1]]
          Oneway ANOVA, p<0.05
            Fold change >= 2.0
              UP - FC ([Var AW] vs [Control])
              DOWN - FC ([Var AW] vs [Control])
              UP - FC ([Var BR] vs [Control])
              DOWN - FC ([Var BR] vs [Control])
              UP - FC ([Var CN] vs [Control])
              DOWN - FC ([Var CN] vs [Control])
  My Favorites
```

Figure 98 Export and Choose Entity List dialog boxes

d Click **OK**.

e Click **Browse** in the **Export** dialog box. Do not type a file name at this location.

f Select the folder to which to save the file.

g Type `FilteredByFrequency_EntityList.cef` for the **File name**.

h Click **Save**.

i Click **OK**.

Recursive feature finding using Qualitative Analysis

The following examples use MassHunter Qualitative Analysis B.06.00 running on 64-bit Windows 7 Professional. If you have TOF and/or Q-TOF data you can alternatively perform a recursive feature finding using Profinder as described beginning at "Recursive feature finding using MassHunter Profinder" on page 117.

1. Start MassHunter Qualitative Analysis Software.

a Double-click the Qualitative Analysis icon located on the desktop,

or (for Qualitative Analysis version B.05.00 or later on Windows 7)

Click **Start > All Programs > Agilent > MassHunter Workstation > Qualitative Analysis B.06.00**,

b Click **Cancel** in the **Open Data File** dialog box to start MassHunter Qualitative Analysis without opening any data files. To open data files later click **File > Open Data File.**

You do not need to open a data file at this time. You are prompted to open a data file in "Confirm the FbF method on a single data file" on page 116.

2. Enable advanced parameters in the user interface.

Advanced parameters must be enabled in MassHunter Qualitative Analysis in order to show tabs labeled Advanced in the Method Editor and to enable compound importing for recursive feature finding of molecular features.

a Check to make sure that **File > Import Compound** is an available command. See Figure 21 on page 43.

b If **File > Import Compound** is not available follow the instructions in "Enable advanced parameters in the user interface." on page 40.

c Continue with the next step.

Create a method to Find Compounds by Formula

Find Compounds by Formula (FbF) involves targeted feature finding using chromatographic deconvolution as shown in Figure 18 on page 41. FbF automatically finds related co-eluting ions, sums the related ion signals into single values, creates compound spectra, and reports results for each molecular feature.

All of the parameters involved in FbF are accessed in the tabs presented in four Method Editor sections that are selected from the Method Explorer window:

Find by Formula - Options: Specify the rules that are applied to match the data based on isotope patterns (*m/z* and abundance) and retention time

Find by Formula - Chromatograms: Enter parameters that are applied to the chromatographic component of the data to extract features.

Find by Formula - Mass Spectra: Enter parameters that are applied to the mass spectral component of the data to extract features.

Find by Formula - Sample Purity: Not used in this workflow.

After the parameters are entered in the Method Editor sections, to Find Compounds by Formula on a single sample data file, click the **Find Compounds by Formula** button from within any one of these Method Editor sections.

Finding the targeted features in your sample data files using Qualitative Analysis involves five sequential steps as shown in Figure 99:

- "Create a method to Find Compounds by Formula"
- "Set the Export CEF Options" on page 114
- "Enable the method to run in MassHunter DA Reprocessor" on page 115
- "Confirm the FbF method on a single data file" on page 116
- "Find compounds using DA Reprocessor" on page 117

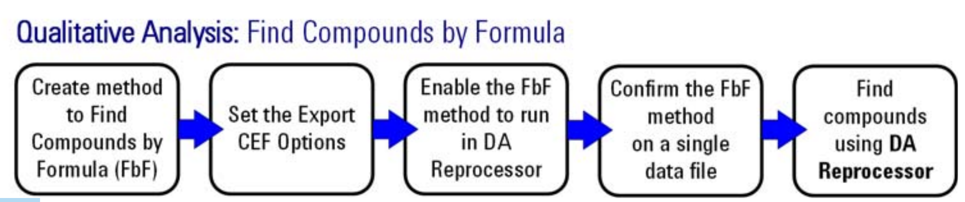

Qualitative Analysis: Find Compounds by Formula
Create method to Find Compounds by Formula (FbF)
Set the Export CEF Options
Enable the FbF method to run in DA Reprocessor
Confirm the FbF method on a single data file
Find compounds using DA Reprocessor

Figure 99 *Steps to find targeted features using Qualitative Analysis*

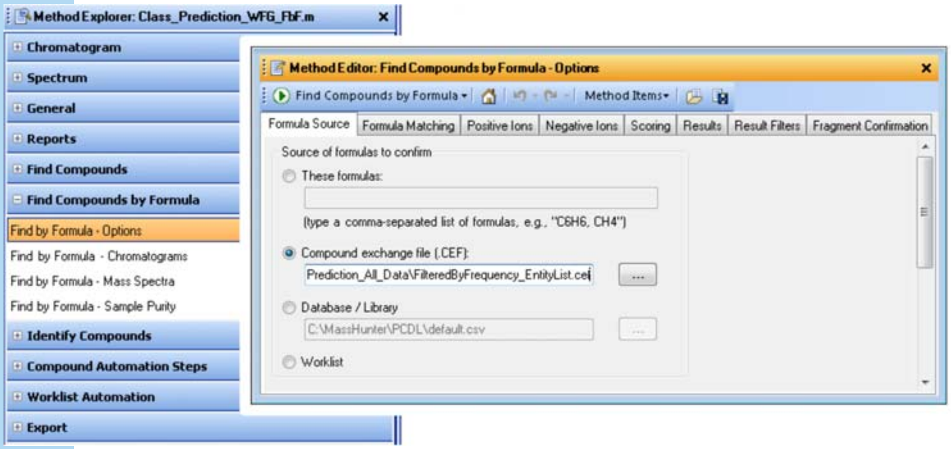
1. Open the Method Editor window for finding compounds by formula.

Open the **Method Editor: Find Compounds by Formula** section from the **Method Explorer** window.

1. Click **Find Compounds by Formula** from within the Method Explorer window.
2. Click **Find by Formula - Options**.

2. Enter the parameters in the Find by Formula - Options section.

The parameters in this section specify the rules that are applied to the formula database to match against the data based on isotope patterns (*m/z* and abundance) and retention time. This is the first part of the recursive refinement of finding features. The input options specify the rules that are applied to the input molecular formula database - the entity list you exported in "Export the entity list for recursion" on page 101.


Method Explorer: Class_Prediction_WFG_FbF.m
Chromatogram
Spectrum
General
Reports
Find Compounds
Find Compounds by Formula
Find by Formula - Options
Find by Formula - Chromatograms
Find by Formula - Mass Spectra
Find by Formula - Sample Purity
Identify Compounds
Compound Automation Steps
Worklist Automation
Export
Method Editor: Find Compounds by Formula - Options
Find Compounds by Formula
Method Items
Formula Source
Formula Matching
Positive Ions
Negative Ions
Scoring
Results
Result Filters
Fragment Confirmation

Source of formulas to confirm

○ These formulas:

(type a comma-separated list of formulas, e.g., "C6H6, CH4")

● Compound exchange file (.CEF):

Prediction_All_Data\FilteredByFrequency_EntityList.cef

○ Database / Library

C:\MassHunter\PCDL\default.csv

○ Worklist

Figure 100 *Overview of the Method Editor tabs associated with Find Compounds by Formula - Options*

Formula Source tab

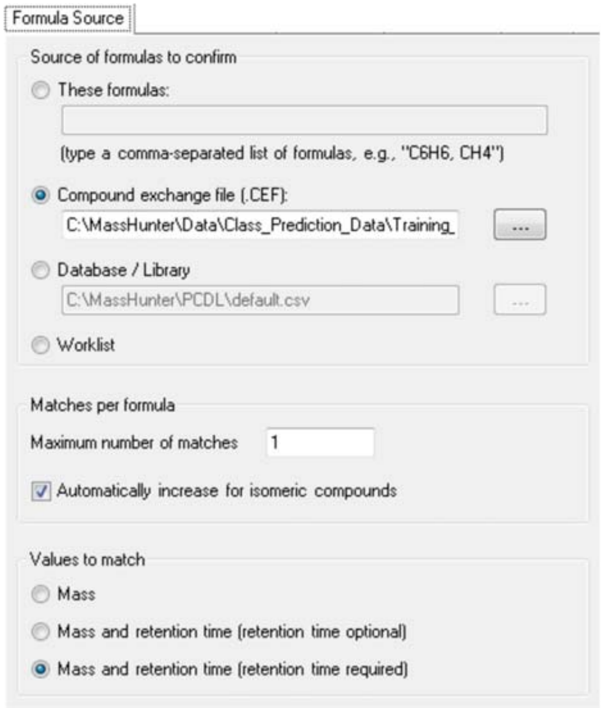
a Enter the parameters on the Formula Source tab.

The parameters on this tab let you use a molecular formula or previously created databases as the source of targeted features to find. In this workflow the source of targeted features is the CEF file you exported for recursion.

1. Click the **Formula Source** tab.
2. Click **Compound exchange file (.CEF)**.
3. Type the folder and file name of the CEF file you saved in "Export the entity list for recursion" on page 101 or click **Browse** and select a CEF file from the **Open CEF file** dialog box.
4. Open the CEF file that contains the most significant features created using Mass Profiler Professional from the "Export the entity list for recursion" on page 101.
5. Click **Open**.
6. Click **Mass and retention time (retention time required)**.

Note: The parameters under the *Values to match* group heading are only active if the **Compound exchange file (.CEF)** button or the **Database** button is clicked.

Note: Mass and retention time (retention time required) is the proper selection for a CEF file. Mass and retention time (retention time optional) is the proper selection for a database source.



Formula Source

Source of formulas to confirm

○ These formulas:

(type a comma-separated list of formulas, e.g., "C6H6, CH4")

● Compound exchange file (.CEF): C:\MassHunter\Data\Class_Prediction_Data\Training_ [...]

○ Database / Library: C:\MassHunter\PCDL\default.csv [...]

○ Worklist

Matches per formula

Maximum number of matches: 1

☑ Automatically increase for isomeric compounds

Values to match

○ Mass

○ Mass and retention time (retention time optional)

● Mass and retention time (retention time required)

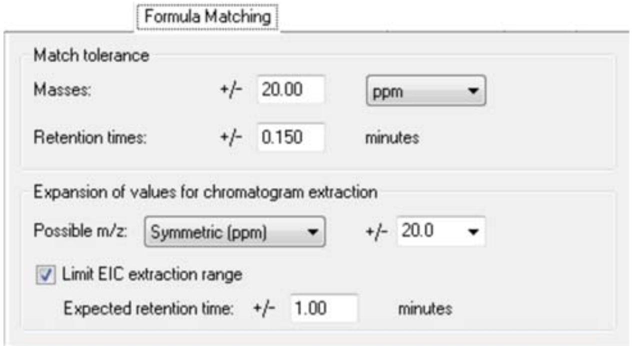
Figure 101 Formula Source tab in the Find by Formula - Options section

Formula Matching tab

b Enter the parameters on the Formula Matching tab.

The parameters in this tab specify the tolerances that are used to match the input values for mass and retention time against those found in the data.

1. Click the **Formula Matching** tab.
2. Type `20` for **Masses** tolerance and select **ppm** as the match tolerance units. Set this value wider than the measured instrumental acquisition mass tolerance to avoid losing valid feature matches.
3. Type `0.15` for **Retention times**. This parameter should be no less than two times the measured retention time tolerance.
4. Select **Symmetric (ppm)** and **± 20** ppm for **Possible m/z**. The parameters under the *Expansion of values for chromatographic extraction* group heading are used to direct the algorithm on how to handle saturated chromatographic data.
5. Mark the **Limit EIC extraction range** check box.
6. Type a value of `1.0` minutes for **Expected retention time**. The value may be between `1.0` and `1.5` minutes.



Formula Matching

Match tolerance

Masses: +/- 20.00 ppm

Retention times: +/- 0.150 minutes

Expansion of values for chromatogram extraction

Possible m/z: Symmetric (ppm) +/- 20.0

☑ Limit EIC extraction range

Expected retention time: +/- 1.00 minutes

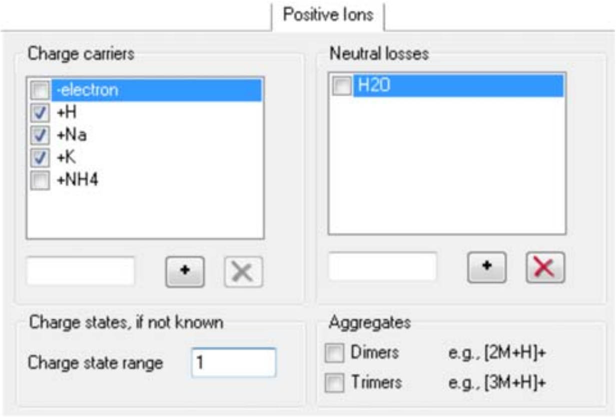
Figure 102 *Formula Matching tab in the Find by Formula - Options section*

Positive Ions tab

c Enter the parameters on the Positive Ions tab.

The parameters in this tab specify the positive ion adducts that the algorithm uses with the molecular formula to confirm that the feature was found in the data. Better results are derived from acquisition methodologies that minimize adducts, especially sodium and potassium.

1. Click the **Positive Ions** tab.
2. Mark the charge carriers **+H**, **+Na**, and **+K** that are known to be present in the data. Typically, positive protonated is the ideal selection. Non-adducted molecular ions, loss of an electron, are an option in Find by Formula.
3. Enter the molecular formulas for specific charge carriers in the input box below the charge carriers selection.
4. Clear **Neutral losses**. Neutral losses are not typically used. An exception is when a facile loss is expected.
5. Enter the molecular formulas for specific neutral losses in the input box below the neutral losses selection.
6. Type `1` for **Charge state range**.
7. Clear the **Dimers** check box.
8. Clear the **Trimers** check box.



Positive Ions

Charge carriers

□ -electron
☑ +H
☑ +Na
☑ +K
□ +NH4

Neutral losses

□ H2O

Charge states, if not known

Charge state range: 1

Aggregates

□ Dimers e.g., [2M+H]+
□ Trimers e.g., [3M+H]+

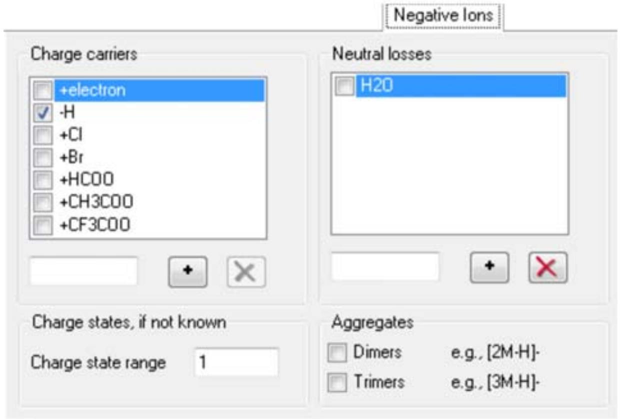
Figure 103 Positive Ions tab in the Find by Formula - Options section

Negative Ions tab

d Enter the parameters on the Negative Ions tab.

The parameters in this tab specify the negative ion adducts that the algorithm uses with the molecular formula to confirm that the feature was found in the data. Better results are derived from acquisition methodologies that minimize adducts.

1. Click the **Negative Ions** tab.
2. Mark the charge carrier **-H** that is known to be present in the data. Typically, negative deprotonated is the ideal selection. Non-adducted molecular ions, attachment of an electron, are an option in Find by Formula.
3. Enter the molecular formulas for specific charge carries in the input box below the charge carriers selection.
4. Clear **Neutral losses**. Neutral losses are not typically used. An exception is when a facile loss is expected.
5. Enter the molecular formulas for specific neutral losses in the input box below the neutral losses selection.
6. Type `1` for **Charge state range**.
7. Clear the **Dimers** check box.
8. Clear the **Trimers** check box.



Negative Ions

Charge carriers

□ +electron
☑ -H
□ +Cl
□ +Br
□ +HCOO
□ +CH3COO
□ +CF3COO

Neutral losses

□ H2O

Charge states, if not known

Charge state range: 1

Aggregates

□ Dimers e.g., [2M-H]-
□ Trimers e.g., [3M-H]-

Figure 104 Negative Ions tab in the Find by Formula - Options section

Scoring tab

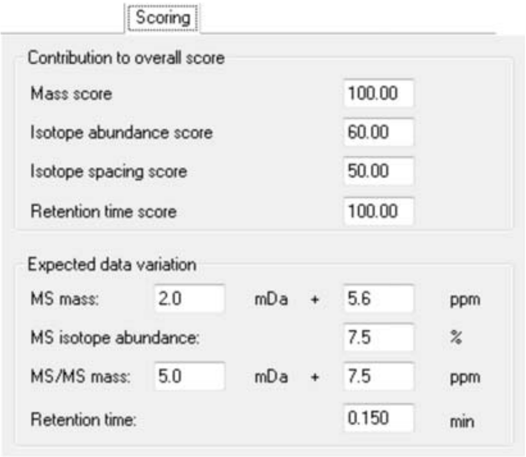
e Enter the parameters on the Scoring tab.

The parameters in this tab specify how to rate whether the spectral pattern is correct for the molecular formula. The scoring determines a goodness of fit between observed ions compared to the expected ions in the database. As signal levels decrease the scoring parameters entered may not match as well. The defaults provided are adequate.

1. Click the **Scoring** tab.
2. Type `100` for the **Mass score.**
3. Type `60` for the **Isotope abundance score**.
4. Type `50` for the **Isotope spacing score.**
5. Type `100` for the **Retention time score**.

Note: If you set values of **100** for the **Mass score** and `0` for **Isotope abundance** score, the **Isotope spacing score**, and the **Retention time score**, then the latter three scores are not included when calculating the **Score**.

6. Type the default values of `2.0` for **mDa** and `5.6` ppm for **MS mass**.
7. Type the default value of `7.5` % for **MS isotope abundance**.
8. Type the default values of `5.0` for **mDa** and `7.5` ppm for **MS/MS mass**.
9. Type the default value of `0.15` min for **Retention time**.



Scoring

Contribution to overall score

Mass score	100.00
Isotope abundance score	60.00
Isotope spacing score	50.00
Retention time score	100.00

Expected data variation

MS mass:	2.0	mDa	+	5.6	ppm
MS isotope abundance:				7.5	%
MS/MS mass:	5.0	mDa	+	7.5	ppm
Retention time:				0.150	min

Figure 105 *Scoring tab in the Find by Formula - Options section*

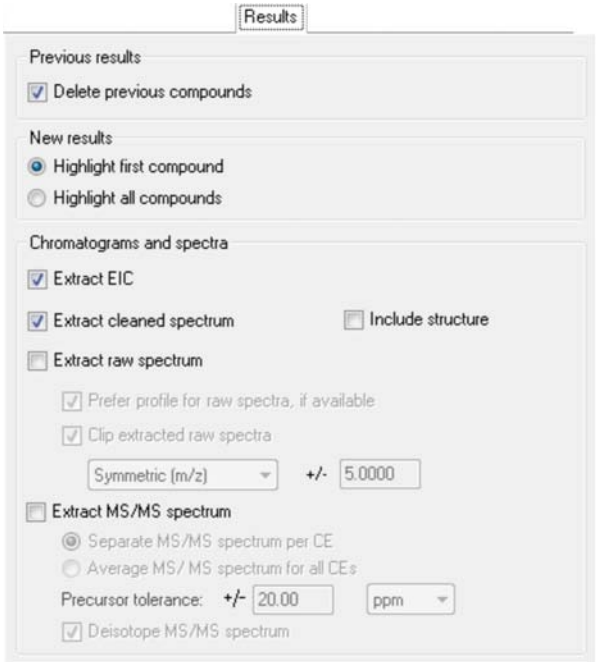
Results tab

f Enter the parameters on the Results tab.

The parameters in this tab specify how the results are saved. This affects the ease reviewing results.

1. Click the **Results** tab.
2. Mark the **Delete previous compounds** check box to delete prior compound results. Clear the **Delete previous compounds** check box when you want to concatenate the Find Compounds by Formula results to the Find by Molecular Feature results and thereby manually review whether the feature was found in both instances.
3. Click **Highlight first compounds** under the *New results* group heading.
4. Mark the **Extract EIC** check box.

5. Mark the **Extract cleaned spectrum** check box. Extracting chromatograms or spectra slows the processing. Once you are comfortable with the results, processing time is reduced by clearing these check boxes.
6. Clear the **Include structure** check box.
7. Clear the **Extract raw spectrum** check box.
8. Clear the **Extract MS/MS spectrum** check box.



Results

Previous results

☑ Delete previous compounds

New results

● Highlight first compound

○ Highlight all compounds

Chromatograms and spectra

☑ Extract EIC

☑ Extract cleaned spectrum   □ Include structure

□ Extract raw spectrum

☑ Prefer profile for raw spectra, if available

☑ Clip extracted raw spectra

Symmetric (m/z)   +/- 5.0000

□ Extract MS/MS spectrum

● Separate MS/MS spectrum per CE

○ Average MS/ MS spectrum for all CEs

Precursor tolerance: +/- 20.00 ppm

☑ Deisotope MS/MS spectrum

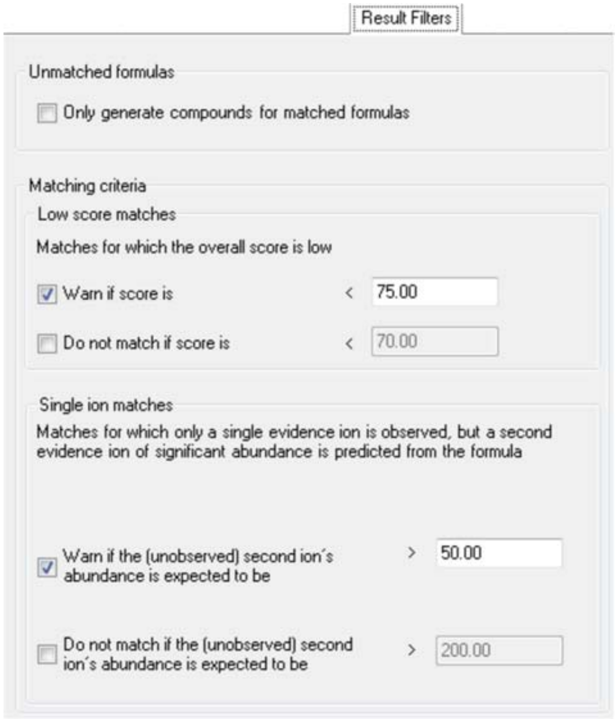
Figure 106 *Results tab in the Find by Formula - Options section*

Result Filters tab

g Enter the parameters on the Result Filters tab.

The parameters in this tab specify whether compounds are generated and/or whether you receive warning notations based on the matching score. Mass Profiler Professional does not need the matching.

1. Click the **Result Filters** tab.
2. Clear the **Only generate compounds for matched formulas** check box.
3. Mark the **Warn if score is** check box.
4. Type `75` for **Warn if score <**.
5. Clear the **Do not match if score is** check box.
6. Mark the **Warn if the (unobserved) second ion's expected abundance is expected to be** check box.
7. Type `50` for **Warn if the (unobserved) second ion's expected abundance is expected to be >**.
8. Clear the **Do not match if the second ion's expected abundance is** check box.



Result Filters

Unmatched formulas

□ Only generate compounds for matched formulas

Matching criteria

Low score matches

Matches for which the overall score is low

☑ Warn if score is < 75.00

□ Do not match if score is < 70.00

Single ion matches

Matches for which only a single evidence ion is observed, but a second evidence ion of significant abundance is predicted from the formula

☑ Warn if the (unobserved) second ion's abundance is expected to be > 50.00

□ Do not match if the (unobserved) second ion's abundance is expected to be > 200.00

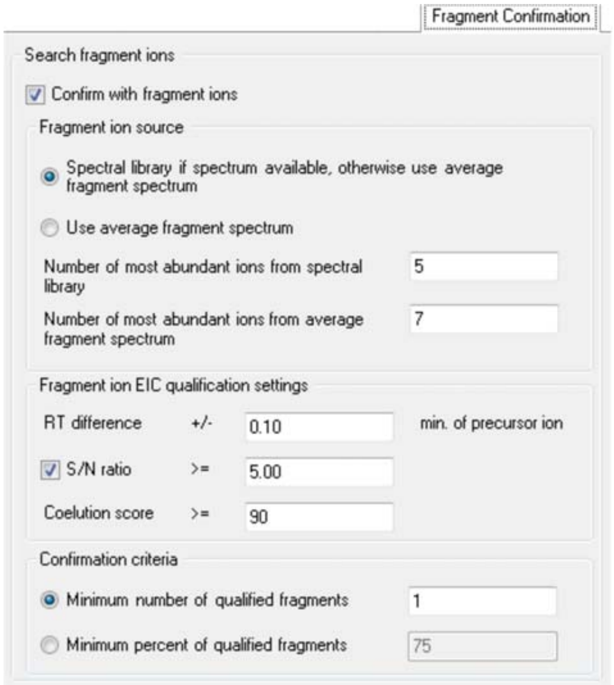
Figure 107 *Result Filters tab in the Find by Formula - Options section*

Fragment Confirmation tab

h Enter the parameters on the Fragment Confirmation tab.

The parameters in this tab specify whether or not to confirm found compounds by comparing the fragment ions in the data file with either the library spectrum or the average fragment spectrum. Fragment confirmation is only possible on data files that are acquired in All Ions MS/MS mode. If you activate this confirmation and the data file was not acquired in All Ions MS/MS mode, then fragment confirmation does not occur.

1. Mark the **Confirm with fragment ions** check box.
2. Click **Spectral library if spectrum available, otherwise use average fragment spectrum**.
3. Type `5` for **Number of most abundant ions from spectral library**.
4. Type `7` for **Number of most abundant ions from average fragment spectrum**.
5. Type `0.10` for **RT difference +/-** min. of precursor ion.
6. Mark the **S/N ratio** check box.
7. Type `5` for **S/N ratio >=**.
8. Type `90` for **Coelution score >=**.
9. Click **Minimum number of qualified fragments**.
10. Type `1` for **Minimum number of qualified fragments**.



Fragment Confirmation

Search fragment ions

☑ Confirm with fragment ions

Fragment ion source

● Spectral library if spectrum available, otherwise use average fragment spectrum

○ Use average fragment spectrum

Number of most abundant ions from spectral library	5
Number of most abundant ions from average fragment spectrum	7

Fragment ion EIC qualification settings

RT difference	+/-	0.10	min. of precursor ion
☑ S/N ratio	>=	5.00	
Coelution score	>=	90	

Confirmation criteria

● Minimum number of qualified fragments	1
○ Minimum percent of qualified fragments	75

Figure 108 Fragment Confirmation tab in the Find by Formula - Options section

3. Enter the parameters in the Find by Formula - Chromatograms section.

In this section, you enter integrator parameters that are applied to the data to extract features for matching to the input formula. This is the second and most critical part of the recursive refinement of the feature finding.

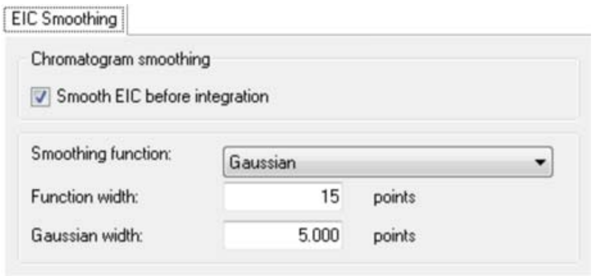
a Click **Find Compounds by Formula > Find by Formula - Chromatograms** in the Method Explorer window. The input options specify the integrator parameters.

EIC Smoothing tab

b Enter the parameters on the EIC Smoothing tab.

The parameters in this tab specify which algorithm to use to smooth the extracted ion chromatogram results.

1. Click the **EIC Smoothing** tab.
2. Mark the **Smooth EIC before integration** check box.
3. Select **Gaussian** from the Smoothing function selection.
4. Type `15` for **Function width** points.
5. Type `5` for **Gaussian width** points.



EIC Smoothing

Chromatogram smoothing

☑ Smooth EIC before integration

Smoothing function:	Gaussian ▾	
Function width:	15	points
Gaussian width:	5.000	points

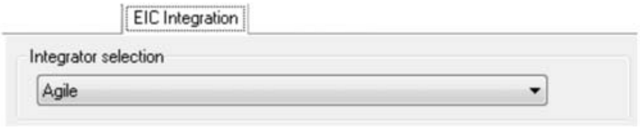
Figure 109 EIC Smoothing tab in the Find by Formula - Chromatograms section

EIC Integration tab

c Enter the parameters on the EIC Integration tab.

The parameters in this tab specify which integrator to use for the data extraction.

1. Click the **EIC Integration** tab.
2. Select **Agile** under the Integrator selection. Agile is the preferred class prediction integrator. No additional user parameters are associated with this integration.



EIC Integration

Integrator selection

Agile ▾

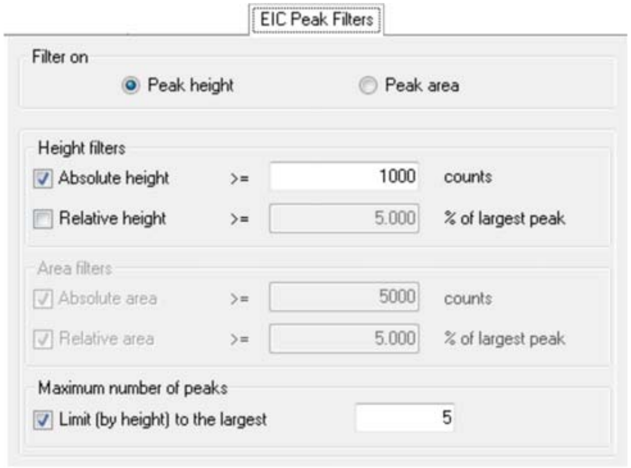
Figure 110 *EIC Integration tab in the Find by Formula - Chromatograms section*

EIC Peak Filters tab

d Enter the parameters on the EIC Peak Filters tab.

The parameters in this tab specify which ions to filter out of the chromatogram integrator results.

1. Click the **EIC Peak Filters** tab.
2. Click **Peak height**.
3. Mark the **Absolute height** check box.
4. Type in a value of `1000` counts for the **Absolute Height**. With targeted feature finding, the minimum **Absolute height** of the feature may be smaller than the absolute height used in the compound filters for untargeted molecular feature extraction.
5. Mark the **Limit (by height) to the largest** check box.
6. Type in a value of `5` for the **Limit (by height) to the largest**. If more than five peaks are found and pass the isotope test and retention times are not used, then the most abundant peaks are reported.



EIC Peak Filters

Filter on

● Peak height ○ Peak area

Height filters

☑ Absolute height >= 1000 counts

□ Relative height >= 5.000 % of largest peak

Area filters

☑ Absolute area >= 5000 counts

☑ Relative area >= 5.000 % of largest peak

Maximum number of peaks

☑ Limit (by height) to the largest 5

Figure 111 *EIC Peak Filters tab in the Find by Formula - Chromatograms section*

4. Enter the parameters in the Find by Formula - Mass Spectra section.

The parameters in the Find by Formula - Mass Spectra section are applied to the data to extract features for matching to the input formula. This is the third and final part of the recursive refinement of the feature finding.

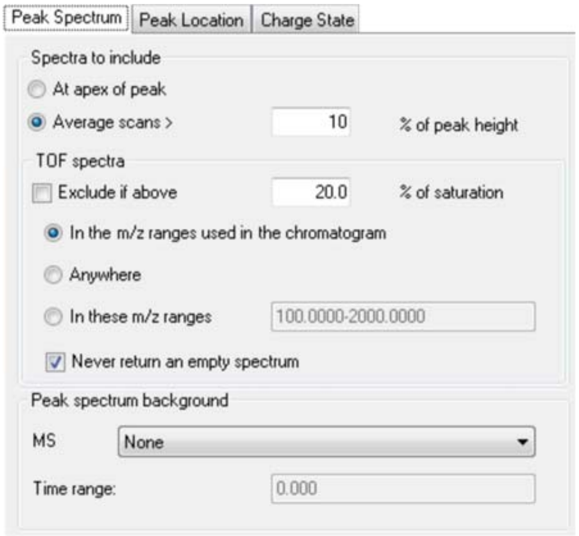
a Click **Find Compounds by Formula > Find by Formula - Mass Spectra** in the Method Explorer window. The input options specify criteria for mass spectra to include in the feature processing.

Peak Spectrum tab

b Enter the parameters on the Peak Spectrum tab.

The parameters in this tab specify which spectra from the extracted ion chromatograms to include in the feature processing and whether to perform background subtraction on the spectra.

1. Click the **Peak Spectrum** tab
2. Click **Average scans >**.
3. Type `10` for the % of peak height. Averaging scans provides mass accuracy.
4. Clear the **Exclude if above X% of saturation** under the *TOF spectra* group heading. If this check box is marked, any spectrum containing a peak within the given percentage of being saturated is excluded from processing for any compound feature.
5. Select **None** for **MS** under the Peak spectrum background group heading.



Spectra to include

○ At apex of peak

● Average scans > 10 % of peak height

TOF spectra

□ Exclude if above 20.0 % of saturation

● In the m/z ranges used in the chromatogram

○ Anywhere

○ In these m/z ranges 100.0000-2000.0000

☑ Never return an empty spectrum

Peak spectrum background

MS None

Time range: 0.000

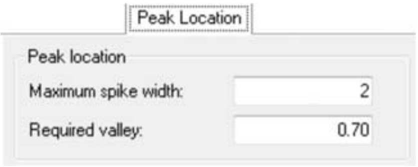
Figure 112 *Peak Spectrum tab in the Find by Formula - Mass Spectra section*

Peak Location tab

c Enter the parameters on the Peak Location tab.

The parameters in this tab specify the *m/z* values in a spectrum that are considered peaks. These parameters are only applicable to profile data files. They are not applicable to centroid collected data files and may be left at the defaults.

1. Click the **Peak Location** tab.
2. Type the default of `2` for **Maximum spike width**.
3. Type the default of `0.70` for **Required valley**.


Peak Location

Peak location

Maximum spike width: 2

Required valley: 0.70

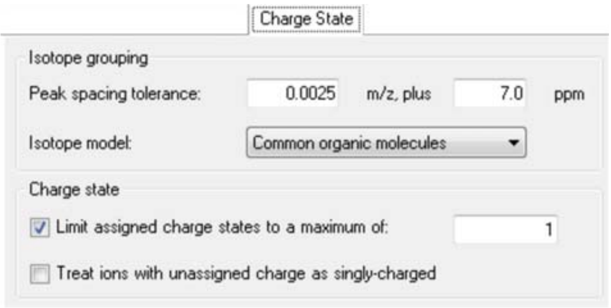
***Figure 113** Peak Location tab in the Find by Formula - Mass Spectra section*

Charge State tab

d Enter the parameters on the Charge State tab.

The parameters in this tab specify isotope grouping tolerances and charge state limits. Set the maximum charge state to one (1). Adjustments to the grouping model can change the compound results.

1. Click the **Charge State** tab.
2. Type `0.0025` for *m/z* and `7.0` ppm for the **Peak spacing tolerance**.
3. Select **Common organic molecules** for the **Isotope model**. You select **Unbiased** if the compounds are known to contain metals.
4. Mark the **Limit assigned charge state to a maximum of** check box.
5. Type `1` for the **Limit assigned charge state to a maximum of**. This parameter should match the value typed into the Charge state range in the “Positive Ions tab” on page 105 and “Negative Ions tab” on page 106 in the “Enter the parameters in the Find by Formula - Options section.”.
6. Clear the **Treat ions with unassigned charge as singly-charged** check box.


Charge State

Isotope grouping

Peak spacing tolerance: 0.0025 m/z, plus 7.0 ppm

Isotope model: Common organic molecules

Charge state

☑ Limit assigned charge states to a maximum of: 1

☐ Treat ions with unassigned charge as singly-charged

***Figure 114** Charge State tab in the Find by Formula - Mass Spectra section*

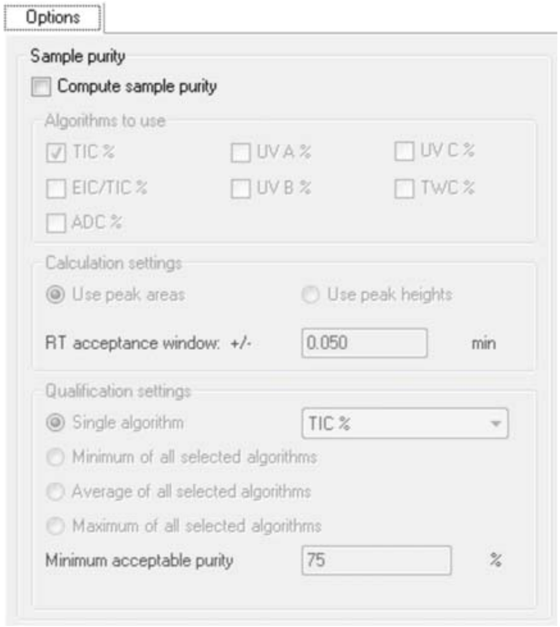
5. Turn off the sample purity calculations in the **Find by Formula - Sample Purity** section.

The Find by Formula - Sample Purity is not used in this example or typical metabolomics analyses.

a Click **Find Compounds by Formula > Find by Formula - Sample Purity** in the Method Explorer window. The input options specify criteria for mass spectra to include in the feature processing.

b Turn off sample purity calculations on the Options tab.

1. Click the **Options** tab.
2. Clear the **Compute sample purity** check box. If this check box is cleared, then sample purity is not calculated. All other options on this tab are unavailable and the entries in the remaining tabs are not considered.



Options

Sample purity

□ Compute sample purity

Algorithms to use

☑ TIC %   □ UV A %   □ UV C %

□ EIC/TIC %   □ UV B %   □ TWC %

□ ADC %

Calculation settings

● Use peak areas   ○ Use peak heights

RT acceptance window: +/- 0.050 min

Qualification settings

● Single algorithm   TIC %

○ Minimum of all selected algorithms

○ Average of all selected algorithms

○ Maximum of all selected algorithms

Minimum acceptable purity 75 %

Figure 115 *Options tab in the Find by Formula - Sample Purity section*

Save your Find Compounds by Formula method

After you have edited your FbF method it is recommended you save the method using a name different from the name you previously used for your MFE method so that you can readily reprocess your data or new data without having to edit the Worklist Automation actions.

a Click **Method > Save As**.

b Select the folder and type a method name in the **Save Method** dialog box. Change the text `MFE` at the end of the method file name used in “Save your Find Compounds by Molecular Feature method” on page 53 to `FbF`.

c Click **Save**.

Set the Export CEF Options

Export CEF Options specifies where MassHunter DA Reprocessor stores the resulting CEF feature files and whether the files replace or overwrite any prior files.

1. Open the Method Editor for exporting CEF options.

a Click **Export** from within the **Method Explorer** window.

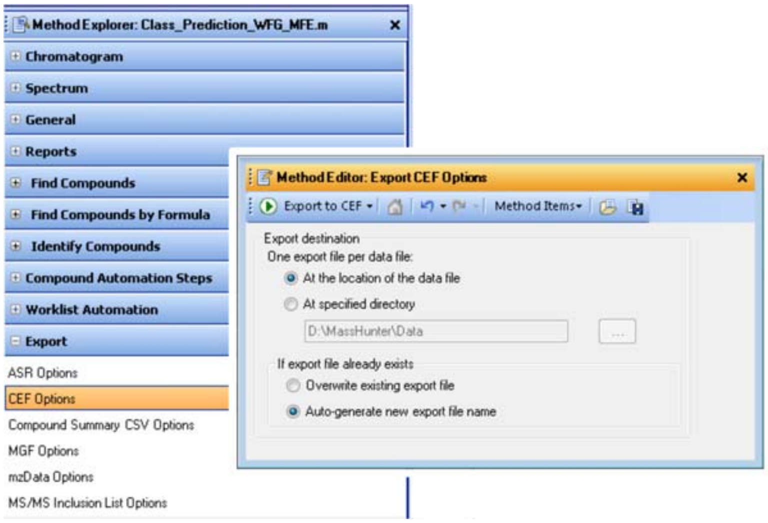
b Click **CEF Options**.

2. Enter the export destination settings for your method.

a Click **At the location of the data file**.

b Click **Auto-generate new export file name**.

c Save your method. Click the save method icon or click **Method > Save**.


Method Explorer: Class_Prediction_WFG_MFE.m
Chromatogram
Spectrum
General
Reports
Find Compounds
Find Compounds by Formula
Identify Compounds
Compound Automation Steps
Worklist Automation
Export
ASR Options
CEF Options
Compound Summary CSV Options
MGF Options
mzData Options
MS/MS Inclusion List Options
Method Editor: Export CEF Options
Export to CEF
Method Items

Export destination
One export file per data file:
● At the location of the data file
○ At specified directory
D:\MassHunter\Data

If export file already exists
○ Overwrite existing export file
● Auto-generate new export file name

Figure 116 *Export CEF Options for use with DA Reprocessor*

Enable the method to run in MassHunter DA Reprocessor

MassHunter software can most efficiently perform computationally intensive tasks, such as feature finding, on multiple data files using MassHunter DA Reprocessor. The following steps enable your method to run using DA Reprocessor.

1. Open the Method Editor to assign actions to run from the worklist.

a Click **Worklist Automation** from within the **Method Explorer** window.

b Click **Worklist Actions** in the **Worklist Automation** section.

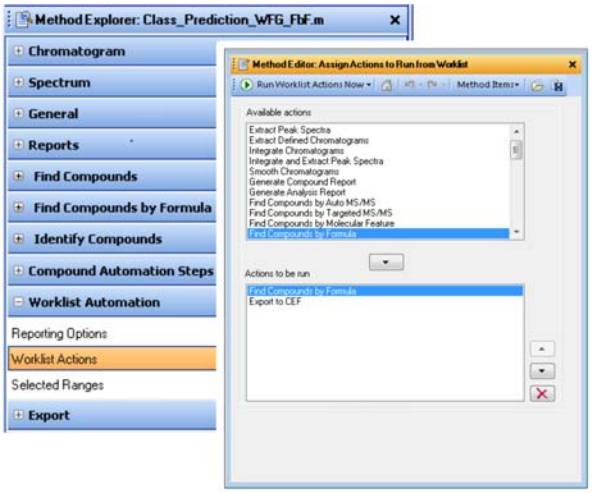
2. Replace the **MFE** action with **FbF** action.

a Double-click the **Find Compounds by Molecular Feature** action in the **Actions to be run** list. The action is automatically removed from the **Actions to be run** list. As an alternate to the double-click, you can click on the action and then click the delete icon ✕.

b Double-click the **Find Compounds by Formula** action in the **Available actions** list. The action is automatically added to the **Actions to be run** list. As an alternate to the double-click, you can click the action and then click the down arrow button ▾ to add the action to the **Actions to be run** list.

c Move the actions in the **Available actions** list so that the **Export to CEF** action is listed after the **Find Compounds by Formula** action as shown in Figure 117 on page 116.

d Save your method. Click the save method icon or click **Method > Save**.



Available actions

- Extract Peak Spectra
- Extract Defined Chromatograms
- Integrate Chromatograms
- Integrate and Extract Peak Spectra
- Smooth Chromatograms
- Generate Compound Report
- Generate Analysis Report
- Find Compounds by Auto MS/MS
- Find Compounds by Targeted MS/MS
- Find Compounds by Molecular Feature
- Find Compounds by Formula

Actions to be run

- Find Compounds by Formula
- Export to CEF

Figure 117 *Assign Actions to Run from Worklist for use with DA Reprocessor*

Confirm the FbF method on a single data file

Metabolomics analyses involves the analysis of a large number of sample files with each sample containing a large number of compounds. Find Compounds by Formula is therefore run on the entire sample data set using MassHunter DA Reprocessor. However, before the entire sample set is run in MassHunter DA Reprocessor, a single file is processed within MassHunter Qualitative Analysis to verify the new parameters.

1. Find Compounds by Formula on a single sample.

a Click **File > Open Data File**.

b Click a single data file in the **Open Data File** dialog box.

c Click **Open**.

d Click **Actions > Find Compounds by Formula**, or click the Find Compounds by Formula button Find Compounds by Formula in the **Method Editor: Find Compounds by Formula** section in the Method Editor window. Feature extraction begins immediately and the progress is shown in an **Operation in Progress** status box (see Figure 36 on page 56).

If no data file is open, or an inappropriate data file is open, a message box appears as shown in Figure 37 on page 56. Click **OK** and open a single data file.

2. Display and review the Compounds List.

When the FbF method finishes processing the data file, MassHunter Qualitative Analysis displays the results in several windows. You may review and arrange the results to meet your preferences. If the window is not displayed, click **View > Compounds List**.

The options for reviewing the are identical to those described in "Confirm the MFE method on a single data file", step 2 - "Display and review the Compound List after running MassHunter DA Reprocessor." on page 60.

3. *Optional* - Export the results for the single sample to a CEF file.

This step is optional and identical to step 4 - "Optional - Export the results for the single sample to a CEF file." on page 58.

Find compounds using DA Reprocessor

Class prediction involves applying your method to a large number of sample files whereby each sample file may contain a large number of compounds. MassHunter Qualitative Analysis can be used to process all of your data sets. However, MassHunter DA Reprocessor provides a more efficient and automated means to run your MassHunter Qualitative Analysis method on multiple sample files. Therefore your method is run on the entire sample data set using DA Reprocessor.

1. Find Compounds by Formula using MassHunter DA Reprocessor.

Follow the same procedure presented in "Find compounds using DA Reprocessor" on page 59 to perform targeted feature finding using FbF.

2. Display and review the after running MassHunter DA Reprocessor.

a Return to MassHunter Qualitative Analysis. If you closed the MassHunter Qualitative Analysis program, do the following:
- Click **Start > All Programs > Agilent > MassHunter Workstation > Qualitative Analysis B.06.00**.
- Click **Cancel** when the **Open Data File** dialog box opens.

b Click **File > Close All** to close the open data files. Do not save any results.

c Click **File > Import Compound** to open one of the CEF files that contains the molecular feature results of Find Compounds by Formula.

Note: Because of the large number of features in a typical sample file, it is recommended to open only one file at a time to review the results. Close the open file and then open the next file.

d Follow the same procedure presented in "Confirm the MFE method on a single data file", step 2 - "Display and review the Compound List after running MassHunter DA Reprocessor." on page 60 to display and review the mass spectral results.

Recursive feature finding using MassHunter Profinder

The following examples use MassHunter Profinder B.06.00 running on 64-bit Windows 7 Professional. If you already performed a recursive feature finding using Qualitative Analysis skip to "Divide the sample data into training and validation data sets" on page 124.

Finding the features in your sample data files using Profinder involves one self-directed wizard that involves five steps as shown in Figure 118 on page 118.

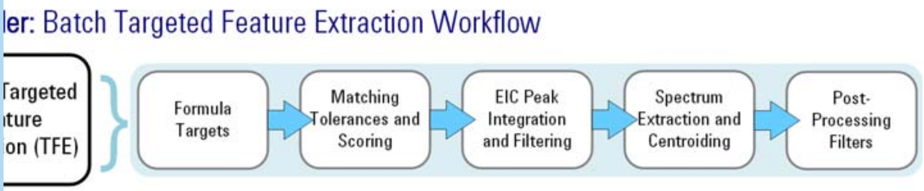

ler: Batch Targeted Feature Extraction Workflow
Targeted
ature
on (TFE)
Formula Targets
Matching Tolerances and Scoring
EIC Peak Integration and Filtering
Spectrum Extraction and Centroiding
Post-Processing Filters

***Figure 118** Steps to find targeted features using Profinder*

The parameters recommended in the following steps are part of the Batch Targeted Feature Extraction workflow wizard in Profinder. See *Agilent G3835AA MassHunter Profinder Software - Quick Start Guide* for additional information on using Profinder.

1. Start MassHunter Profinder software.

MassHunter Profinder a software tool used to perform the function of finding molecular features in TOF and Q-TOF data files. After the molecular features are found they are imported into Mass Profiler Professional for statistical analysis. Feature finding is an essential prerequisite to using Mass Profiler Professional.

a Double-click the Profinder icon located on the desktop,

Click **Start > All Programs > Agilent > MassHunter Workstation > Profinder B.06.00**.

b Click the **File > Add/Remove Sample Files** or Add/Remove Sample Files... in the toolbar to begin the workflow.

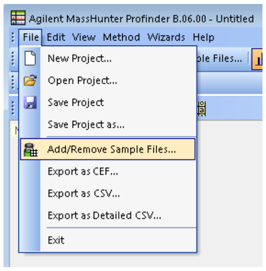


***Figure 119** Add sample files to begin feature finding in Profinder*

2. Add all of the sample data files to Profinder.

a Click **Add file(s)** in the **Add/Remove Sample Files** dialog box.

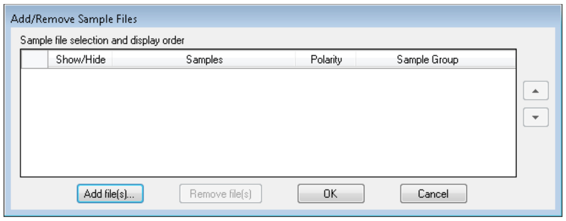
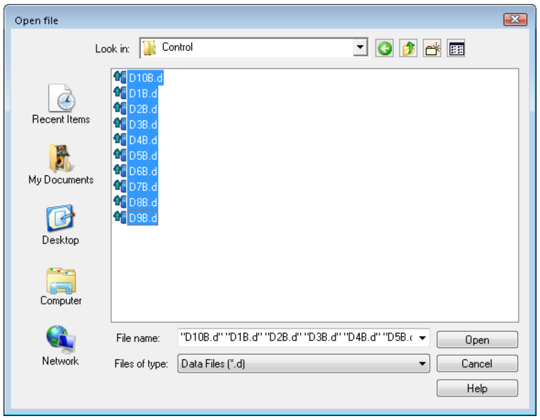


***Figure 120** Add/Remove Sample Files dialog box*

b Navigate to the folder containing your raw sample data files and in the **Open File** dialog box.

c Select your raw sample data files in the **Open File** dialog box.

d Click **Open**.



Look in: Control

D10B.d
D1B.d
D2B.d
D3B.d
D4B.d
D5B.d
D6B.d
D7B.d
D8B.d
D9B.d

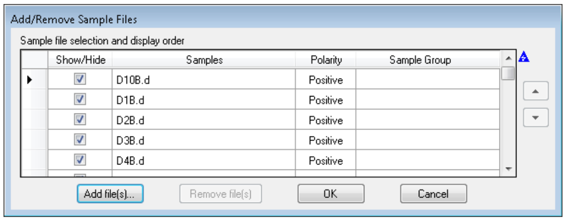
File name: "D10B.d" "D1B.d" "D2B.d" "D3B.d" "D4B.d" "D5B.(

Files of type: Data Files (*.d)

Figure 121 *Add/remove Sample Files dialog box*

e Repeat steps a through d if your sample data files reside in multiple folders.

f Click **OK** in the **Add/Remove Sample Files** dialog box when all of your sample data files are selected.



Show/Hide	Samples	Polarity	Sample Group
☑	D10B.d	Positive	
☑	D1B.d	Positive	
☑	D2B.d	Positive	
☑	D3B.d	Positive	
☑	D4B.d	Positive	

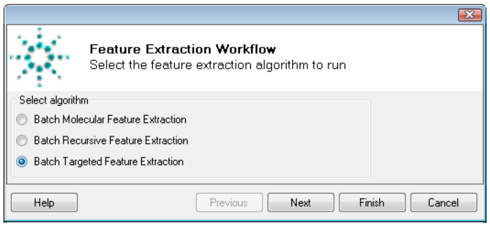
Figure 122 *Samples added to the Add/remove Sample Files dialog box*

3. Select the Batch Targeted Feature Extraction workflow.

Begin the Batch Molecular Feature Extraction (MFE) workflow.

a Click **Batch Molecular Feature Extraction**.

b Click **Next**.


Feature Extraction Workflow
Select the feature extraction algorithm to run
Help
Previous
Next
Finish
Cancel

Select algorithm

○ Batch Molecular Feature Extraction

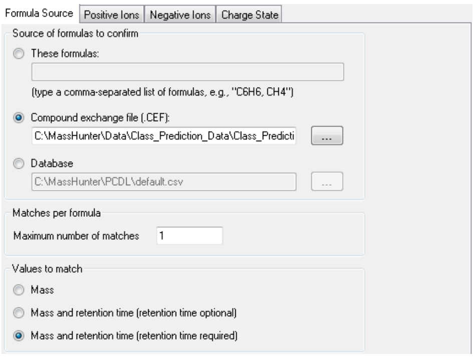
○ Batch Recursive Feature Extraction

● Batch Targeted Feature Extraction

Figure 123 *Select Batch Molecular Feature Extraction*

4. Enter the formula targets parameters in **(Step 1 of 5)** of the TFE workflow.

a Enter the parameters in the **Formula Source** tab.



Source of formulas to confirm

○ These formulas:

(type a comma-separated list of formulas, e.g., "C6H6, CH4")

● Compound exchange file (.CEF):

C:\MassHunter\Data\Class_Prediction_Data\Class_Predicti ...

○ Database

C:\MassHunter\PCDL\default.csv ...

Matches per formula

Maximum number of matches 1

Values to match

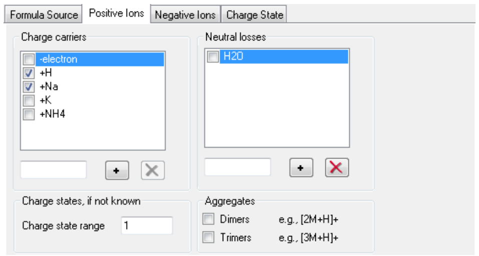
○ Mass

○ Mass and retention time (retention time optional)

● Mass and retention time (retention time required)

Figure 124 *Formula Targets Step 1 of 5 - Formula Source tab*

b Enter the parameters in the **Positive Ions** tab.



Charge carriers

□ -electron
☑ +H
☑ +Na
□ +K
□ +NH4

Neutral losses

□ H2O

Charge states, if not known

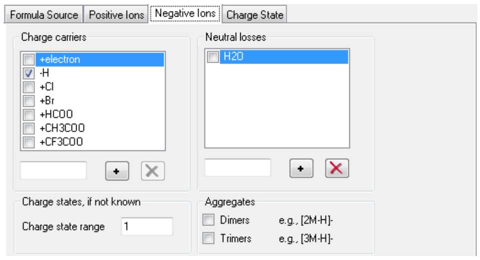
Charge state range 1

Aggregates

□ Dimers e.g., [2M+H]+
□ Trimers e.g., [3M+H]+

Figure 125 *Formula Targets Step 1 of 5 - Positive Ions tab*

c Enter the parameters in the **Negative Ions** tab.



Charge carriers

□ +electron
☑ -H
□ +Cl
□ +Br
□ +HCOO
□ +CH3COO
□ +CF3COO

Neutral losses

□ H2O

Charge states, if not known

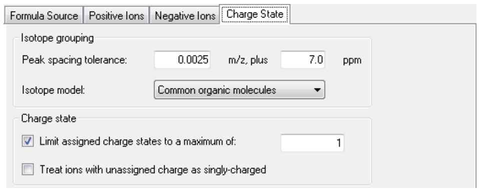
Charge state range 1

Aggregates

□ Dimers e.g., [2M-H]-
□ Trimers e.g., [3M-H]-

Figure 126 *Formula Targets Step 1 of 5 - Negative Ions tab*

d Enter the parameters in the **Charge State** tab.



Isotope grouping

Peak spacing tolerance: 0.0025 m/z, plus 7.0 ppm

Isotope model: Common organic molecules

Charge state

☑ Limit assigned charge states to a maximum of: 1

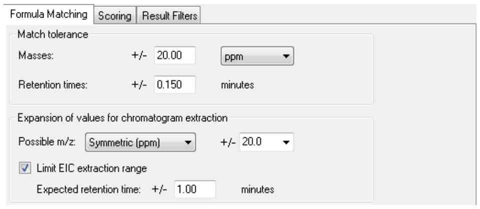
□ Treat ions with unassigned charge as singly-charged

Figure 127 *Formula Targets Step 1 of 5 - Charge State tab*

e Click **Next**.

5. Enter the matching tolerance and scoring parameters in **(Step 2 of 5)** of the TFE workflow.

a Enter the parameters in the **Formula Matching** tab.



Match tolerance

Masses: +/- 20.00 ppm

Retention times: +/- 0.150 minutes

Expansion of values for chromatogram extraction

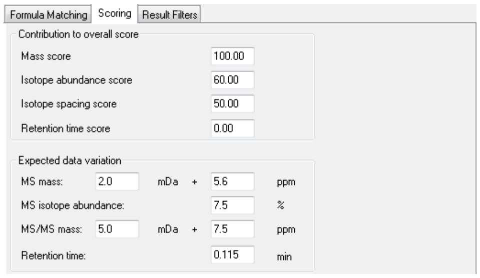
Possible m/z: Symmetric (ppm) +/- 20.0

☑ Limit EIC extraction range

Expected retention time: +/- 1.00 minutes

Figure 128 Matching Tolerances and Scoring Step 2 of 5 - Formula Matching tab

b Enter the parameters in the **Scoring** tab.



Contribution to overall score

Mass score 100.00

Isotope abundance score 60.00

Isotope spacing score 50.00

Retention time score 0.00

Expected data variation

MS mass: 2.0 mDa + 5.6 ppm

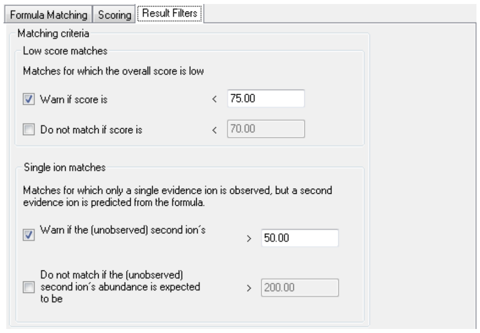
MS isotope abundance: 7.5 %

MS/MS mass: 5.0 mDa + 7.5 ppm

Retention time: 0.115 min

Figure 129 Matching Tolerances and Scoring Step 2 of 5 - Scoring tab

c Enter the parameters in the **Results Filters** tab.



Matching criteria

Low score matches

Matches for which the overall score is low

☑ Warn if score is < 75.00

☐ Do not match if score is < 70.00

Single ion matches

Matches for which only a single evidence ion is observed, but a second evidence ion is predicted from the formula.

☑ Warn if the (unobserved) second ion's > 50.00

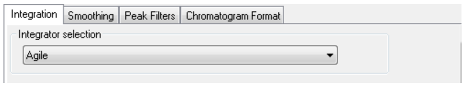
☐ Do not match if the (unobserved) second ion's abundance is expected to be > 200.00

Figure 130 Matching Tolerances and Scoring Step 2 of 5 - Results Filters tab

d Click **Next**.

6. Enter the EIC peak integration and filtering parameters in **(Step 3 of 5)** of the TFE workflow.

a Enter the parameters in the **Integration** tab.

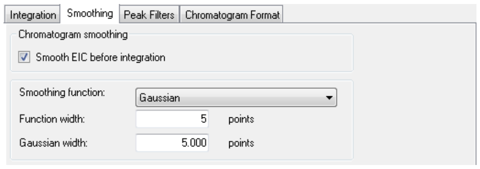


Integrator selection

Agile

Figure 131 EIC Peak Integration and Filtering Step 3 of 5 - Integration tab

b Enter the parameters in the **Smoothing** tab.



Integration | Smoothing | Peak Filters | Chromatogram Format

Chromatogram smoothing

☑ Smooth EIC before integration

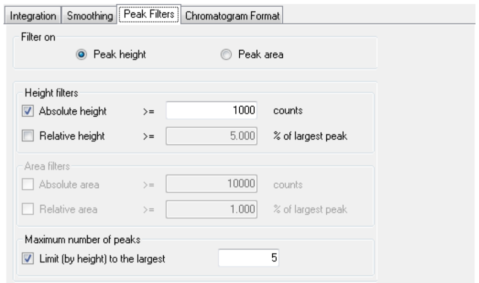
Smoothing function: Gaussian

Function width: 5 points

Gaussian width: 5.000 points

Figure 132 *EIC Peak Integration and Filtering Step 3 of 5 - Smoothing tab*

c Enter the parameters in the **Peak Filters** tab.



Integration | Smoothing | Peak Filters | Chromatogram Format

Filter on

● Peak height ○ Peak area

Height filters

☑ Absolute height >= 1000 counts

□ Relative height >= 5.000 % of largest peak

Area filters

□ Absolute area >= 10000 counts

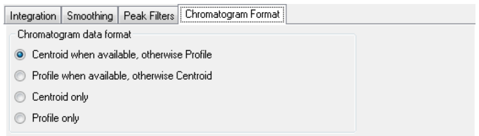
□ Relative area >= 1.000 % of largest peak

Maximum number of peaks

☑ Limit (by height) to the largest 5

Figure 133 *EIC Peak Integration and Filtering Step 3 of 5 - Peak Filters tab*

d Enter the parameters in the **Chromatogram Format** tab.



Integration | Smoothing | Peak Filters | Chromatogram Format

Chromatogram data format

● Centroid when available, otherwise Profile

○ Profile when available, otherwise Centroid

○ Centroid only

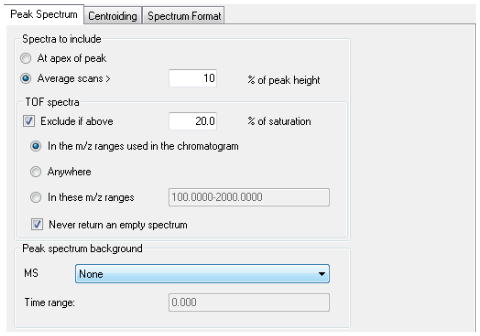
○ Profile only

Figure 134 *EIC Peak Integration and Filtering Step 3 of 5 - Chromatogram Format tab*

e Click **Next**.

7. Enter the spectrum extraction and centroiding parameters in **(Step 4 of 5)** of the TFE workflow.

a Enter the parameters in the **Peak Spectrum** tab.



Peak Spectrum | Centroiding | Spectrum Format

Spectra to include

○ At apex of peak

● Average scans > 10 % of peak height

TOF spectra

☑ Exclude if above 20.0 % of saturation

● In the m/z ranges used in the chromatogram

○ Anywhere

○ In these m/z ranges 100.0000-2000.0000

☑ Never return an empty spectrum

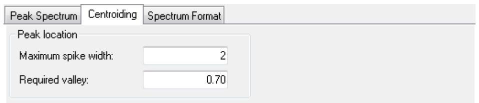
Peak spectrum background

MS None

Time range: 0.000

Figure 135 *Spectrum Extraction and Centroiding Step 4 of 5 - Peak Spectrum tab*

b Enter the parameters in the **Centroiding** tab.



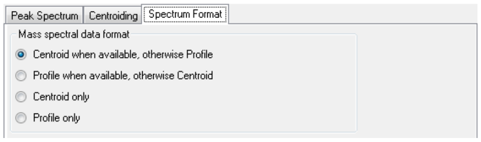
Peak location

Maximum spike width: 2

Required valley: 0.70

Figure 136 *Spectrum Extraction and Centroiding Step 4 of 5 - Centroiding tab*

c Enter the parameters in the **Spectrum Format** tab.



Mass spectral data format

● Centroid when available, otherwise Profile

○ Profile when available, otherwise Centroid

○ Centroid only

○ Profile only

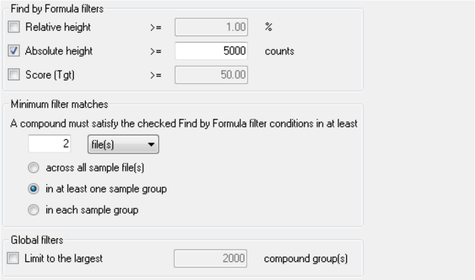
Figure 137 *Spectrum Extraction and Centroiding Step 4 of 5 - Spectrum Format tab*

d Click **Next**.

8. Enter the parameters for the post-processing filters in **(Step 5 of 5)** of the TFE workflow.

a Enter the parameters for the post-processing filters.

b Click **Finish**. Feature finding in the sample data files begins immediately.



Find by Formula filters

□ Relative height	>=	1.00	%
☑ Absolute height	>=	5000	counts
□ Score (Tgt)	>=	50.00	

Minimum filter matches

A compound must satisfy the checked Find by Formula filter conditions in at least

2 file(s)

○ across all sample file(s)

● in at least one sample group

○ in each sample group

Global filters

□ Limit to the largest 2000 compound group(s)

Figure 138 *Post-Processing filters Step 5 of 5*

9. Save your method.

a Click **Method > Save As** to save your method.

b Navigate to the appropriate folder.

c Enter your file name.

d Click **Save**.

Divide the sample data into training and validation data sets

Training sample data files are used to build your class prediction model. Validation data files are used as a final quality inspection of your class prediction model. Provided that you have sufficient replicates for each of your classes, you begin training by setting aside a subset of your sample data, typically one or more sample data files representing each of the classifications, to use as a validation data set. The remaining majority of the data, the training data set, is used to train your prediction model.

In this workflow the recursive feature data files (CEF files) generated by Qualitative Analysis/DA Reprocessor or Profinder are copied into separate folders. Placing the data files into separate folders prevents potential confusion among the CEF files used for training versus validation and, more importantly, treats the validation files identical with how new samples files are analyzed.

Note: If you plan to use your class prediction model with data already collected instead of with data to be collected in the future, an alternate process for managing training versus validation data involves retaining all of the CEF files in the same directory and importing all of the files into a single MPP experiment so that all of the features across both the training and validation data files are subject to the same normalization and alignment (see the initial experiment grouping described in "Group samples based on the independent variables and replicate structure of your experiment in the MS Experiment Creation Wizard (Step 6 of 11)." on page 74). When you import all of the sample data files into a single experiment, it is recommended to create a new Parameter Name (i.e., "Class Prediction") and designate at least one sample representing each of the experimental classifications a value of "Validation" and the remaining samples a value of "Training." Use the samples designated "Validation" in the "Class Prediction" parameter to recreate you differential analysis and build your class prediction model.

Skip ahead to "Recreate your differential analysis using the recursive features" on page 125 if you are importing all of your sample data into one experiment.

1. Create folders for your training and validation CEF files.

a Click **Start > All Programs > Accessories > Windows Explorer** to open Windows Explorer.

b Navigate to **C:\MassHunter\Data**.

c Create a new folder named `Class_Prediction_Data`.

d Navigate to your new folder *C:\MassHunter\Data\Class_Prediction_Data*.

e Create a new folder named `Training_Sample_Data`.

f Create a new folder named `Validation_Sample_Data`.

2. Move most of the replicate CEF files to the training folder.

Move most of the replicate CEF files for each condition to the training data folder **C:\MassHunter\Data\Class_Prediction_Data\Training_Sample_Data**. For the example files ("Features of the example experiment" on page 21), nine of the ten replicate CEF files for each condition represent the training data set.

All of the data files *except* D2B.cef, D13B.cef, D24B.cef, and D35B.cef from "Features of the example experiment" on page 21 are the training sample files.

3. Move at least one replicate CEF file to the validation folder.

Move one CEF file for each condition to the validation data folder *C:\MassHunter\Data\Class_Prediction_Data\Validation_Sample_Data*. For the example files one CEF file representing each condition represents the validation data set.

Data files D2B.cef, D13B.cef, D24B.cef, and D35B.cef from "Features of the example experiment" on page 21 are the validation sample files in this example.

Import and organize your recursive data

Recreate your project setup, experiment up, and import filters following the procedures presented in "Create a new project and experiment" on page 68 through "Filter, align, and normalize the sample data" then return to the next step of this workflow.

Note: If you plan to use your class prediction model with data already collected, instead of with data to be collected in the future, an alternate process is to create an experiment with all of the sample data and use grouping to identify the *training* versus the *validation* data as described in "Order and group the sample data files" on page 73.

Recreate your differential analysis using the recursive features

Recreate your differential analysis using your recursive features following the procedures presented in "Perform a differential analysis" on page 84 then return to the next step of this workflow.

The number of entities during your analysis using recursively found features is smaller than the number features in your initial differential analysis using untargeted features. In both differential analyses the number of features decreases as you progress through the steps of the *Analysis: Significance Testing and Fold Change workflow*. A comparison using this example data is shown in Table 2. The features in the "Training data" column began with recursive features from Filter by Frequency entities from the "All sample data" column.

Table 2 *Feature count decreases as the data is process during the Analysis: Significance Testing and Fold Change workflow.*

	All sample data files, untargeted features	Training data files, recursive features
Step 1 of 8: Summary Report	3,763	1,138
Step 3 of 8: Filter Flags	2,885	1,072
Step 4 of 8: Filter by Frequency *these entities were exported for recursion*	907	925
Step 6 of 8: Significance Analysis	793	835
Step 7 of 8: Fold Change	738	706

The compound alignment values you enter for the retention time window and the mass window in "Select and enter the retention time and mass alignment parameters in the MS Experiment Creation Wizard (Step 8 of 11)." on page 79 determine the number of unique features at the beginning of your differential analysis.

Save the project containing your recursive analysis

If you did not save your analysis at the end of "Perform a differential analysis", save your analysis at this time.

a Click **Project > Export Project**.

b Mark the check box next to the experiment you wish to save.

c Click **OK**.

d Select or create the file folder.

e Type the **File name**.

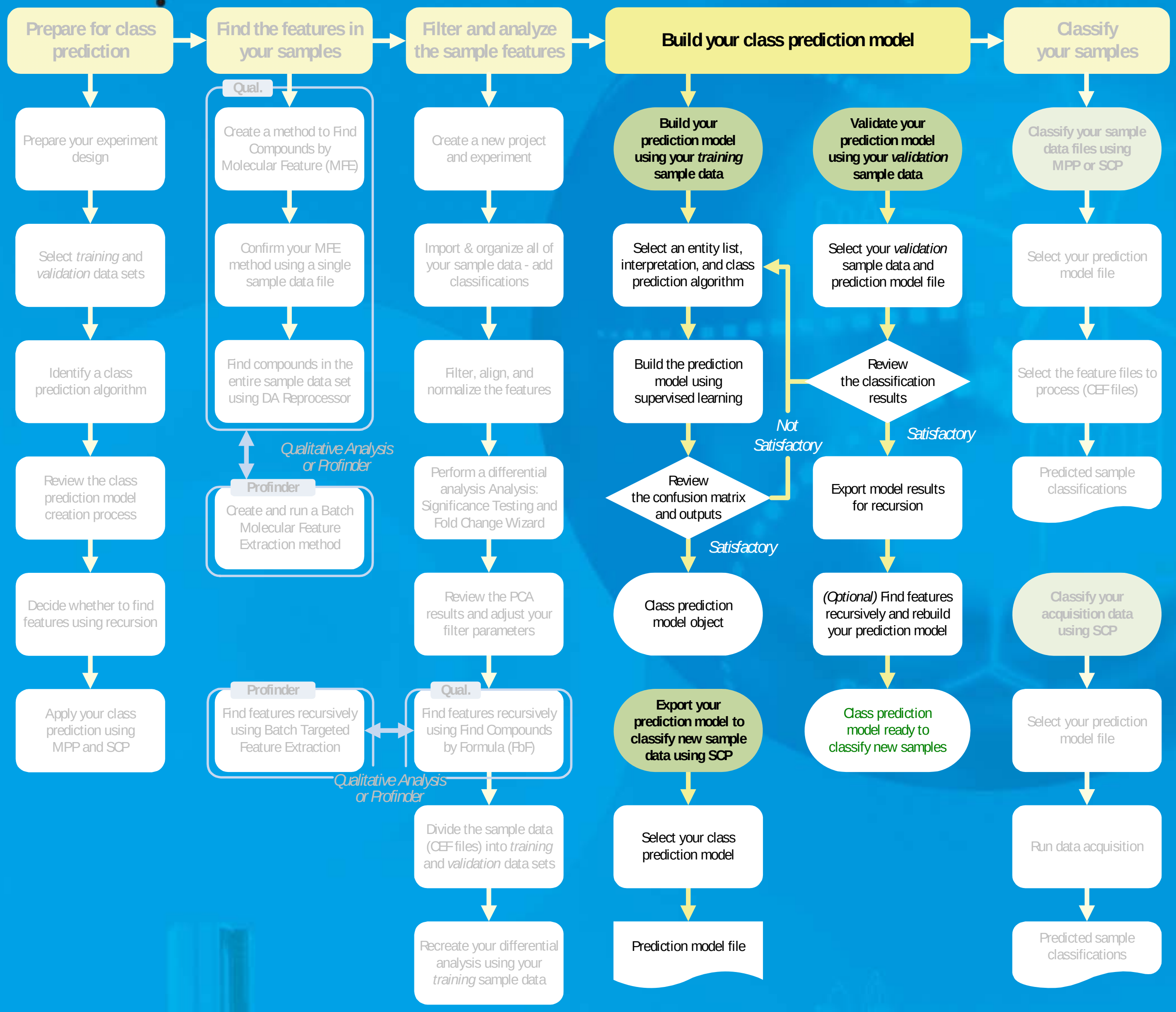
f Click **Save**.

Next step...

You have now completed the third step of the class prediction workflow. In the next workflow step you build and validate your prediction model using Mass Profiler Professional.

Build your class prediction model

Build your prediction model using the training data set and one or more class prediction algorithms. When you obtain satisfactory results, validate your prediction model using the validation data set.



Prepare for class prediction

Prepare your experiment design

Select *training* and *validation* data sets

Identify a class prediction algorithm

Review the class prediction model creation process

Decide whether to find features using recursion

Apply your class prediction using MPP and SCP

Find the features in your samples

Qual.

Create a method to Find Compounds by Molecular Feature (MFE)

Confirm your MFE method using a single sample data file

Find compounds in the entire sample data set using DA Reprocessor

Qualitative Analysis or Profinder

Profinder

Create and run a Batch Molecular Feature Extraction method

Profinder

Find features recursively using Batch Targeted Feature Extraction

Qualitative Analysis or Profinder

Filter and analyze the sample features

Create a new project and experiment

Import & organize all of your sample data - add classifications

Filter, align, and normalize the features

Perform a differential analysis Analysis: Significance Testing and Fold Change Wizard

Review the PCA results and adjust your filter parameters

Qual.

Find features recursively using Find Compounds by Formula (FbF)

Divide the sample data (CEF files) into *training* and *validation* data sets

Recreate your differential analysis using your *training* sample data

Build your class prediction model

Build your prediction model using your *training* sample data

Select an entity list, interpretation, and class prediction algorithm

Build the prediction model using supervised learning

Review the confusion matrix and outputs

Not Satisfactory

Satisfactory

Class prediction model object

Export your prediction model to classify new sample data using SCP

Select your class prediction model

Prediction model file

Validate your prediction model using your *validation* sample data

Select your *validation* sample data and prediction model file

Review the classification results

Satisfactory

Export model results for recursion

(Optional) Find features recursively and rebuild your prediction model

Class prediction model ready to classify new samples

Classify your samples

Classify your sample data files using MPP or SCP

Select your prediction model file

Select the feature files to process (CEF files)

Predicted sample classifications

Classify your acquisition data using SCP

Select your prediction model file

Run data acquisition

Predicted sample classifications

Agilent Technologies

Build your prediction model

After you collect replicate sample data, find the features in all of your sample data files, and create a differential analysis that is able to differentiate your training sample data into the known classifications, you can build your prediction model. Building a prediction model begins with the assumption that you have created a successful initial differential analysis and respective entity lists as described in “Filter and analyze the sample features” on page 67.

Your prediction model can be used to predict functional classes like diseases and conditions from the abundance profile of the entities in your sample data. For example, you can predict the class or parameter of a sample, identify signatures that discriminate well among classes, and identify samples that are outliers and provide a quality control to your sampling.

Select an interpretation, entity list, and class prediction algorithm

Your prediction model is built based upon the selection of a non-averaged interpretation and the selection of an appropriately filtered entity list. The interpretation you select provides the model with the known classifications and whether to average the feature intensity values across replicates or normalize the feature intensity values within a sample.

The entity list you select provides the model with the initial quality control parameters with which to filter the features, leaving the features that contribute the most change among your classifications.

Averaged and non-averaged interpretations

You can choose whether or not to select an interpretation that uses averaging across replicates within each condition (classification) for your experiment. In this workflow guide the class prediction model is created using a non-averaged interpretation, where the intensity of each feature within a data file is normalized among all of the features within the same data file instead of averaging the features across replicate data files.

Averaged Interpretation: Average the feature intensities across replicate data files. The mean intensity value for each entity across the replicates is used for analysis and visualization, and the interpretation is simply listed by its parameter name. Use an averaged interpretation if you plan to use your class prediction model with data that is already collected per your experiment design.

Non-averaged Interpretation: Normalize the feature intensities within a sample to all of the features within the same sample instead of averaging the feature intensities across replicate data files. The normalized intensity value for each entity within a sample is used for analysis and visualization, and the interpretation is listed by its parameter name followed by “(Non-averaged).” Use a non-averaged interpretation if you plan to use your class prediction model to evaluate data that is acquired at a future date, such as part of a real-time sample classification process within your acquisition method.

Filtered entity list

You build your class prediction model using a list of features (entities) that has been filtered to a smaller set that contains the features contributing to the best separation of your experimental conditions (classifications). The 3D PCA scores plot displayed during the “Review the sample quality in QC on samples in the Analysis: Significance Testing and Fold Change (Step 5 of 8) workflow.” on page 91 provided evidence that the **Filtered by frequency** entity list contains a sufficient set of features that are able to differentiate the samples based on the classifications contained in the experiment interpretations. You can select an alternate entity list for your model; select an entity list that has at least been **Filtered on Flags** to remove one-hit wonders and a **Fold change** entity list, if available, is preferable.

Class prediction algorithm

The class prediction algorithm you select determines which of the available learning tools are used by MPP to train the model. The prediction algorithm uses the known functional classifications (interpretation) in your training sample data to build a prediction model to classify new samples of unknown class. Each algorithm has different traits that may or may not be the most beneficial to building your prediction model. See “Identify appropriate class prediction algorithms” on page 23 to learn more about the algorithms available to build your class prediction model.

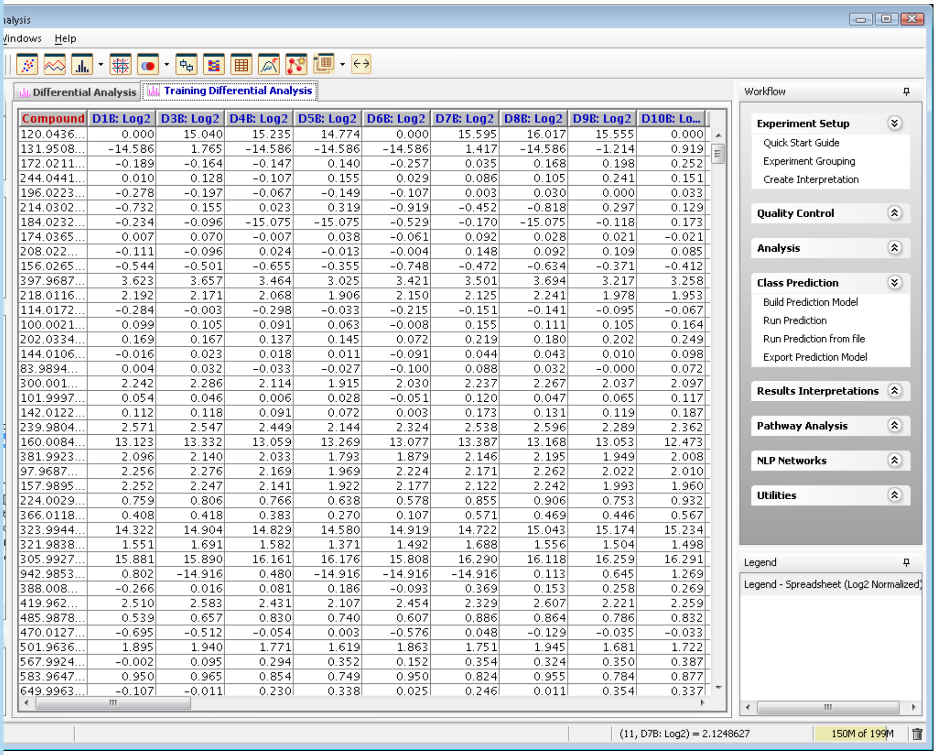
The following examples use Mass Profiler Professional B.12.61 running on 64-bit, Windows 7 Professional.

Launch Mass Profiler Professional

Double-click the Mass Profiler Professional icon located on the desktop, or click **Start > All Programs > Agilent > MassHunter Workstation > Mass Profiler Professional > Mass Profiler Professional**.

If MPP is already open, select your recursive experiment created in “Recreate your differential analysis using the recursive features” on page 125.

The example MPP project created during this workflow contains two experiments, the original analysis using untargeted features from the training data set and the analysis using targeted features found recursively from the same training data set. Figure 139 on page 130 shows how the example project appears with both experiments displayed at the same time.



Compound	D1B: Log2	D3B: Log2	D4B: Log2	D5B: Log2	D6B: Log2	D7B: Log2	D8B: Log2	D9B: Log2	D10B: Lo...
120.0436...	0.000	15.040	15.235	14.774	0.000	15.595	16.017	15.555	0.000
131.9508...	-14.586	1.765	-14.586	-14.586	-14.586	1.417	-14.586	-1.214	0.919
172.0211...	-0.189	-0.164	-0.147	0.140	-0.257	0.035	0.168	0.198	0.252
244.0441...	0.010	0.128	-0.107	0.155	0.029	0.086	0.105	0.241	0.151
196.0223...	-0.278	-0.197	-0.067	-0.149	-0.107	0.003	0.030	0.000	0.033
214.0302...	-0.732	0.155	0.023	0.319	-0.919	-0.452	-0.818	0.297	0.129
184.0232...	-0.234	-0.096	-15.075	-15.075	-0.529	-0.170	-15.075	-0.118	0.173
174.0365...	0.007	0.070	-0.007	0.038	-0.061	0.092	0.028	0.021	-0.021
208.022...	-0.111	-0.096	0.024	-0.013	-0.004	0.148	0.092	0.109	0.085
156.0265...	-0.544	-0.501	-0.655	-0.355	-0.748	-0.472	-0.634	-0.371	-0.412
397.9687...	3.623	3.657	3.464	3.025	3.421	3.501	3.694	3.217	3.258
218.0116...	2.192	2.171	2.068	1.906	2.150	2.125	2.241	1.978	1.953
114.0172...	-0.284	-0.003	-0.298	-0.033	-0.215	-0.151	-0.141	-0.095	-0.067
100.0021...	0.099	0.105	0.091	0.063	-0.008	0.155	0.111	0.105	0.164
202.0334...	0.169	0.167	0.137	0.145	0.072	0.219	0.180	0.202	0.249
144.0106...	-0.016	0.023	0.018	0.011	-0.091	0.044	0.043	0.010	0.098
83.9894...	0.004	0.032	-0.033	-0.027	-0.100	0.088	0.032	-0.000	0.072
300.001...	2.242	2.286	2.114	1.915	2.030	2.237	2.267	2.037	2.097
101.9997...	0.054	0.046	0.006	0.028	-0.051	0.120	0.047	0.065	0.117
142.0122...	0.112	0.118	0.091	0.072	0.003	0.173	0.131	0.119	0.187
239.9804...	2.571	2.547	2.449	2.144	2.324	2.538	2.596	2.289	2.362
160.0084...	13.123	13.332	13.059	13.269	13.077	13.387	13.168	13.053	12.473
381.9923...	2.096	2.140	2.033	1.793	1.879	2.146	2.195	1.949	2.008
97.9687...	2.256	2.276	2.169	1.969	2.224	2.171	2.262	2.022	2.010
157.9895...	2.252	2.247	2.141	1.922	2.177	2.122	2.242	1.993	1.960
224.0029...	0.759	0.806	0.766	0.638	0.578	0.855	0.906	0.753	0.932
366.0118...	0.408	0.418	0.383	0.270	0.107	0.571	0.469	0.446	0.567
323.9944...	14.322	14.904	14.829	14.580	14.919	14.722	15.043	15.174	15.234
321.9838...	1.551	1.691	1.582	1.371	1.492	1.688	1.556	1.504	1.498
305.9927...	15.881	15.890	16.161	16.176	15.808	16.290	16.118	16.259	16.291
942.9853...	0.802	-14.916	0.480	-14.916	-14.916	-14.916	0.113	0.645	1.269
388.008...	-0.266	0.016	0.081	0.186	-0.093	0.369	0.153	0.258	0.269
419.962...	2.510	2.583	2.431	2.107	2.454	2.329	2.607	2.221	2.259
485.9878...	0.539	0.657	0.830	0.740	0.607	0.886	0.864	0.786	0.832
470.0127...	-0.695	-0.512	-0.054	0.003	-0.576	0.048	-0.129	-0.035	-0.033
501.9636...	1.895	1.940	1.771	1.619	1.863	1.751	1.945	1.681	1.722
567.9924...	-0.002	0.095	0.294	0.352	0.152	0.354	0.324	0.350	0.387
583.9647...	0.950	0.965	0.854	0.749	0.950	0.824	0.955	0.784	0.877
649.9963...	-0.107	-0.011	0.230	0.338	0.025	0.246	0.011	0.354	0.337

Figure 139 *MPP project containing two experiments, the original differential analysis of all 40 samples and the recreated differential analysis of the 36 training samples after the features were found recursively.*

Build a prediction model

You can build your class prediction model using either the original analysis based on the untargeted features found in the training data set or the analysis based on the targeted features found recursively from the same training data set.

1. Select your recursive experiment, interpretation, and analysis.

a Double-click the recursive experiment in the Project Navigator. The recursive experiment opens and becomes the active experiment as shown in Figure 140.

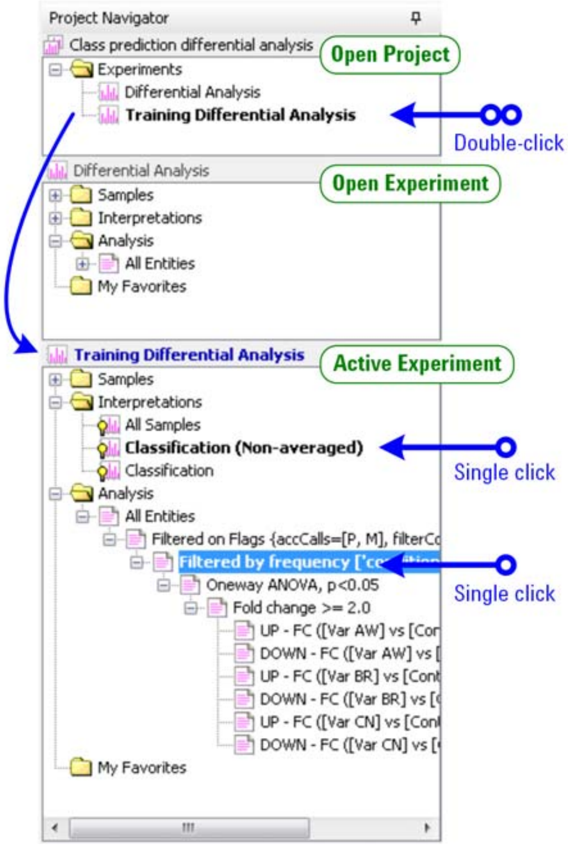


Figure 140 *Project Navigator with both experiments created during the class prediction workflow*

b Click the **Classification (Non-averaged)** interpretation for the recursive experiment in the Project Navigator.

c Click the **Filtered by frequency** analysis entity list. Figure 139 on page 130 shows the MPPP screen for the example recursive data set. You can select an alternate entity list for your model; select an entity list that has at least been **Filtered on Flags** to remove one-hit wonders and a **Fold change** entity list, if available, is preferable.

2. Launch the **Build Prediction Model** wizard.

Click **Build Prediction Model** in the Workflow Browser. The **Class Prediction** wizard is launched.

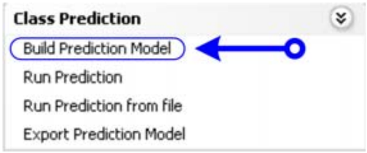


Figure 141 *Selecting Build Prediction Model in the Workflow Browser*

The **Build Prediction Model** wizard has five (5) steps and involves optional dialog boxes in choosing an entity list, interpretation, and parameters for your class prediction algorithm. A flow chart of the Class Prediction wizard is shown in Figure 142.

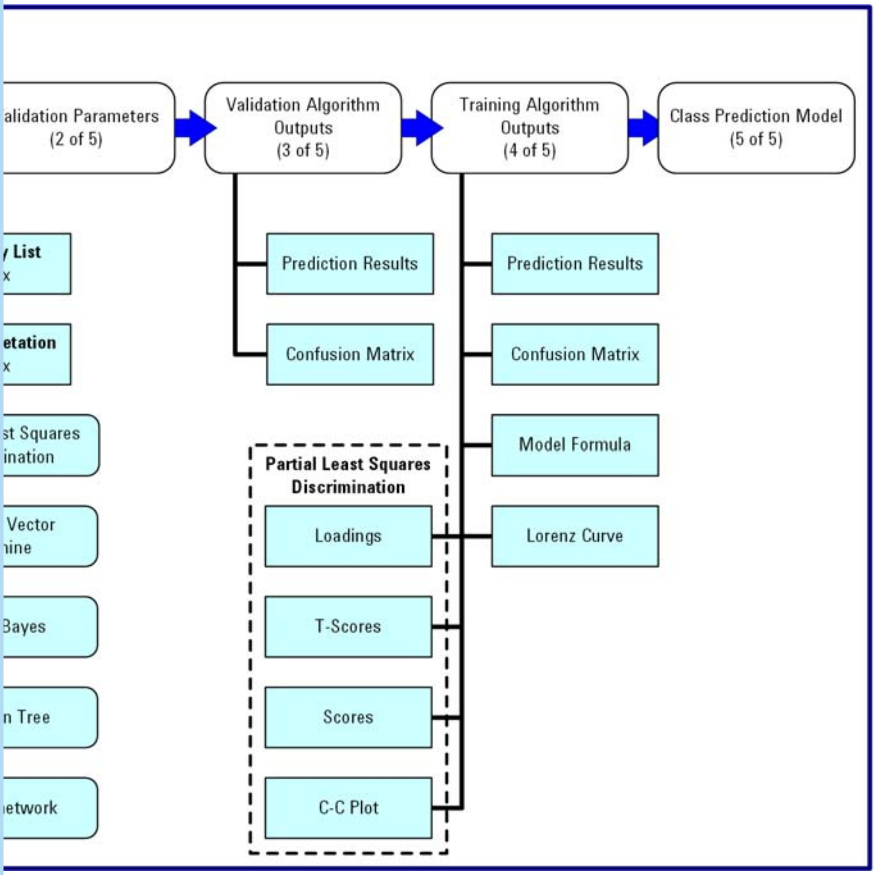


Figure 142 Flow chart of the Build Prediction Model wizard

3. Enter the source parameters in **Class Prediction (Step 1 of 5)**.

a Click **Choose** to select the **Entity List** that you want to use as the source for building your class prediction model if the default entity list is not as described in step 1 - “Select your recursive experiment, interpretation, and analysis.” on page 130.

Select an entity list that has at least been **Filtered on Flags** to remove one-hit wonders. A **Fold change** entity list, if available, is preferable. The **Fold Change** entity list is selected in this example.

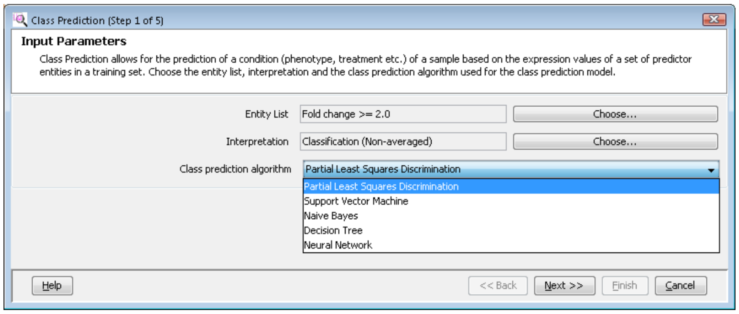
b Click **Choose** to select the **Interpretation** that you want to use as the source for building your class prediction model if the default interpretation is not as described in step 1 - “Select your recursive experiment, interpretation, and analysis.” on page 130.

The optimal interpretation to build your prediction model is a non-averaged interpretation.

c Select a **Class Prediction Algorithm**. Partial Least Squares Discrimination is selected for the example training data set.

Select the algorithm appropriate for your experiment. The available class prediction algorithms are **Partial Least Squares Discrimination**, **Support Vector Machine**, **Naïve Bayes**, **Decision Tree**, and **Neural Network**. Each of the algorithms is described in “Identify appropriate class prediction algorithms” on page 23.

d Click **Next**.



Class Prediction (Step 1 of 5)

Input Parameters

Class Prediction allows for the prediction of a condition (phenotype, treatment etc.) of a sample based on the expression values of a set of predictor entities in a training set. Choose the entity list, interpretation and the class prediction algorithm used for the class prediction model.

Entity List: Fold change >= 2.0

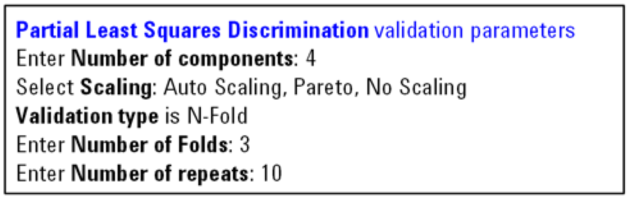
Interpretation: Classification (Non-averaged)

Class prediction algorithm: Partial Least Squares Discrimination

Figure 143 *Input Parameters page (Class Prediction (Step 1 of 5))*

4. Enter the validation parameters in **Class Prediction (Step 2 of 5)**.

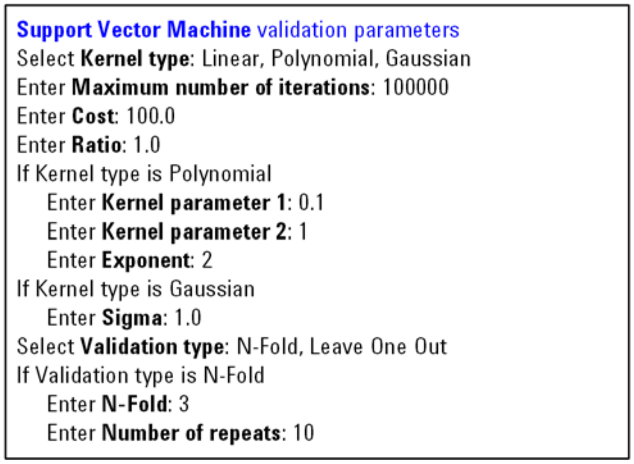
The parameters available in this page of the wizard depend on the class prediction algorithm you selected in the prior step of the wizard (see Figure 144 through Figure 148 for the parameters available and default values for each algorithm).



Partial Least Squares Discrimination validation parameters

- Enter **Number of components**: 4
- Select **Scaling**: Auto Scaling, Pareto, No Scaling
- **Validation type** is N-Fold
- Enter **Number of Folds**: 3
- Enter **Number of repeats**: 10

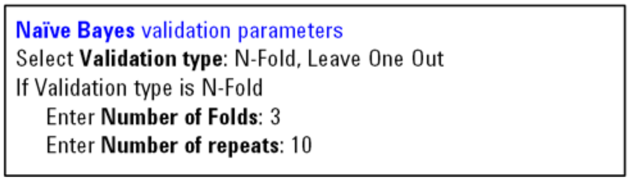
Figure 144 *Validation parameters available for Partial Least Squares Discrimination*



Support Vector Machine validation parameters

- Select **Kernel type**: Linear, Polynomial, Gaussian
- Enter **Maximum number of iterations**: 100000
- Enter **Cost**: 100.0
- Enter **Ratio**: 1.0
- If Kernel type is Polynomial
  - Enter **Kernel parameter 1**: 0.1
  - Enter **Kernel parameter 2**: 1
  - Enter **Exponent**: 2
- If Kernel type is Gaussian
  - Enter **Sigma**: 1.0
- Select **Validation type**: N-Fold, Leave One Out
- If Validation type is N-Fold
  - Enter **N-Fold**: 3
  - Enter **Number of repeats**: 10

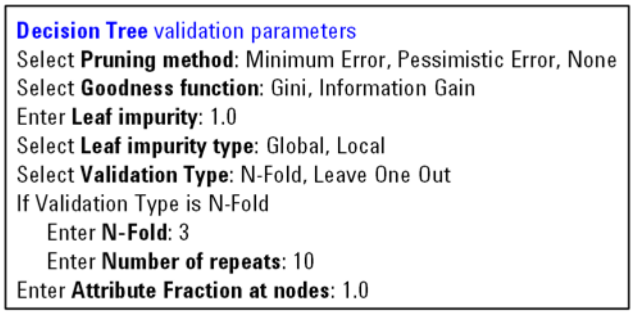
Figure 145 *Validation parameters available for Support Vector Machine*



Naïve Bayes validation parameters

- Select **Validation type**: N-Fold, Leave One Out
- If Validation type is N-Fold
  - Enter **Number of Folds**: 3
  - Enter **Number of repeats**: 10

Figure 146 *Validation parameters available for Naïve Bayes*



Decision Tree validation parameters

Select **Pruning method**: Minimum Error, Pessimistic Error, None

Select **Goodness function**: Gini, Information Gain

Enter **Leaf impurity**: 1.0

Select **Leaf impurity type**: Global, Local

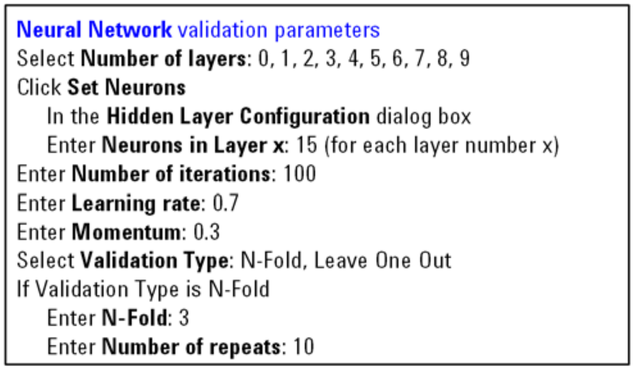
Select **Validation Type**: N-Fold, Leave One Out

If Validation Type is N-Fold

- Enter **N-Fold**: 3
- Enter **Number of repeats**: 10

Enter **Attribute Fraction at nodes**: 1.0

Figure 147 *Validation parameters available for Decision Tree*



Neural Network validation parameters

Select **Number of layers**: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

Click **Set Neurons**

- In the **Hidden Layer Configuration** dialog box
- Enter **Neurons in Layer x**: 15 (for each layer number x)

Enter **Number of iterations**: 100

Enter **Learning rate**: 0.7

Enter **Momentum**: 0.3

Select **Validation Type**: N-Fold, Leave One Out

If Validation Type is N-Fold

- Enter **N-Fold**: 3
- Enter **Number of repeats**: 10

Figure 148 *Validation parameters available for Neural Network*

a Click **Back** to select a different algorithm.

Partial Lease Squares Discrimination is the selected algorithm for the example experiment; the default values are shown in Figure 144 on page 133.

b Type `4` for the **Number of components**.

c Select a **Scaling** to apply to the features in your entity list. Select **No Scaling** if a feature is not present in any of your samples.

The available Scaling options are **Auto Scaling**, **Pareto**, and **No Scaling**.

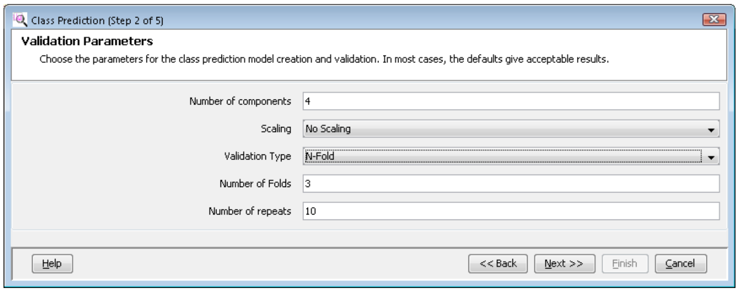
d Select the **Validation type** to apply to your class prediction training.

The available Validation types are **Leave One Out**, and **N-Fold**. The validation types are part of the supervised learning employed to build your prediction model and are described in "Review the steps to create a class prediction model" on page 31. No parameters are available to enter for **Leave One Out**. For **N-Fold** you continue with the following parameters.

e Type `4` for the **Number of folds** associated with the N-Fold Validation type.

f Type `10` for the **Number of repeats** associated with the N-Fold Validation type.

g Click **Next**. The supervised learning model validation is run at this time. It may take a few seconds to several minutes before the next page of the wizard is displayed.



Class Prediction (Step 2 of 5)

Validation Parameters

Choose the parameters for the class prediction model creation and validation. In most cases, the defaults give acceptable results.

Number of components: 4

Scaling: No Scaling

Validation Type: N-Fold

Number of Folds: 3

Number of repeats: 10

Figure 149 *Validation Parameters page (Class Prediction (Step 2 of 5))*

5. Review the validation algorithm outputs in **Class Prediction (Step 3 of 5)**.

The parameters available in this page of the wizard depend on the class prediction algorithm you selected in the prior step of the wizard (see Figure 144 through Figure 148 for the parameters available and default values for each algorithm).

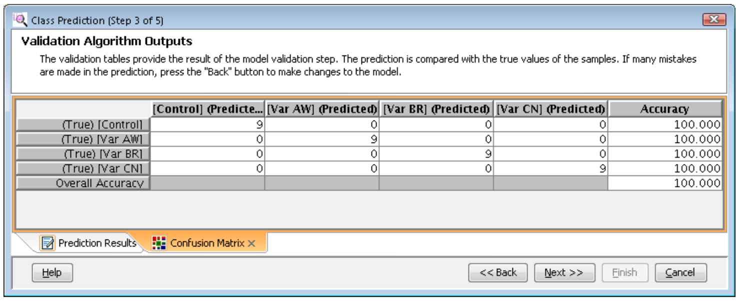
a Review the results of your class prediction model.

Compare the true value of the sample classifications to the predicted classifications based on the prediction model. The Confusion Matrix (Figure 150) provides a quick means to view how the true classifications compare to the predicted classification for replicates among the samples. For validation algorithm outputs the Confusion Matrix results show a cumulative Confusion Matrix, which is the sum of confusion matrices for individual runs of the learning algorithm.

The table of each sample presented in the Prediction Results (Figure 151 on page 136) allows you to review the actual classification and predicted classification for each individual sample.

b Click **Back** to adjust your class prediction algorithm parameters or select a different algorithm.

c Click **Next**. The supervised learning performs training at this time. It may take a few seconds to several minutes before the next page of the wizard is displayed.



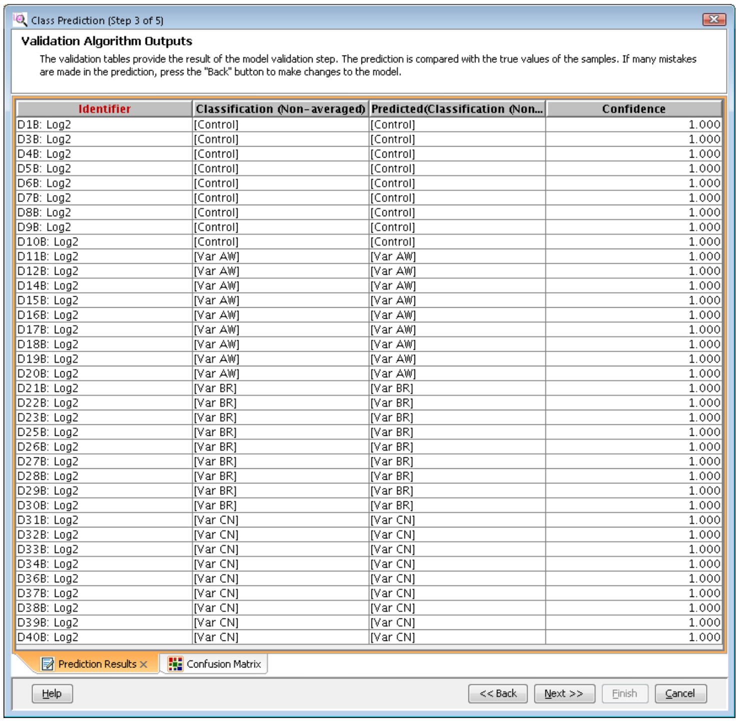
Class Prediction (Step 3 of 5)

Validation Algorithm Outputs

The validation tables provide the result of the model validation step. The prediction is compared with the true values of the samples. If many mistakes are made in the prediction, press the "Back" button to make changes to the model.

	[Control] (Predicte...	[Var AW] (Predicted)	[Var BR] (Predicted)	[Var CN] (Predicted)	Accuracy
(True) [Control]	9	0	0	0	100.000
(True) [Var AW]	0	9	0	0	100.000
(True) [Var BR]	0	0	9	0	100.000
(True) [Var CN]	0	0	0	9	100.000
Overall Accuracy					100.000

Figure 150 *The Confusion Matrix on the Validation Algorithm Outputs page (Class Prediction (Step 3 of 5))*



Validation Algorithm Outputs

The validation tables provide the result of the model validation step. The prediction is compared with the true values of the samples. If many mistakes are made in the prediction, press the "Back" button to make changes to the model.

Identifier	Classification (Non-averaged)	Predicted(Classification (Non...	Confidence
D1B: Log2	[Control]	[Control]	1.000
D3B: Log2	[Control]	[Control]	1.000
D4B: Log2	[Control]	[Control]	1.000
D5B: Log2	[Control]	[Control]	1.000
D6B: Log2	[Control]	[Control]	1.000
D7B: Log2	[Control]	[Control]	1.000
D8B: Log2	[Control]	[Control]	1.000
D9B: Log2	[Control]	[Control]	1.000
D10B: Log2	[Control]	[Control]	1.000
D11B: Log2	[Var AW]	[Var AW]	1.000
D12B: Log2	[Var AW]	[Var AW]	1.000
D14B: Log2	[Var AW]	[Var AW]	1.000
D15B: Log2	[Var AW]	[Var AW]	1.000
D16B: Log2	[Var AW]	[Var AW]	1.000
D17B: Log2	[Var AW]	[Var AW]	1.000
D18B: Log2	[Var AW]	[Var AW]	1.000
D19B: Log2	[Var AW]	[Var AW]	1.000
D20B: Log2	[Var AW]	[Var AW]	1.000
D21B: Log2	[Var BR]	[Var BR]	1.000
D22B: Log2	[Var BR]	[Var BR]	1.000
D23B: Log2	[Var BR]	[Var BR]	1.000
D25B: Log2	[Var BR]	[Var BR]	1.000
D26B: Log2	[Var BR]	[Var BR]	1.000
D27B: Log2	[Var BR]	[Var BR]	1.000
D28B: Log2	[Var BR]	[Var BR]	1.000
D29B: Log2	[Var BR]	[Var BR]	1.000
D30B: Log2	[Var BR]	[Var BR]	1.000
D31B: Log2	[Var CN]	[Var CN]	1.000
D32B: Log2	[Var CN]	[Var CN]	1.000
D33B: Log2	[Var CN]	[Var CN]	1.000
D34B: Log2	[Var CN]	[Var CN]	1.000
D36B: Log2	[Var CN]	[Var CN]	1.000
D37B: Log2	[Var CN]	[Var CN]	1.000
D38B: Log2	[Var CN]	[Var CN]	1.000
D39B: Log2	[Var CN]	[Var CN]	1.000
D40B: Log2	[Var CN]	[Var CN]	1.000

Figure 151 *The Prediction Results on the Validation Algorithm Outputs page (Class Prediction (Step 3 of 5))*

6. Review the training algorithm outputs in **Class Prediction (Step 4 of 5)**.

On this page of the wizard, you review the results of training the prediction model and the parameters associated with the prediction model. The available views on this page depend on the class prediction algorithm you selected. This same information can also be viewed after you complete your prediction model; double-click on the model in the Project Navigator.

a Review the results and training algorithm of your class prediction model.

For training algorithm outputs, the Confusion Matrix results show the result of applying the final prediction model to the training sample data. In addition to the tabs containing the Confusion Matrix and Prediction Results, two additional tabs containing the training algorithm outputs are added for the Model Formula and Lorenz Curve.

The Lorenz Curve (Figure 152 on page 137) is used to visualize the *class belongingness measure*, a value from 0 to 1 assigned to each sample for a particular class. When the samples are sorted in decreasing class belongingness along the x-axis, the red line in the Lorenz Curve creates a trace that identifies the samples that are predicted to belong to the selected class. The y-axis is the cumulative fraction of samples that fit in the selected class, and therefore in conjunction with the x-axis, shows the total number of samples included in the classification. In this case, a line of equality is obtained because the correlation of the *True Positive Fraction* is 1:1 with the *Rank*; each sample in the population contributes equally to the accuracy of the model prediction.

The point along the Lorenz curve where the cumulative *True Positive Fraction* reaches its maximum value of 1 indicates the number of samples that would be predicted to be in the selected *Class Fraction* if all the items actually belonging to this class are classified correctly. In this example 9 samples account for 100% of the samples that were actually assigned to the Control class. Based on our replicate sample structure, the same Lorenz curve is displayed for each classification.

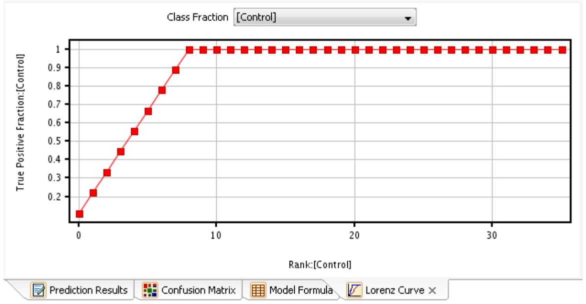
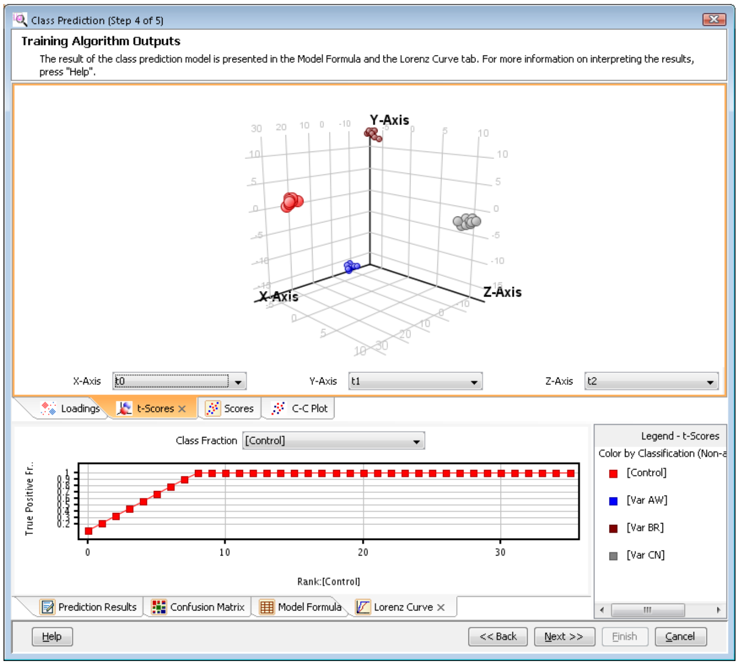


Figure 152 *Lorenz curve for the example training data set*

b Click **Back** to adjust your class prediction algorithm parameters or select a different algorithm.

c Click **Next**.



Training Algorithm Outputs

The result of the class prediction model is presented in the Model Formula and the Lorenz Curve tab. For more information on interpreting the results, press "Help".

X-Axis t0 Y-Axis t1 Z-Axis t2

Class Fraction [Control]

Legend - t-Scores
Color by Classification (Non-a
🟥 [Control]
🟦 [Var AW]
🟫 [Var BR]
⬛ [Var CN]

Figure 153 *Training Algorithm Outputs page (Class Prediction (Step 4 of 5))*

7. Save the class prediction model in **Class Prediction (Step 4 of 5)**.

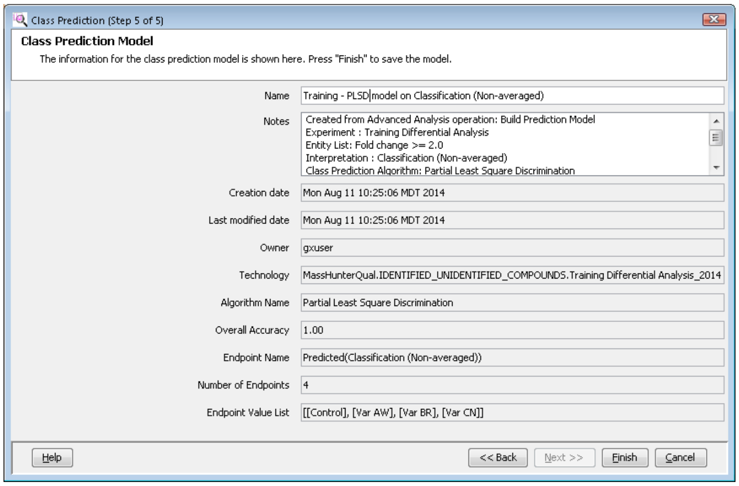
The parameters available in this page of the wizard depend on the class prediction algorithm you selected in the prior step of the wizard (see Figure 144 through Figure 148 for the parameters available and default values for each algorithm).

a Review your results.

b Add or edit descriptive information that is stored with the class prediction model in the **Name** and **Notes** parameters. A clear, simple descriptive name is recommended to help your project organization.

c Click **Back** to review and adjust your class prediction algorithm parameters or select a different algorithm.

d Click **Finish**.



Class Prediction Model

The information for the class prediction model is shown here. Press "Finish" to save the model.

Name: Training - PLSD model on Classification (Non-averaged)

Notes:
Created from Advanced Analysis operation: Build Prediction Model
Experiment : Training Differential Analysis
Entity List: Fold change >= 2.0
Interpretation : Classification (Non-averaged)
Class Prediction Algorithm: Partial Least Square Discrimination

Creation date: Mon Aug 11 10:25:06 MDT 2014

Last modified date: Mon Aug 11 10:25:06 MDT 2014

Owner: gxuser

Technology: MassHunterQual.IDENTIFIED_UNIDENTIFIED_COMPOUNDS.Training Differential Analysis_2014

Algorithm Name: Partial Least Square Discrimination

Overall Accuracy: 1.00

Endpoint Name: Predicted(Classification (Non-averaged))

Number of Endpoints: 4

Endpoint Value List: [[Control], [Var AW], [Var BR], [Var CN]]

Figure 154 *Class Prediction Model page (Class Prediction (Step 5 of 5))*

Export your prediction model

Exporting your prediction model begins with the assumption that you have created a satisfactory prediction model using your training data. You export a prediction model to validate your model with your validation data set and to use the prediction model with Sample Class Predictor.

1. Launch **Export Prediction Model** in the Workflow Browser.

Click **Export Prediction Model** in the Workflow Browser.

Export Prediction Model has three (3) steps. A flow chart of exporting your class prediction model is shown in Figure 155.

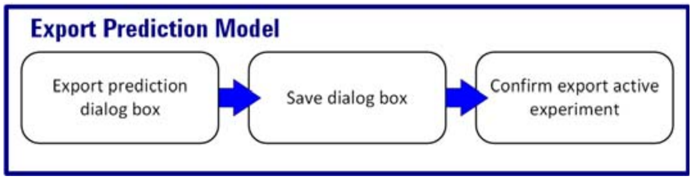
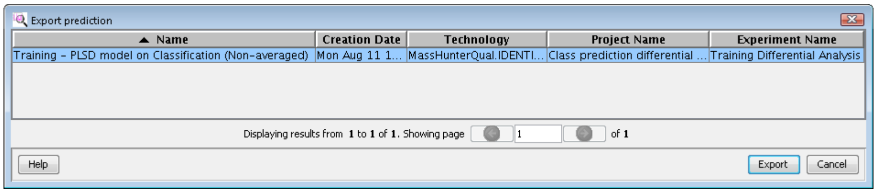

Export Prediction Model
Export prediction dialog box
Save dialog box
Confirm export active experiment

Figure 155 *Flow chart of the steps to export your class prediction model*

2. Select your prediction model.

a Select the prediction model to export in the **Export prediction** dialog box. The prediction model row is highlighted when selected.

b Click **Export**.


Export prediction

▲ Name	Creation Date	Technology	Project Name	Experiment Name
Training - PLSD model on Classification (Non-averaged)	Mon Aug 11 1...	MassHunterQual.IDENTI...	Class prediction differential ...	Training Differential Analysis

Displaying results from **1** to **1** of **1**. Showing page 1 of **1**

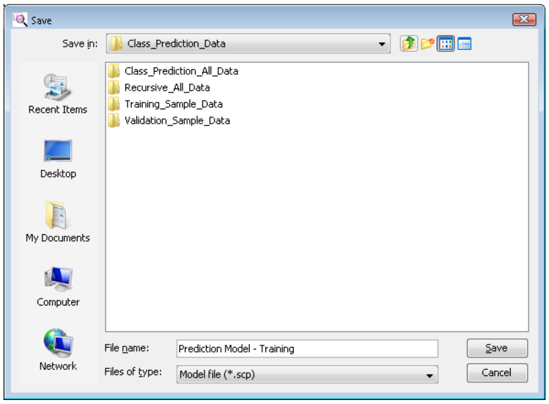
Figure 156 *Export prediction dialog box*

3. Enter the export file name and folder.

a Navigate to the folder containing your class prediction training feature files (CEF files) in the **Save** dialog box.

b Type the **File name**. For this example prediction model type `Prediction Model - Training`.

c Click **Save**.


Save

Save in: Class_Prediction_Data

Class_Prediction_All_Data
Recursive_All_Data
Training_Sample_Data
Validation_Sample_Data

File name: Prediction Model - Training

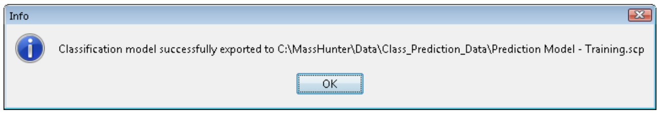
Files of type: Model file (*.scp)

Figure 157 *Save dialog box*

4. Review the confirmation of the exported prediction model.

a Review the folder and file name of the exported prediction model in the **Info** dialog box.

b Click **OK**.



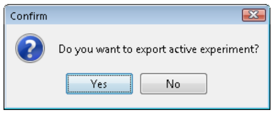
Classification model successfully exported to C:\MassHunter\Data\Class_Prediction_Data\Prediction Model - Training.scp

Figure 158 *Save dialog box*

5. Save your project.

The Export Prediction Model workflow automatically encourages you to save (export) your current project as a TAR file for archiving, restoration of any future analysis to the current results, sharing the data with a collaborator, or sharing the data with Agilent customer support.

Click **Yes**. The exported project file name is the same name you entered for the prediction file name.



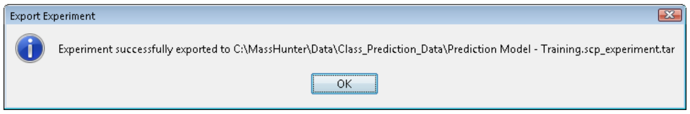
Do you want to export active experiment?

Figure 159 *Confirm dialog box*

6. Review the confirmation of the saved project.

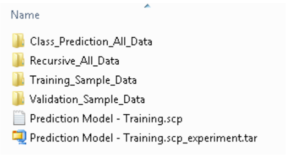
a Review the folder and file name of the exported MPP project in the **Export Experiment** dialog box.

b Click **OK**. The project and experiment(s) are exported to the same location you saved your prediction model as shown in Figure 161 on page 140.



Experiment successfully exported to C:\MassHunter\Data\Class_Prediction_Data\Prediction Model - Training.scp_experiment.tar

Figure 160 *Export Experiment dialog box*



Name

Class_Prediction_All_Data
Recursive_All_Data
Training_Sample_Data
Validation_Sample_Data
Prediction Model - Training.scp
Prediction Model - Training.scp_experiment.tar

Figure 161 *Folder containing example class prediction model and project*

Validate your prediction model

Validating your prediction model begins with the assumption that you have created a satisfactory prediction model and exported the model to a prediction model file. Before using your model to classify new data, the next step is to validate the model using the validation data set aside in the section "Import and organize your recursive data" on page 125. If you did not perform a recursive feature finding, use the validation data set aside in section "Next step..." on page 66.

1. Launch **Run Prediction from file** in the Workflow Browser.

Click **Run Prediction from file** in the Workflow Browser.

The **Run Prediction from file** wizard has four (4) steps and involves additional dialog boxes to open the validation data files and reorder the samples. A flow chart of validating your class prediction model using Run Prediction from file is shown in Figure 162.

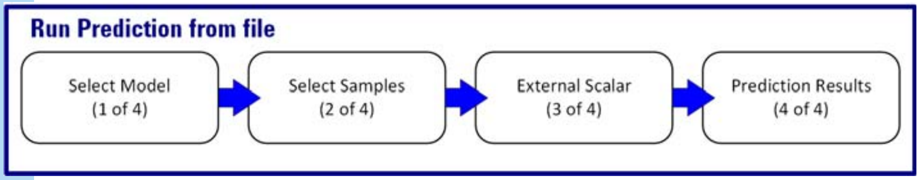
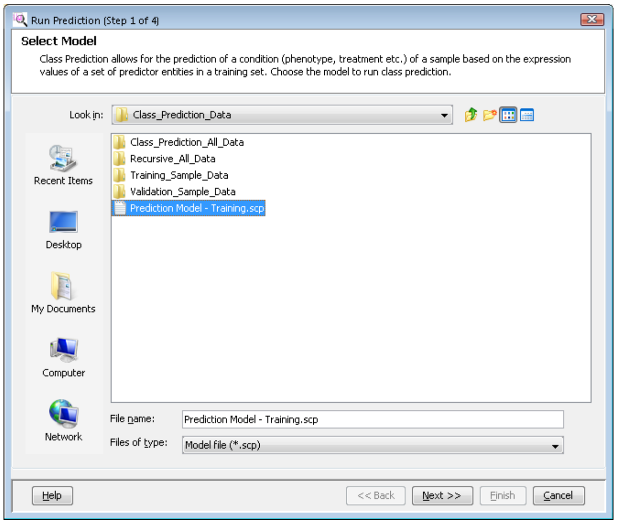


Figure 162 *Flow chart of the Run Prediction from file wizard*

2. Select your prediction model in **Run Prediction (Step 1 of 4)**.

a Select the prediction model to export. The prediction model is highlighted when selected.

b Click **Next**.



Run Prediction (Step 1 of 4)

Select Model

Class Prediction allows for the prediction of a condition (phenotype, treatment etc.) of a sample based on the expression values of a set of predictor entities in a training set. Choose the model to run class prediction.

Look in: Class_Prediction_Data

- Class_Prediction_All_Data
- Recursive_All_Data
- Training_Sample_Data
- Validation_Sample_Data
- Prediction Model - Training.scp

File name: Prediction Model - Training.scp

Files of type: Model file (*.scp)

Figure 163 *Select Model page (Run Prediction (Step 1 of 4))*

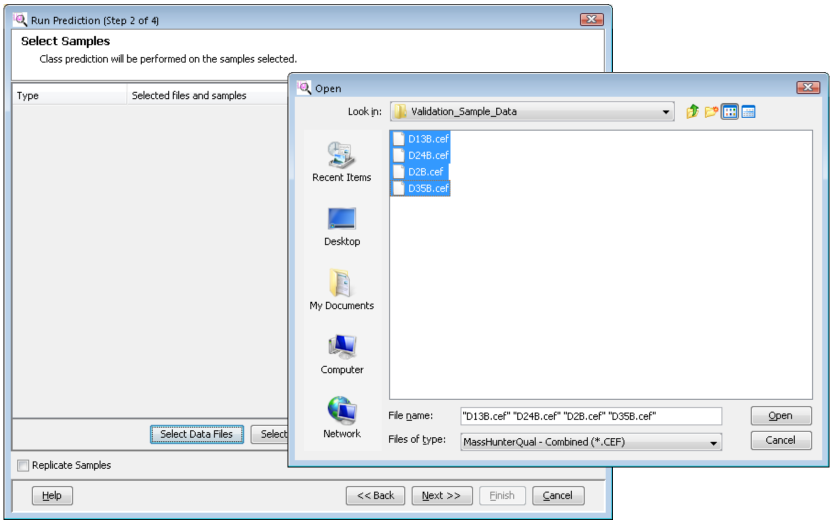
3. Select the validation data to import in **Run Prediction (Step 2 of 4)**.

a Click **Select Data Files**.

b Select the validation data files to open.

Note: Orderly naming of the data files with respect to the parameters related to the independent variables helps you make sure that all of the data is selected.

c Click **Open**.



Look in: Validation_Sample_Data

D13B.cef
D24B.cef
D2B.cef
D35B.cef

File name: "D13B.cef" "D24B.cef" "D2B.cef" "D35B.cef"

Files of type: MassHunterQual - Combined (*.CEF)

Figure 164 *Selection of the validation CEF files from MassHunter Qualitative Analysis*

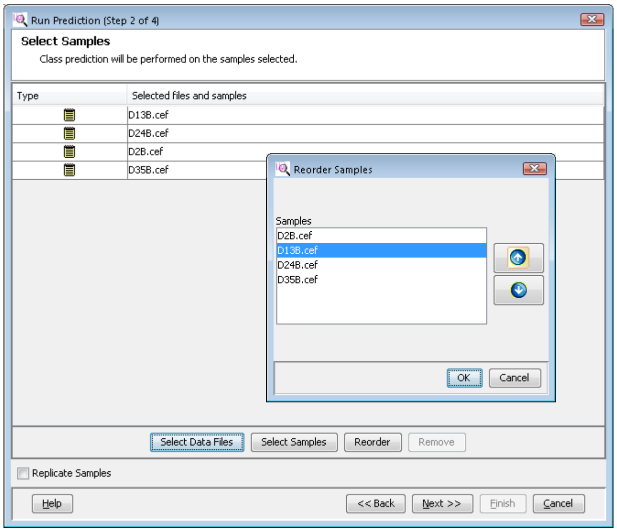
d Click **Reorder** to arrange the file list different from alphabetical order.

You can arrange the files, for example, in the same order of your original sample grouping (classifications) described in section "Review and order the selected files that are imported in the MS Experiment Creation Wizard (Step 5 of 11)." on page 73.

e Click the **Up** or **Down** buttons to reorder the selected sample or samples.

f Click **OK**.

g Click **Next**.



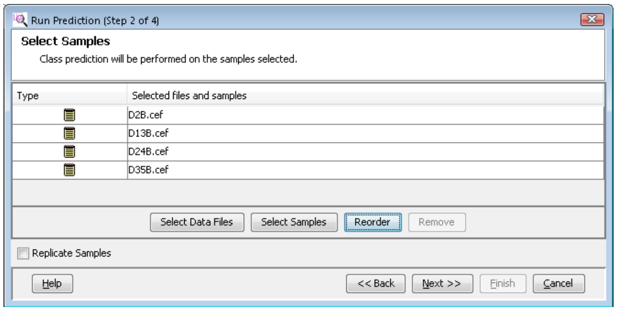
Selected files and samples:

D13B.cef
D24B.cef
D2B.cef
D35B.cef

Samples:

D2B.cef
D13B.cef
D24B.cef
D35B.cef

Figure 165 *Selected sample files in the Reorder Samples dialog box are initially arranged in alphabetical order. This order is not necessarily the experimental order.*


Run Prediction (Step 2 of 4)

Select Samples

Class prediction will be performed on the samples selected.

Type	Selected files and samples
▤	D2B.cef
▤	D13B.cef
▤	D24B.cef
▤	D35B.cef

□ Replicate Samples

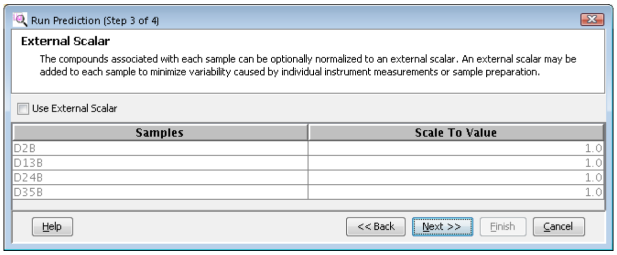
Figure 166 Select Samples page (Run Prediction (Step 2 of 4))

4. Select whether to normalize the data in **Run Prediction (Step 3 of 4)**.

Normalizing the data reduces the variability caused by sample preparation and instrument response. You can use manually set external scaling value for each sample file. No scalar is used for this example class prediction model.

a Clear the **Use External Scalar** check box.

b Click **Next**. Your class prediction model is immediately run to classify the validation data files. A **Progress** dialog box is displayed during the classification (Figure 168).


Run Prediction (Step 3 of 4)

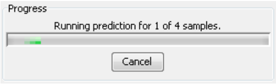
External Scalar

The compounds associated with each sample can be optionally normalized to an external scalar. An external scalar may be added to each sample to minimize variability caused by individual instrument measurements or sample preparation.

□ Use External Scalar

Samples	Scale To Value
D2B	1.0
D13B	1.0
D24B	1.0
D35B	1.0

Figure 167 External Scalar page (Run Prediction (Step 3 of 4))


Progress

Running prediction for 1 of 4 samples.

Figure 168 Progress dialog box

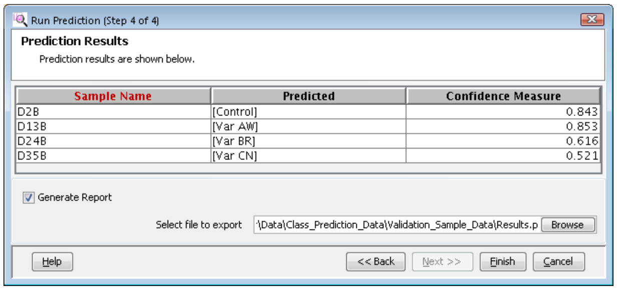
5. Review the prediction results in **Run Prediction (Step 4 of 4)**.

a Review the results of applying your class prediction model to the validation data files. The validation data files were not used to create your prediction model but, in this example, were collected experimentally at the same time as the training data files.

b Mark the **Generate Report** check box. This saves the results of running the class prediction model on the validation data files in PDF format.

c Type a meaningful file name for the PDF report in **Select file to export**. By default the report is saved in the same folder with the validation data files and the file name is the first sample name.

d Click **Finish**. A **Progress** dialog box is displayed while the report is generated.



Run Prediction (Step 4 of 4)

Prediction Results

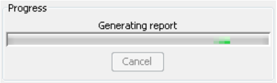
Prediction results are shown below.

Sample Name	Predicted	Confidence Measure
D2B	[Control]	0.843
D13B	[Var AW]	0.853
D24B	[Var BR]	0.616
D35B	[Var CN]	0.521

☑ Generate Report

Select file to export `\Data\Class_Prediction_Data\Validation_Sample_Data\Results.p`

Figure 169 *Prediction Results page (Run Prediction (Step 4 of 4))*



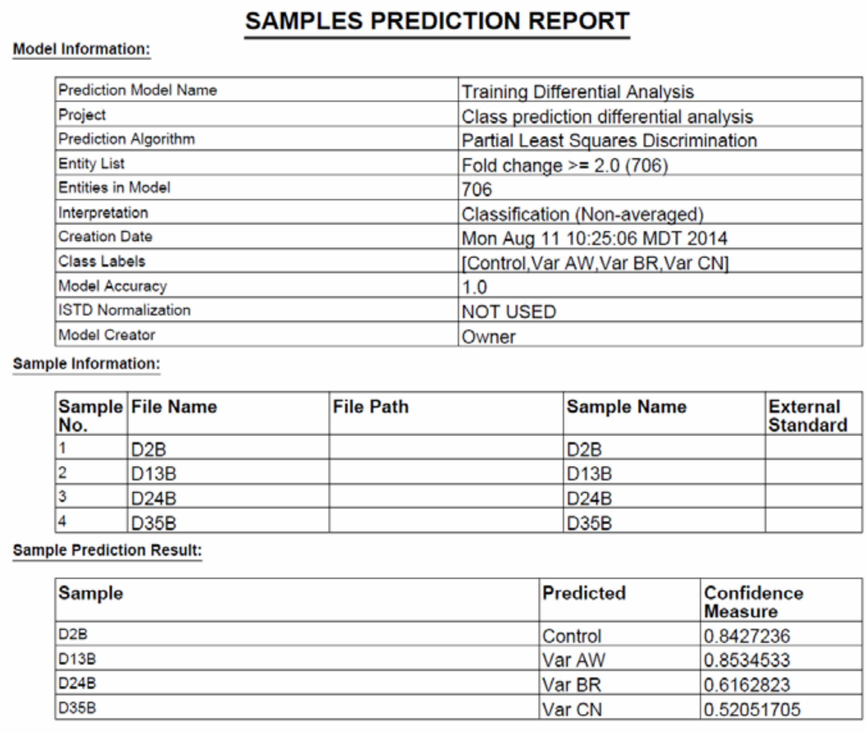
Progress

Generating report

Figure 170 *Progress dialog box*

6. Review the saved report.

Open the saved report using an appropriate program that can open and view PDF files. The first page (Figure 171) of the report contains the results of the prediction model applied to your validation data files. The subsequent pages contain PCA Score plots for each sample file placed with the training PCA Scores. Figure 172 on page 145 shows an example page for a validation data file.



SAMPLES PREDICTION REPORT

Model Information:

Prediction Model Name	Training Differential Analysis
Project	Class prediction differential analysis
Prediction Algorithm	Partial Least Squares Discrimination
Entity List	Fold change >= 2.0 (706)
Entities in Model	706
Interpretation	Classification (Non-averaged)
Creation Date	Mon Aug 11 10:25:06 MDT 2014
Class Labels	[Control,Var AW,Var BR,Var CN]
Model Accuracy	1.0
ISTD Normalization	NOT USED
Model Creator	Owner

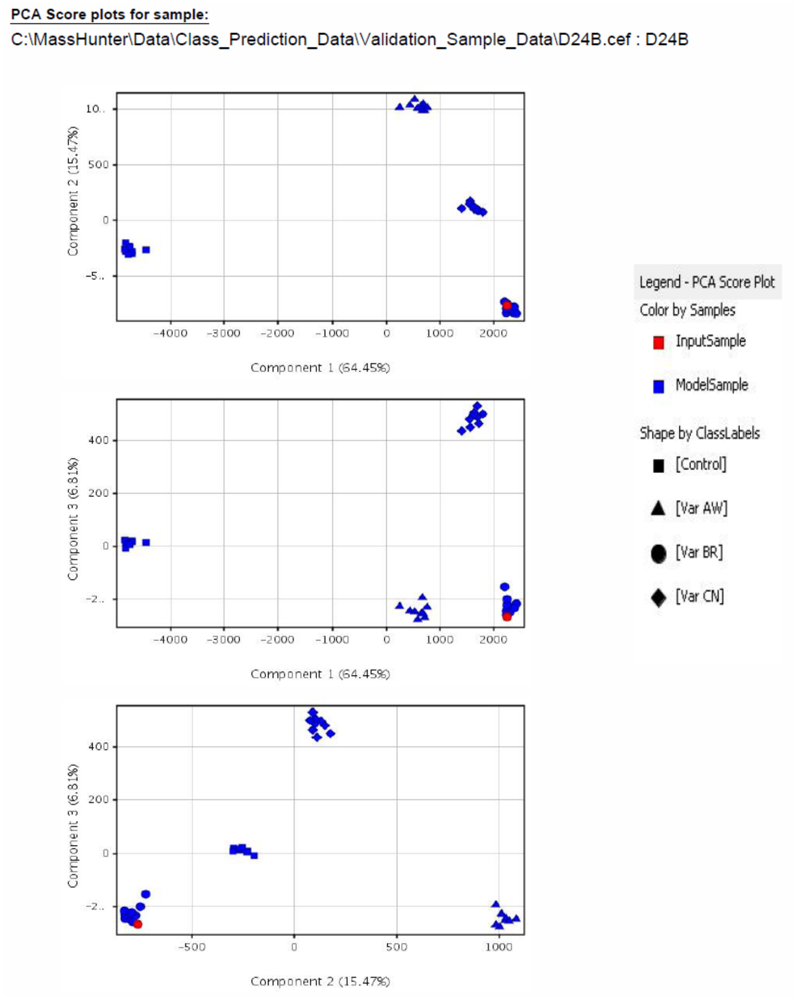
Sample Information:

Sample No.	File Name	File Path	Sample Name	External Standard
1	D2B		D2B	
2	D13B		D13B	
3	D24B		D24B	
4	D35B		D35B	

Sample Prediction Result:

Sample	Predicted	Confidence Measure
D2B	Control	0.8427236
D13B	Var AW	0.8534533
D24B	Var BR	0.6162823
D35B	Var CN	0.52051705

Figure 171 *First page of the PDF prediction report for the example validation data files*



PCA Score plots for sample:

C:\MassHunter\Data\Class_Prediction_Data\Validation_Sample_Data\D24B.cef : D24B

Figure 172 *Example sample PCA Score plot from the PDF prediction report*

Recreate your prediction model using recursion

This is another opportunity during the class prediction workflow to perform a recursive feature finding in your sample data files. Recursive feature finding at this step uses the features (entities) in your class prediction model as the targeted list of features. If you performed recursive feature finding after your initial differential analysis you do not need to perform recursive feature finding again. If you did not perform a recursive feature finding after your initial differential analysis, export the features from your class prediction model for recursive feature finding.

By using the entity list containing the class prediction model features to perform a targeted feature finding, you improve the statistical accuracy (measure of confidence) of your differential analysis and improve the accuracy of your subsequent class prediction model.

Combined with collecting replicate samples in your experiment, recursive feature finding improves the statistical accuracy (confidence) of your analysis and reduces the potential for obtaining a false positive or a false negative answer to your hypothesis and sample classification. For an overview of finding features recursively, see "Review recursive feature finding" on page 34.

If you are unsure whether to perform recursive feature finding, review "Decide whether or not to perform recursive feature finding" on page 99.

Create the prediction model entity list

Export the features that are used by your class prediction model. These prediction model features are used to perform targeted feature finding from the original data files.

a Right-click the prediction model in the Experiment Navigator.

b Click **Expand as Entity List**. The entity list is automatically created and placed in the Experiment Navigator with the same name as the class prediction model.

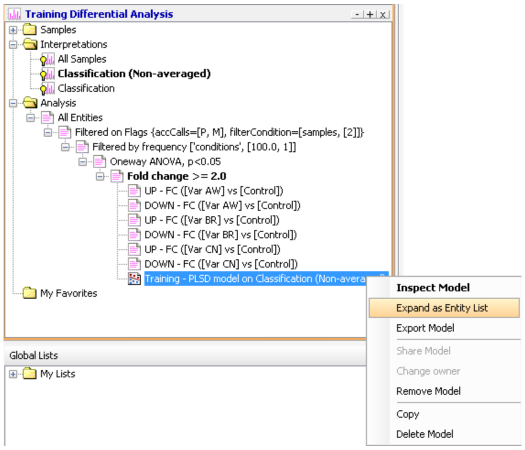

Training Differential Analysis
- + x
Samples
Interpretations
All Samples
Classification (Non-averaged)
Classification
Analysis
All Entities
Filtered on Flags {accCalls=[P, M], filterCondition=[samples, [2]]}
Filtered by frequency ['conditions', [100.0, 1]]
Oneway ANOVA, p<0.05
Fold change >= 2.0
UP - FC ([Var AW] vs [Control])
DOWN - FC ([Var AW] vs [Control])
UP - FC ([Var BR] vs [Control])
DOWN - FC ([Var BR] vs [Control])
UP - FC ([Var CN] vs [Control])
DOWN - FC ([Var CN] vs [Control])
Training - PLSD model on Classification (Non-avera
My Favorites
Global Lists
My Lists
Inspect Model
Expand as Entity List
Export Model
Share Model
Change owner
Remove Model
Copy
Delete Model

Figure 173 *Expand as Entity list*

c Click **Export for Recursion** in the Workflow Browser and follow the steps in section "Export the entity list for recursion" on page 101 using the prediction model entity list in place of the filtered by frequency entity list.

Rebuild your prediction model using recursive feature finding

During recursive feature finding you export the features used by your class prediction model as a targeted list of features for finding in your original sample data files. This step improves the quality of finding the features in the original sample files; targeted feature finding focuses on finding a specific set of features with less emphasis on feature strength.

Substituting the prediction model entity list as the list of targeted features, you repeat the same set of steps described in section "Perform a recursive feature finding" on page 99. The key steps for Qualitative Analysis are:

Create a method to Find Compounds by Formula (FbF) in Qualitative Analysis.

1. "Recursive feature finding using Qualitative Analysis" on page 102
2. "Create a method to Find Compounds by Formula" on page 102
3. "Save your Find Compounds by Formula method" on page 114
4. "Set the Export CEF Options" on page 114
5. "Enable the method to run in MassHunter DA Reprocessor" on page 115

Confirm your FbF method using a single sample data file.

6. "Confirm the FbF method on a single data file" on page 116

Recursively find compounds in the entire sample data set using DA Reprocessor.

7. "Find compounds using DA Reprocessor" on page 117

Recreate your project experiment, filters, and differential analysis in MPP.

8. "Import and organize your recursive data" on page 125
9. "Recreate your differential analysis using the recursive features" on page 125
10. "Save the project containing your recursive analysis" on page 126

Rebuild and validate your class prediction model in MPP.

11. "Build your prediction model" on page 128
12. "Export your prediction model" on page 139
13. "Validate your prediction model" on page 141

If you are using Profinder refer to "Recursive feature finding using MassHunter Profinder" on page 117.

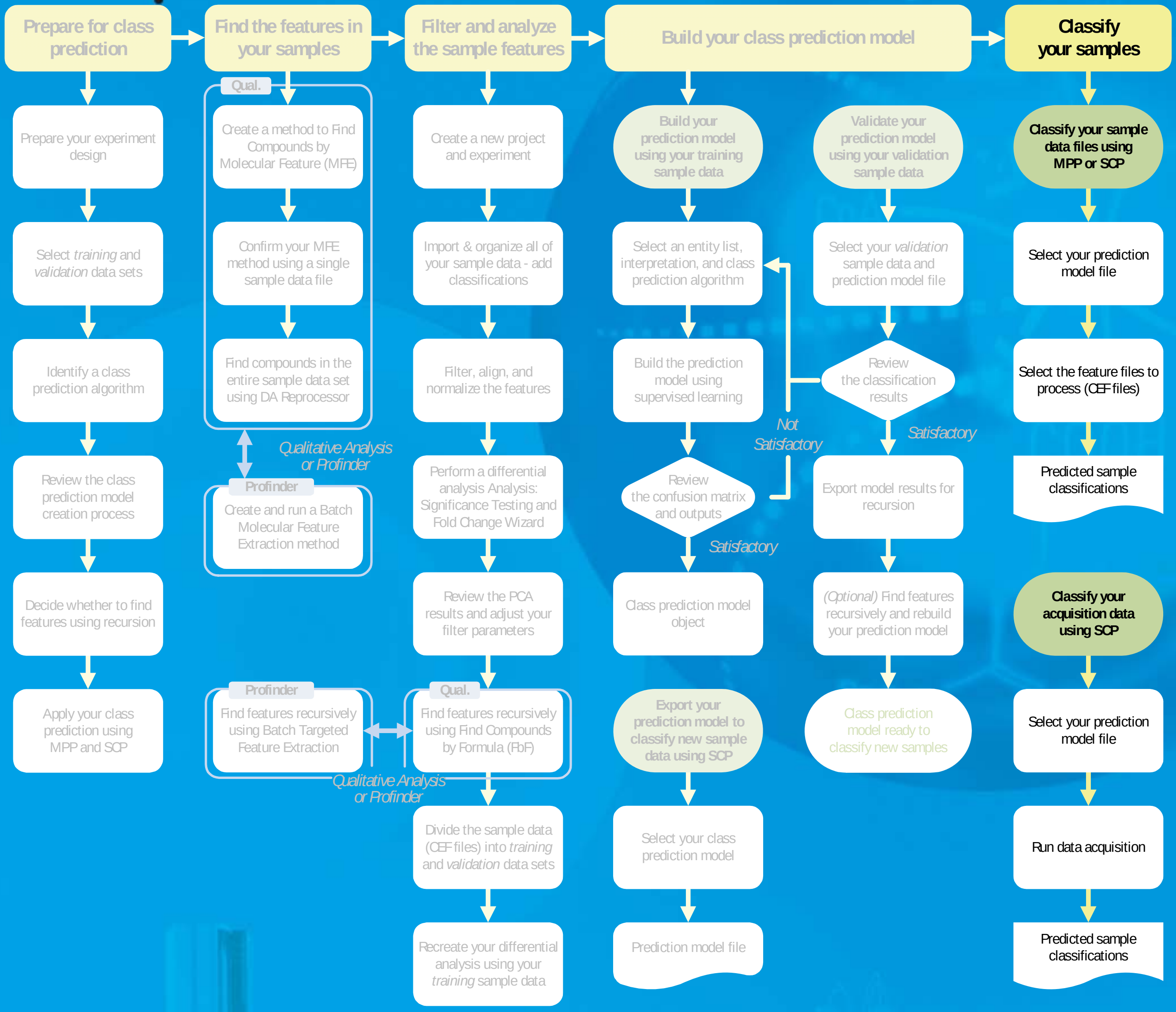
Your class prediction model is now ready to classify new samples.

Next Step...

You have now completed the fourth step of the class prediction workflow. In the next workflow step you apply your prediction model using Mass Profiler Professional and Sample Class Predictor.

Classify your samples

Your prediction model can be used to predict functional classes like diseases and conditions from abundance profile of the entities in your sample data.



Prepare for class prediction

Prepare your experiment design

Select *training* and *validation* data sets

Identify a class prediction algorithm

Review the class prediction model creation process

Decide whether to find features using recursion

Apply your class prediction using MPP and SCP

Find the features in your samples

Qual.

Create a method to Find Compounds by Molecular Feature (MFE)

Confirm your MFE method using a single sample data file

Find compounds in the entire sample data set using DA Reprocessor

Qualitative Analysis or Profinder

Profinder

Create and run a Batch Molecular Feature Extraction method

Profinder

Find features recursively using Batch Targeted Feature Extraction

Qualitative Analysis or Profinder

Filter and analyze the sample features

Create a new project and experiment

Import & organize all of your sample data - add classifications

Filter, align, and normalize the features

Perform a differential analysis Analysis: Significance Testing and Fold Change Wizard

Review the PCA results and adjust your filter parameters

Qual.

Find features recursively using Find Compounds by Formula (FbF)

Divide the sample data (CEF files) into *training* and *validation* data sets

Recreate your differential analysis using your *training* sample data

Build your class prediction model

Build your prediction model using your training sample data

Select an entity list, interpretation, and class prediction algorithm

Build the prediction model using supervised learning

Review the confusion matrix and outputs

Satisfactory

Not Satisfactory

Class prediction model object

Export your prediction model to classify new sample data using SCP

Select your class prediction model

Prediction model file

Validate your prediction model using your validation sample data

Select your *validation* sample data and prediction model file

Review the classification results

Satisfactory

Export model results for recursion

(Optional) Find features recursively and rebuild your prediction model

Class prediction model ready to classify new samples

Classify your samples

Classify your sample data files using MPP or SCP

Select your prediction model file

Select the feature files to process (CEF files)

Predicted sample classifications

Classify your acquisition data using SCP

Select your prediction model file

Run data acquisition

Predicted sample classifications

Agilent Technologies

Overview of classifying new samples

Classifying new samples begins with the assumption that you have created a satisfactory prediction model using your training data, validated the prediction model using your validation data set, and exported your prediction model to use the model with Mass Profiler Professional and Sample Class Predictor. MPP and SCP are separately licensed programs. You can only use SCP with a valid Sample Class Predictor OrderID.

New sample data files are classified using Mass Profiler Professional or Sample Class Predictor. Both software packages can classify individual or multiple sample data files that were previously acquired. Sample Class Predictor can be integrated with your acquisition software to automate classification of new samples as part of your acquisition method. Class prediction integrated with acquisition adds a means to perform real-time sample quality assurance and control to your acquisition method. During acquisition, your class prediction model is a targeted analysis that is applied to your sample data files.

Sample Class Predictor interfaces with both ChemStation and MassHunter acquisition software:

- With ChemStation, Sample Class Prediction runs within acquisition.
- With MassHunter, Sample Class Prediction is run using MPP.

The supported data file types are AMDIS, ChemStation, MassHunter, GC Scan, and Generic. For class prediction on an AMDIS data source, both the FIN and corresponding ELU file must be present in the same working directory.

Samples must be from the same data source as those with which the prediction model was created, but the data does not need to be from the same technology. When class prediction is run from a file, SCP and MPP use the alignment values included in the prediction model file to perform alignments on the features in the new data.

Classify your sample data files

Three ways are available to classify new sample data files that have previously been acquired:

1. MPP - **Run Prediction**: Classify samples that are part of an experiment but were not part of building or validating the prediction model. This classification is described below in “MPP - Run Prediction”.
2. MPP - **Run Prediction from File**: Classify samples that reside in a folder on your computer or on a network drive. This classification is described in “MPP - Run Prediction from File” on page 152.
3. SCP - **Project > Run Prediction**: Classify samples that reside in a folder on your computer or on a network drive using the same **Run Prediction from File** wizard used by MPP. SCP is a separately licensed program. This classification is described in “SCP - Project > Run Prediction” on page 153.

MPP - Run Prediction

Classify samples that are part of an experiment but were not part of building or validating the prediction model. Run Prediction classifies all of the samples in the current experiment. Run Prediction can be used to generate a PDF report for your training data files.

1. Launch **Run Prediction** in the Workflow Browser.

Click **Run Prediction** in the Workflow Browser.

The **Run Prediction** wizard has three (3) steps. A flow chart of classifying new sample using Run Prediction is shown in Figure 174.

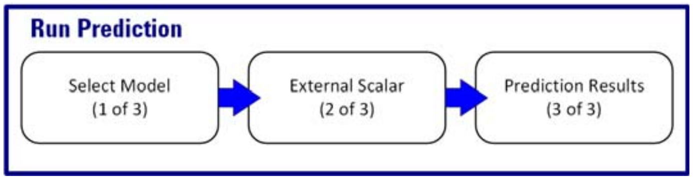
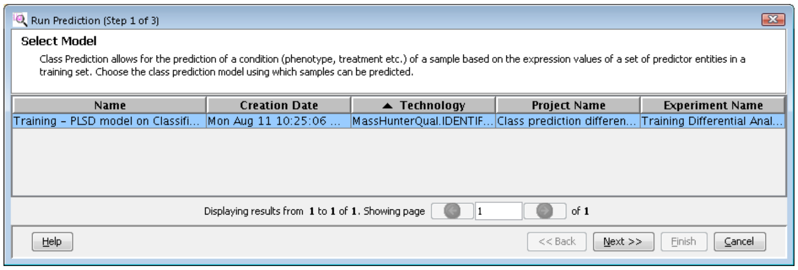


Figure 174 *Flow chart of the Run Prediction wizard*

2. Select your prediction model in **Run Prediction (Step 1 of 3)**.

a Select the prediction model to export. The prediction model is highlighted when selected.

b Click **Next**.



Select Model

Class Prediction allows for the prediction of a condition (phenotype, treatment etc.) of a sample based on the expression values of a set of predictor entities in a training set. Choose the class prediction model using which samples can be predicted.

Name	Creation Date	▲ Technology	Project Name	Experiment Name
Training – PLSD model on Classifi...	Mon Aug 11 10:25:06 ...	MassHunterQual.IDENTIF...	Class prediction differen...	Training Differential Anal...

Displaying results from **1** to **1** of **1**. Showing page 1 of **1**

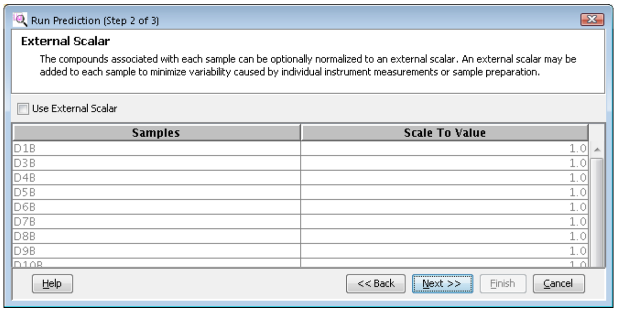
Figure 175 *Select Model page (Run Prediction (Step 1 of 3))*

3. Select whether to normalize the data in **Run Prediction (Step 2 of 3)**.

Normalizing the data reduces the variability caused by sample preparation and instrument response. You can use manually set external scaling value for each sample file. No scalar is used for this example class prediction model.

a Clear the **Use External Scalar** check box.

b Click **Next**. Your class prediction model is immediately run to classify the validation data files and a progress bar is displayed.



Run Prediction (Step 2 of 3)

External Scalar

The compounds associated with each sample can be optionally normalized to an external scalar. An external scalar may be added to each sample to minimize variability caused by individual instrument measurements or sample preparation.

□ Use External Scalar

Samples	Scale To Value
D1B	1.0
D3B	1.0
D4B	1.0
D5B	1.0
D6B	1.0
D7B	1.0
D8B	1.0
D9B	1.0
D10B	1.0

Figure 176 *External Scalar page (Run Prediction (Step 2 of 3))*

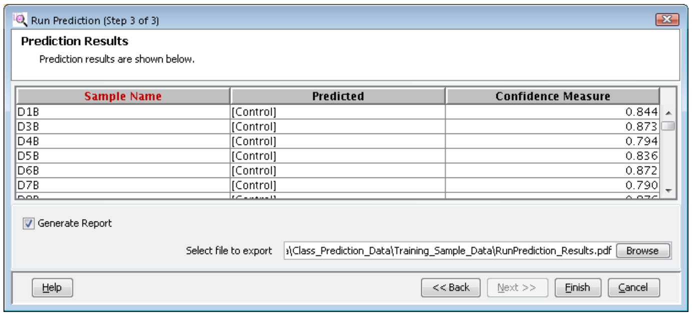
4. Review the prediction results in **Run Prediction (Step 3 of 3)**.

a Review the results of applying your class prediction model to the sample data files.

b Mark the **Generate Report** check box. This saves the results of running the class prediction model on the sample data files in PDF format.

c Type a meaningful file name for the PDF report in **Select file to export**. By default the report is saved in the same folder with the validation data files and the file name is the first sample name.

d Click **Finish**. A **Progress** dialog box is displayed while the report is generated.



Run Prediction (Step 3 of 3)

Prediction Results

Prediction results are shown below.

Sample Name	Predicted	Confidence Measure
D1B	[Control]	0.844
D3B	[Control]	0.873
D4B	[Control]	0.794
D5B	[Control]	0.836
D6B	[Control]	0.872
D7B	[Control]	0.790

☑ Generate Report

Select file to export a\Class_Prediction_Data\Training_Sample_Data\RunPrediction_Results.pdf

Figure 177 *Prediction Results page (Run Prediction (Step 3 of 3))*

MPP - Run Prediction from File

Classify samples that reside in a folder on your computer or on a network drive.

This operation was used when you validated your class prediction model. For a complete set of steps for this operation, see "Validate your prediction model" on page 141.

SCP - Project > Run Prediction

Classify samples that reside in a folder on your computer or on a network drive using the same **Run Prediction from File** wizard used by MPP. SCP is a separately licensed program.

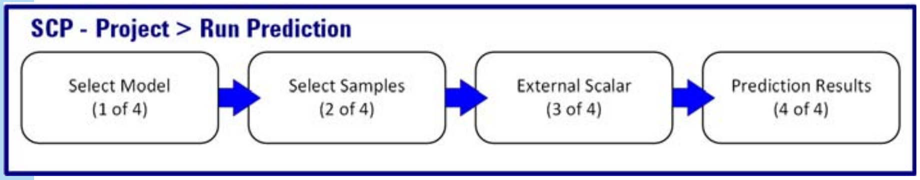
1. Launch Sample Class Predictor.

Double-click the Sample Class Predictor icon located on the desktop, or click **Start > All Programs > Agilent > SampleClassPredictor > Sample Class Predictor**.

2. Launch **Run Prediction**.

Click **Project > Run Prediction**.

The **Run Prediction** wizard has four (4) steps that are identical to the same four steps used by **Run Prediction from file** used by MPP. A flow chart of Run Prediction from within SCP is shown in Figure 178.



SCP - Project > Run Prediction

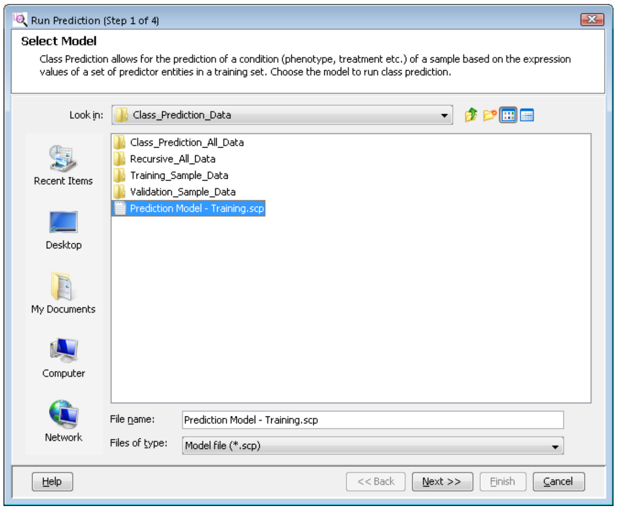
Select Model (1 of 4) → Select Samples (2 of 4) → External Scalar (3 of 4) → Prediction Results (4 of 4)

Figure 178 *Flow chart of the Run Prediction within SCP*

3. Select your prediction model in **Run Prediction (Step 1 of 4)**.

a Select the prediction model. The prediction model is highlighted when selected.

b Click **Next**.



Run Prediction (Step 1 of 4)

Select Model

Class Prediction allows for the prediction of a condition (phenotype, treatment etc.) of a sample based on the expression values of a set of predictor entities in a training set. Choose the model to run class prediction.

Look in: Class_Prediction_Data

- Class_Prediction_All_Data
- Recursive_All_Data
- Training_Sample_Data
- Validation_Sample_Data
- Prediction Model - Training.scp

File name: Prediction Model - Training.scp

Files of type: Model file (*.scp)

Figure 179 *Select Model page (Run Prediction (Step 1 of 4))*

4. Select the validation data to import in **Run Prediction (Step 2 of 4)**.

a Click **Select Data Files**.

b Select the sample data files to open.

Note: Orderly naming of the data files with respect to the parameters related to the independent variables helps you make sure that all of the data is selected.

c Click **Open**.

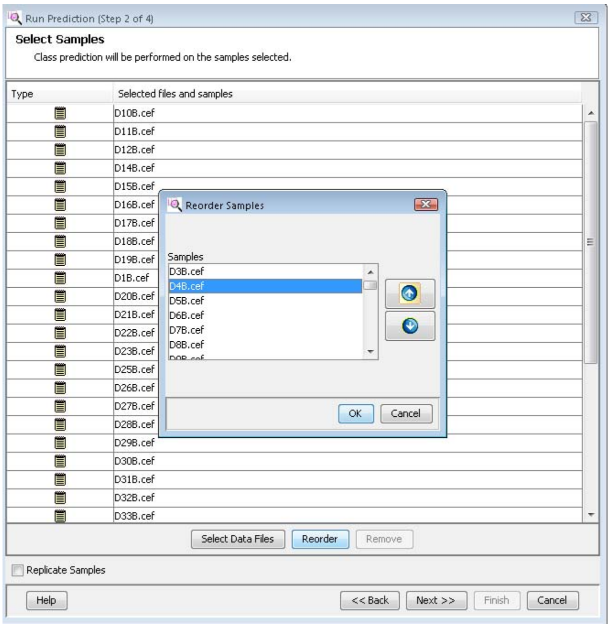
d Click **Reorder** to arrange the file list different from alphabetical order.

You can arrange the files, for example, in the same order of your original sample grouping (classifications) described in section "Review and order the selected files that are imported in the MS Experiment Creation Wizard (Step 5 of 11)." on page 73.

e Click the **Up** ⬆ or **Down** ⬇ buttons to reorder the selected sample or samples.

f Click **OK**.

g Click **Next**.



Run Prediction (Step 2 of 4)

Select Samples

Class prediction will be performed on the samples selected.

Type	Selected files and samples
▤	D10B.cef
▤	D11B.cef
▤	D12B.cef
▤	D14B.cef
▤	D15B.cef
▤	D16B.cef
▤	D17B.cef
▤	D18B.cef
▤	D19B.cef
▤	D1B.cef
▤	D20B.cef
▤	D21B.cef
▤	D22B.cef
▤	D23B.cef
▤	D25B.cef
▤	D26B.cef
▤	D27B.cef
▤	D28B.cef
▤	D29B.cef
▤	D30B.cef
▤	D31B.cef
▤	D32B.cef
▤	D33B.cef

Reorder Samples

Samples

D3B.cef
D4B.cef
D5B.cef
D6B.cef
D7B.cef
D8B.cef

□ Replicate Samples

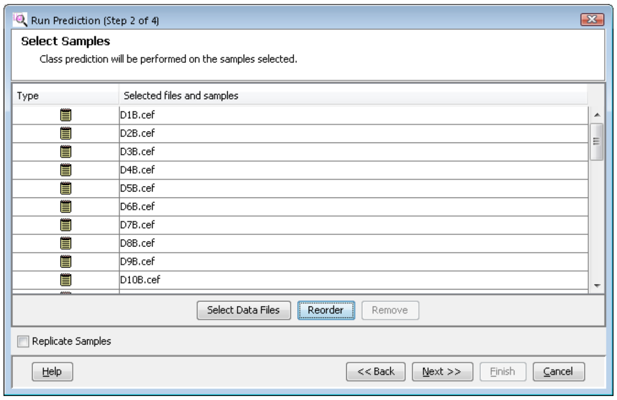
Figure 180 Select Samples page (Run Prediction (Step 2 of 4))

5. Select whether to normalize the data in **Run Prediction (Step 3 of 4)**.

Normalizing the data reduces the variability caused by sample preparation and instrument response. You can use manually set external scaling value for each sample file. No scalar is used for this example class prediction model.

a Clear the **Use External Scalar** check box.

b Click **Next**. Your class prediction model is immediately run to classify the sample data files. A **Progress** dialog box is displayed during the classification.



Run Prediction (Step 2 of 4)

Select Samples

Class prediction will be performed on the samples selected.

Type	Selected files and samples
	D1B.cef
	D2B.cef
	D3B.cef
	D4B.cef
	D5B.cef
	D6B.cef
	D7B.cef
	D8B.cef
	D9B.cef
	D10B.cef

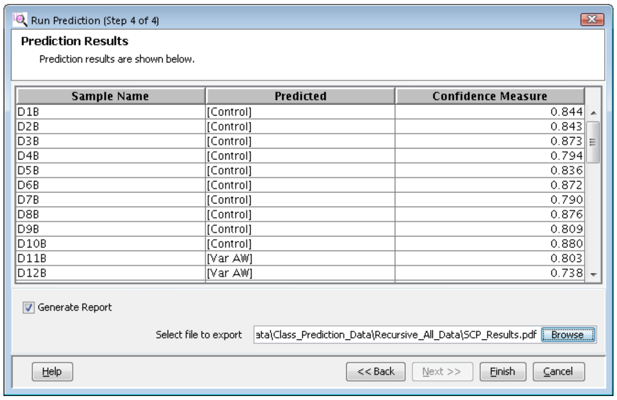
Figure 181 *External Scalar page (Run Prediction (Step 3 of 4))*

6. Review the prediction results in **Run Prediction (Step 4 of 4)**.

a Review the results of applying your class prediction model to the sample data files.

b Mark the **Generate Report** check box. This saves the results of running the class prediction model on the sample data files in PDF format.

c Type a meaningful file name for the PDF report in **Select file to export**. By default the report is saved in the same folder with the sample data files and the file name is the first sample name.



Run Prediction (Step 4 of 4)

Prediction Results

Prediction results are shown below.

Sample Name	Predicted	Confidence Measure
D1B	[Control]	0.844
D2B	[Control]	0.843
D3B	[Control]	0.873
D4B	[Control]	0.794
D5B	[Control]	0.836
D6B	[Control]	0.872
D7B	[Control]	0.790
D8B	[Control]	0.876
D9B	[Control]	0.809
D10B	[Control]	0.880
D11B	[Var AW]	0.803
D12B	[Var AW]	0.738

☑ Generate Report

Select file to export: ata\Class_Prediction_Data\Recursive_All_Data\SCP_Results.pdf

Figure 182 *Prediction Results page (Run Prediction (Step 4 of 4))*

d Click **Finish**. A **Progress** dialog box is displayed while the report is generated.

7. Review the saved report.

Open the saved report using an appropriate program that can open and view PDF files. The first page (Figure 171 on page 144) of the report contains the results of the prediction model applied to your sample data files. The subsequent pages contain PCA Score plots for each sample file placed with the training PCA Scores. Figure 172 on page 145 shows an example page for a sample data file.

Classify your acquisition data

Sample Class Predictor interfaces directly with ChemStation and MassHunter acquisition applications. During acquisition, your class prediction model is a targeted analysis that is applied to your sample data files as an automated classification tool.

Sample Class Predictor can run models on data acquired and processed in Agilent MassHunter or Agilent OpenLAB ChemStation Edition. The supported data file types are AMDIS, ChemStation, MassHunter, GC Scan and Generic. For Prediction on AMDIS data sources, both the FIN and corresponding ELU file should be present in the same working directory.

Samples must be from the same data source as those the model was built on. But they do not need to be from the same technology. When a prediction is run from a data file during acquisition, the targeted feature extraction uses the alignment values included in the file to perform alignments on entities in the new data.

1. Launch Acquisition.

Start your MassHunter or ChemStation data acquisition software.

The following example is for MassHunter Workstation Data Acquisition.

2. View the Method Editor window and SCP tab.

a Click **View > Method Editor** to display the method editor window.

b Click on the tab for the **Method Editor** window in the **Method Editor** if the Method Editor window is not active. The tab for the **Method Editor** is arranged with the tabs for **Worklist** and **Sample Run** at the bottom of the **Method Editor**.

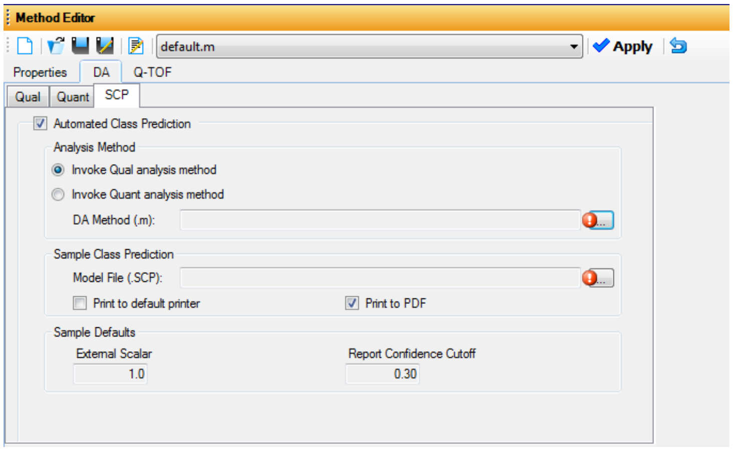
c Click the **DA** tab.

d Click on the **SCP** tab within the DA tab (see Figure 183).

3. Setup automated class prediction.

a Mark **Automatic Class Prediction** in the SCP tab in the Method Editor.

When **Automatic Class Prediction** is marked you must enter an analysis method and sample class prediction model (see Figure 183).



default.m

☑ Automated Class Prediction

Analysis Method

● Invoke Qual analysis method

○ Invoke Quant analysis method

DA Method (.m):

Sample Class Prediction

Model File (.SCP):

□ Print to default printer ☑ Print to PDF

Sample Defaults

External Scalar 1.0

Report Confidence Cutoff 0.30

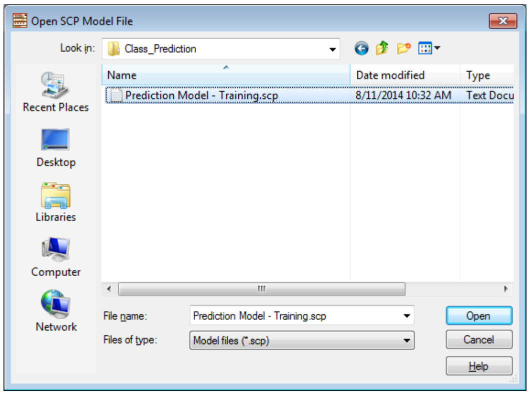
Figure 183 *SCP tab immediately after marking Automated Class Prediction*

b Click **Invoke Qual analysis method** to enable SCP to operate with the Qualitative Analysis Find Compounds by Formula that you used to perform a recursive feature finding (see “Recursive feature finding using Qualitative Analysis” on page 102).

c Click **Open** in the **Open Qual/Quant Method** dialog box.

d Browse to select your **Model File (SCP)** to apply on your acquired sample data.

e Click **Open** in the **Open SCP Model File** dialog box.



Look in: Class_Prediction

Name	Date modified	Type
Prediction Model - Training.scp	8/11/2014 10:32 AM	Text Docu

File name: Prediction Model - Training.scp

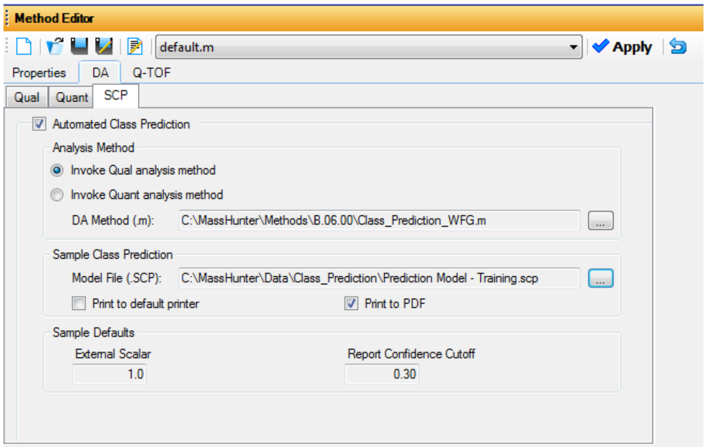
Files of type: Model files (*.scp)

Figure 184 *Open SCP Model File dialog box*

f Mark **Print to PDF** to save your class prediction results to a file. You can review and print the results after your acquisition has ended.

g Type an appropriate value for **External Scalar**.

h Type an appropriate value for **Report Confidence Cutoff**.



default.m

☑ Automated Class Prediction

Analysis Method

● Invoke Qual analysis method

○ Invoke Quant analysis method

DA Method (.m): C:\MassHunter\Methods\B.06.00\Class_Prediction_WFG.m

Sample Class Prediction

Model File (.SCP): C:\MassHunter\Data\Class_Prediction\Prediction Model - Training.scp

□ Print to default printer ☑ Print to PDF

Sample Defaults

External Scalar: 1.0

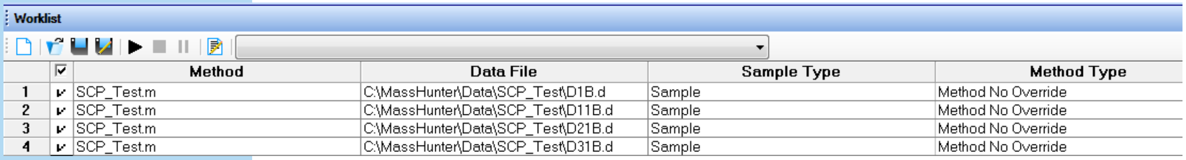
Report Confidence Cutoff: 0.30

Figure 185 *SCP tab with proper parameters for the example class prediction model*

4. *(Optional)* Test your SCP setup using a worklist.

a Click on the tab for the **Worklist** window at the bottom of the **Method Editor**.

b Create a worklist to process one of more of your original sample data files.



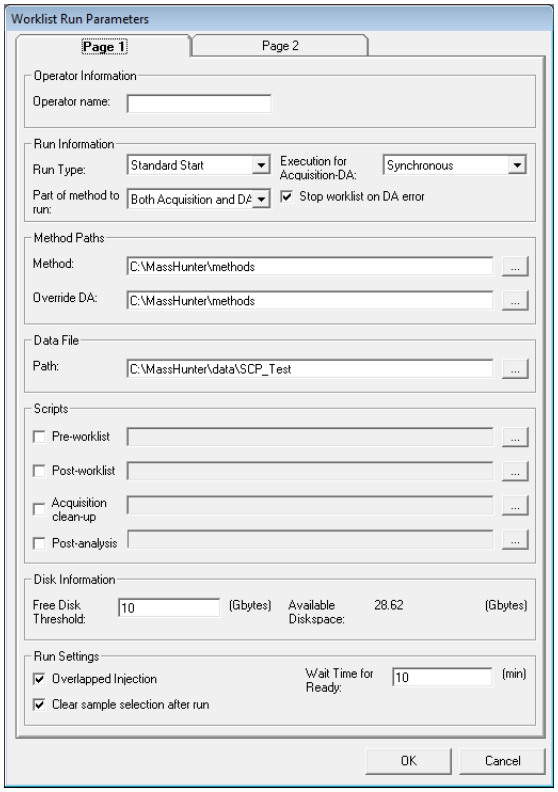
	☑	Method	Data File	Sample Type	Method Type
1	✓	SCP_Test.m	C:\MassHunter\Data\SCP_Test\D1B.d	Sample	Method No Override
2	✓	SCP_Test.m	C:\MassHunter\Data\SCP_Test\D11B.d	Sample	Method No Override
3	✓	SCP_Test.m	C:\MassHunter\Data\SCP_Test\D21B.d	Sample	Method No Override
4	✓	SCP_Test.m	C:\MassHunter\Data\SCP_Test\D31B.d	Sample	Method No Override

Figure 186 *Example workflow for class prediction*

c Click **Worklist > Worklist Run Parameters** to edit the worklist run parameters.

d Select **DA Only** for P**art of the method to run**.

e Click **OK**.



Worklist Run Parameters

Page 1 | Page 2

Operator Information

Operator name:

Run Information

Run Type: Standard Start

Execution for Acquisition-DA: Synchronous

Part of method to run: Both Acquisition and DA

☑ Stop worklist on DA error

Method Paths

Method: C:\MassHunter\methods

Override DA: C:\MassHunter\methods

Data File

Path: C:\MassHunter\data\SCP_Test

Scripts

☐ Pre-worklist

☐ Post-worklist

☐ Acquisition clean-up

☐ Post-analysis

Disk Information

Free Disk Threshold: 10 (Gbytes)

Available Diskspace: 28.62 (Gbytes)

Run Settings

☑ Overlapped Injection

☑ Clear sample selection after run

Wait Time for Ready: 10 (min)

OK Cancel

Figure 187 *Worklist Run Parameters - Page 1*

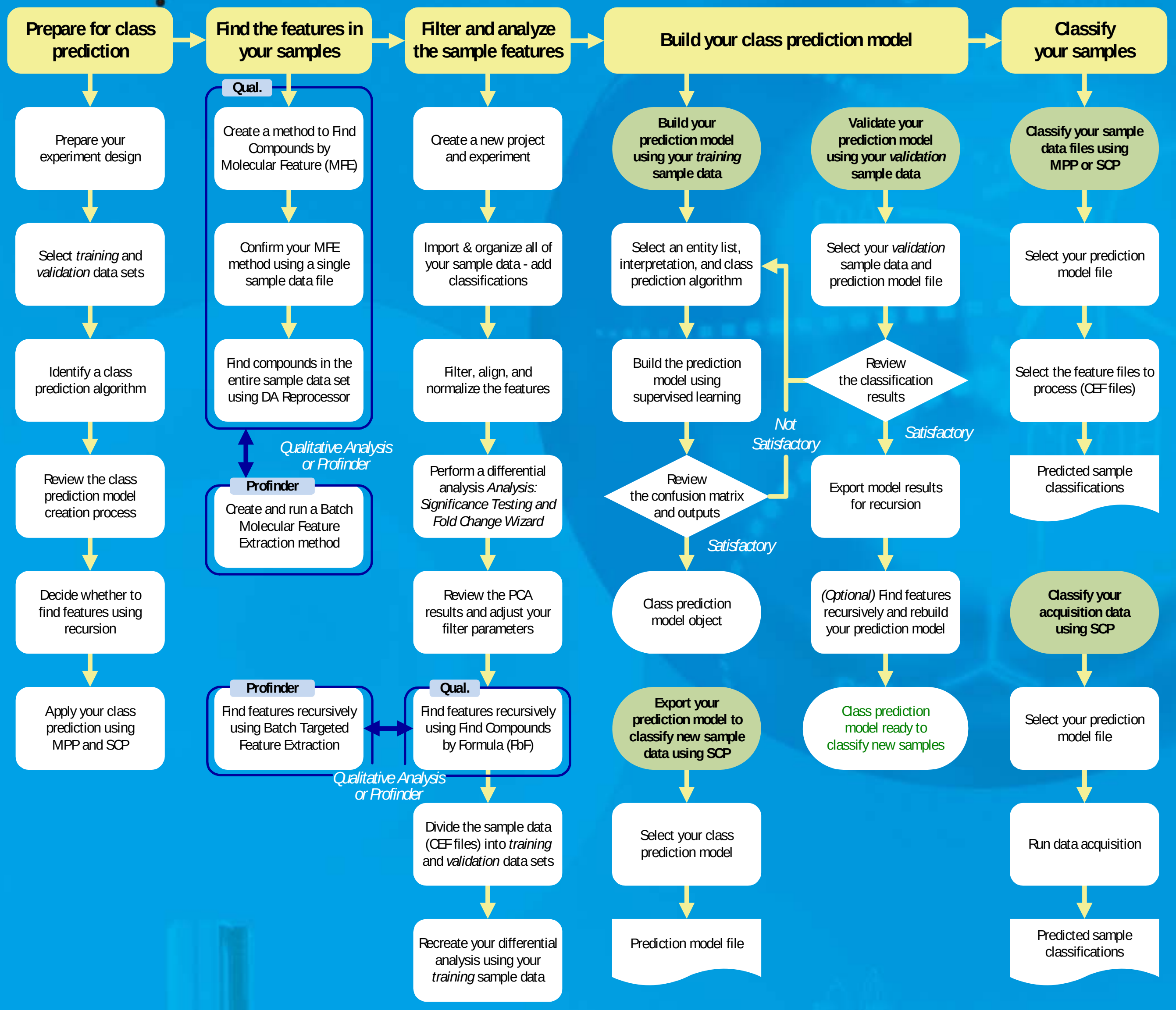
f Click **Worklist > Run** to run the worklist and verify operation of automated SCP.

Conclusion

You have now completed the class prediction workflow.

Reference information

This chapter consists of definitions and references. The definitions section includes a list of terms and their definitions as used in this workflow. The references section includes citations to Agilent publications that help you use Agilent products and perform class prediction analyses.



Prepare for class prediction
- Prepare your experiment design
- Select *training* and *validation* data sets
- Identify a class prediction algorithm
- Review the class prediction model creation process
- Decide whether to find features using recursion
- Apply your class prediction using MPP and SCP

Find the features in your samples

Qual.
- Create a method to Find Compounds by Molecular Feature (MFE)
- Confirm your MFE method using a single sample data file
- Find compounds in the entire sample data set using DA Reprocessor

Qualitative Analysis or Profinder

Profinder
- Create and run a Batch Molecular Feature Extraction method

Profinder
- Find features recursively using Batch Targeted Feature Extraction

Qualitative Analysis or Profinder

Filter and analyze the sample features
- Create a new project and experiment
- Import & organize all of your sample data - add classifications
- Filter, align, and normalize the features
- Perform a differential analysis *Analysis: Significance Testing and Fold Change Wizard*
- Review the PCA results and adjust your filter parameters

Qual.
- Find features recursively using Find Compounds by Formula (FbF)
- Divide the sample data (CEF files) into *training* and *validation* data sets
- Recreate your differential analysis using your *training* sample data

Build your class prediction model
- Build your prediction model using your *training* sample data
- Select an entity list, interpretation, and class prediction algorithm
- Build the prediction model using supervised learning
- Review the confusion matrix and outputs
- *Satisfactory*
- Class prediction model object
- Export your prediction model to classify new sample data using SCP
- Select your class prediction model
- Prediction model file

Not Satisfactory

- Validate your prediction model using your *validation* sample data
- Select your *validation* sample data and prediction model file
- Review the classification results
- *Satisfactory*
- Export model results for recursion
- *(Optional)* Find features recursively and rebuild your prediction model
- Class prediction model ready to classify new samples

Classify your samples
- Classify your sample data files using MPP or SCP
- Select your prediction model file
- Select the feature files to process (CEF files)
- Predicted sample classifications
- Classify your acquisition data using SCP
- Select your prediction model file
- Run data acquisition
- Predicted sample classifications

Agilent Technologies

Definitions

This section contains a list of terms and their definitions as used in this workflow. Review of the terms and definitions presented in this section helps you understand the software wizards and the class prediction workflow.

Abundance profile

Relative or absolute signal intensities of the extracted compounds identified in your sample derived from the chromatographic/mass spectral data.

Algorithm

Mathematical calculations and related parameters that are applied to sample data to produce a result that may be used to represent or classify a sample.

Alignment

Adjustment of the chromatographic retention time of eluting components to improve the correlation among data sets, based on the elution of specific component(s) that are (1) naturally present in each sample or (2) deliberately added to the sample through spiking the sample with a known compound or set of compounds that does not interfere with the sample.

AMDIS

Acronym for automated mass spectral deconvolution and identification system developed by NIST (http://www.amdis.net).

Amino acid

Biologically significant molecules that contain a core carbon positioned between a carboxyl and amine group in addition to an organic substituent. Dual carboxyl and amine functionalities facilitate the formation of peptides and proteins.

ANOVA

Abbreviation for analysis of variance which is a statistical method that simultaneously compares the means between two or more attributes or parameters of a data set. ANOVA is used to determine if a statistical difference exists between the means of two or more data sets and thereby prove or disprove the hypothesis. See also t-Test.

Attribute

Another term for an independent variable. Referred to as a parameter and is assigned a parameter name during the various steps of the metabolomic data analysis.

Attribute value

Another term for one of several values within an attribute for which exist correlating samples. Referred to as a condition or a parameter value and given an assigned value during the various steps of the metabolomic data analysis.

Baselining

A technique used to view and compare data that involves converting the original data values to values that are expressed as changes relative to a calculated statistical value derived from the data. The calculated statistical value is referred to as the baseline.

Bayesian

A term used to refer to statistical techniques named after the Reverend Thomas Bayes (ca. 1702 - 1761).

Bayesian inference

The use of statistical reasoning, instead of direct facts, to calculate the probability that a hypothesis may be true. Also known as Bayesian statistics.

Bioinformatics

The use of computers, statistics, and informational techniques to increase the understanding of biological processes.

Biomarker

An organic molecule whose presence and concentration in a biological sample indicates a normal or altered function of higher level biological activity.

Carbohydrate

An organic molecule consisting entirely of carbon, hydrogen, and oxygen that is important to living organisms.

CEF file

A binary file format called a compound exchange file (CEF) that is used to exchange data between Agilent software. In the class prediction and metabolomics workflows CEF files are used to share molecular features between MassHunter Qualitative Analysis and Mass Profiler Professional.

Cell

The fundamental unit of an organism consisting of several sets of biochemical functions within an enclosing membrane. Animals and plants are made of one or more cells that combine to form tissues and perform living functions.

Census

Collection of a sample from every member of a population.

Cheminformatics

The use of computers and informational techniques (such as analysis, classification, manipulation, storage, and retrieval) to analyze and solve problems in the field of chemistry.

Chemometrics

A science employing mathematical and analytical processes to extract information from chemical data sets. The processes involve interactive applications of techniques employed in disciplines such as multivariate statistics, applied mathematics, and computer science to obtain meaningful information from complex data sets. Chemometrics is typically used to obtain meaningful information from data derived from chemistry, biochemistry and chemical engineering. Mass Profiler Professional is designed to employ chemometrics processes to GC/MS and LC/MS data sets to obtain useful information.

Child

A subset of information that is created by an algorithm from an original set of information. An entity list created using Mass Profiler Professional is a child. An original entity list is referred to as the parent of one or more child entity lists.

Class

A grouping of samples organized for a study based on having a similar or identical likeness such as origin, condition (age, disease, treatment, extraction procedure), or another trait that is relevant to an experiment.

Class prediction

Steps employed using Mass Profiler Professional and Sample Class Predictor to classify samples from mass spectrometry data. Class prediction is a supervised learning method that involves three steps: build your model using known samples, validate your model using known samples that were not used during the model creation, and apply your prediction model to samples with unknown class membership.

Classification

The process of assigning a class membership to a sample.

Classify

Assign class membership to a sample.

Co-elution

When compounds elute from a chromatographic column at nominally the same time making the assignment of the observed ions to each compound difficult.

Complex

Class of compounds consisting of two or more proteins that physically bind each other. Their combined form is biologically active and stable.

Composite spectrum

A compound spectrum generated to represent the molecular feature that includes more than one ion, isotope, or adduct (not just M + H) and is used by Mass Profiler Professional for recursive analysis and ID Browser.

Compound

A metabolite that may be individually referred to as a compound, descriptor, element, entity, feature, or metabolite during the various steps presented during various workflows used by MPP.

Condition

Another term for one of several values within a parameter for which exist correlating samples. Condition may also be referred to as a parameter value during the various steps of the metabolomic data analysis. See also attribute value.

Confusion matrix

A data table used to assess the ability of a prediction model to correctly predict the classification of validation sample data. The rows of the table represent the actual classification of the validation samples and the columns of the table represent the predicted classifications. For validation algorithm outputs, the results show a cumulative Confusion Matrix, which is the sum of confusion matrices for individual runs of the learning algorithm. For training algorithm outputs, the Confusion Matrix is the result of applying the prediction model to the training sample data.

Data

Information that represents in a form suitable for storing and processing by a computer the qualitative or quantitative attributes of a subject. Examples include GC/MS and LC/MS data consisting fundamentally of time, ion m/z, and ion abundance from a chemical sample.

Data processing

Conversion of data into meaningful information. Computers are employed to enable rapid recording and handling of large amounts of data, i.e., MassHunter Workstation and Mass Profiler Professional.

Data reduction

See reduction.

Deconvolution

The technique of reconstructing individual mass and mass spectral data from co-eluting compounds.

Dependent variable

An element in a data set that can only be observed as a result of the influence from the variation of an independent variable. For example, a pharmaceutical compound structure and quantity may be controlled as two independent variables while the metabolite profile presents a host of small-molecule products that make up the dependent variables of a study.

Determinate

Having exact and definite limits on an analytical result that provides a conclusive degree of correlation of the subject to the specimen.

Element

A metabolite that may be individually referred to as a compound, molecular feature, element, or entity during the various steps of the metabolomic data analysis.

Endogenous

Pertaining to cause, development, or origination from within an organism.

Entity

A compound that may be individually referred to as a descriptor, element, feature, or metabolite during the various steps presented during various workflows used by MPP.

Entity list

The compounds (also referred to as elements, entities, or features) that satisfy the parameters of an analysis specified by each experiment performed on your data. Entity lists are viewed in the Experiment Navigator. When you create a prediction model, an entity list provides the model with the features that you have previously determined contribute the most change among your classifications.

Environment

The natural and/or controlled conditions that surround an organism as it lives and operates.

Enzyme

Proteins acting as biocatalysts in a metabolomic reaction. These entities are particularly important in depicting a biochemical network.

Experiment

Data acquired in an attempt to understand causality where tests or analyses are defined and performed on an organism to discover something that is not yet known, to demonstrate as proof of something that is known, or to find out whether something is effective.

Externality

A quality, attribute, or state that originates and/or is established independently from the specimen under evaluation.

Extraction

The process of retrieving a deliberate subset of data from a larger data set whereby the subset of the data preserves the meaningful information, not the redundant and less meaningful information. Also known as data extraction.

Feature

Independent, distinct characteristic of a phenomena and data under observation. Features are an important part of the identification of patterns - pattern recognition - within data whether processed by a human or by artificial intelligence, such as MassHunter Workstation and Mass Profiler Professional. During the various workflows used by MPP a feature may be individually referred to as a compound, descriptor, element, entity, or metabolite.

Feature extraction

The reduction of data size and complexity through the removal of redundant and non-specific data by using the important variables (features) associated with the data. Careful feature extraction yields a smaller data set that is more easily processed without any compromise in the information quality.

Feature selection

The identification of important, or non-important, variables and the variable relationships in a data set using both analytical and a priori knowledge about the data.

Filter

The process of establishing criteria by which entities are removed (filtered) from further analysis during the class prediction and metabolomics workflows.

Filter by flag

A flag is a term used to denote a quality of an entity within a sample. A flag indicates if the entity was detected in each sample as follows: Present means the entity was detected, Absent means the entity was not detected, and Marginal means the signal for the entity was saturated.

Find minimal entities

A machine learning technique that identifies the entities within a list that best distinguish the conditions within an interpretation. Find minimal entities is best applied when a large number of features and a strong inter-correlations exists among the entities within the conditions. The algorithms employed to identify the minimal entities that distinguish the conditions are identical to those used in supervised learning within class prediction. An entity list created using find minimal entities is not recommended for use with class prediction to prevent the creation of an under-constrained prediction model unless the number of replicates is large and/or you created the minimal entity list using the genetic algorithm.

Function

Biological purpose or activity.

Functional classification

Classification of samples that relate to a particular biological action that is under evaluation in your experiment.

Genotype

The genetic makeup of an individual organism.

Hypothesis

A proposition made to explain certain facts and tentatively accepted to provide a basis for further investigation. A proposed explanation for observable phenomena may or may not be supported by the analytical data. Statistical data analysis is performed to quantify the probability that the hypothesis is true. Also known as the scientific hypothesis.

Hypothetical

A statement based on, involving, or having the nature of a hypothesis for the purposes of serving as an example and not necessarily based on an actuality.

ID Browser

Software that automatically annotates the entity list with the compound names and adds them to any of the various visualization and pathway analysis tools.

Identified compound

Chromatographic components that have an assigned, exact identity, such as compound name and molecular formula, based on prior assessment or comparison with a database. See also Unidentified Compound.

Independent variable

An essential element, constituent, attribute, or quality in a data set that is deliberately controlled in an experiment. For example, a pharmaceutical compound structure and quantity may be controlled as two independent variables while the metabolite profile presents a host of independent small molecule products that make up the dependent variables of a study. An independent variable may be referred to as a parameter and is assigned a parameter name during the various steps of the metabolomic data analysis.

Interpretation

An interpretation specifies how samples are related (grouped) into experimental conditions for display and treatment by the analysis. Samples with the same parameter value are grouped into a single experimental condition. The “All Samples” interpretation contains all of the samples used in the creation of the experiment. Custom interpretations include only the samples that meet selected experimental conditions (parameters) and their respective values (parameter values). When you create a prediction model, an interpretation provides the model with the known classifications.

Latent variables

Variables in a mathematical model that are often not themselves observable or measurable in an experiment (also referred to as hidden variables). Latent variables may represent multiple physical or measurable characteristics of a sample that reduce the dimensionality of the analysis and therefore help present a less complex representation of the underlying relationships among classifications.

Lipidomics

Identification and quantification of cellular lipids from an organism in a specified biological situation. The study of lipids is a subset of metabolomics.

Membership

A sample having known or identified attributes that make the sample part of a specific class.

Mass variation

Using the mass to charge (m/z) resolution to improve compound identification. Compounds with nearly identical and identical chromatographic behavior are deconvoluted by adjusting the m/z range for extracting ion chromatograms.

Mean

The numerical result of dividing the sum of the data values by the number of individual data observations.

Metabolism

The chemical reactions and physical processes whereby living organisms convert ingested compounds into other compounds, structures, energy and waste.

Metabolite

Small organic molecules that are intermediate compounds and products produced as part of metabolism. Metabolites are important modulators, substrates, byproducts, and building blocks of many different biological processes. In order to distinguish metabolites from larger biological molecules, known as macromolecules such as proteins, DNA and others, metabolites are typically under 1000 Da. A metabolite may be individually referred to as a compound, molecular feature, element, or entity during the various steps of the metabolomic data analysis.

Metabolome

The complete set of small-molecule metabolites that may be found within a biological sample. Small molecules are typically in the range of 50 to 600 Da.

Metabolomics

The process of identification and quantification of all metabolites of an organism in a specified biological situation. The study of the metabolites of an organism presents a chemical “fingerprint” of the organism under the specific situation. See Metabonomics for the study of the change in the metabolites in response to externalities.

Metabonomics

The metabolic response to externalities such as drugs, environmental factors, and disease. The study of metabonomics by the medical community may lead to more efficient drug discovery and to individualized patient treatment. Meaningful information learned from the metabolite response can be used for clinical diagnostics or for understanding the onset and progression of human diseases. See Metabolomics for the identification and quantitation of metabolites.

Model

A mathematical and statistical algorithm that produces a recommendation for the sample class membership. A model can be transported from the software that generated the algorithm to software that can apply the algorithm to new samples.

NLP

Natural Language Processing (NLP) algorithm that extracts information from published literature.

Normalization

A technique used to adjust the ion intensity of mass spectral data from an absolute value based on the signal measured at the detector to a relative intensity of 0 to 100 percent based on the signal of either (1) the ion of the greatest intensity or (2) a specific ion in the mass spectrum.

Null hypothesis

The default position taken by the hypothesis that no effect or correlation of the independent variables exists with respect to the measurements taken from the samples. An example null hypothesis: “No effect or correlation exits between a change in the independent variables (e.g., treatment) and a change in the dependent variables (e.g., metabolic profile).”

Observation

Data acquired in an attempt to understand causality where no ability exists to (1) control how subjects are sampled and/or (2) control the exposure each sample group receives.

One-hit wonder

An entity that appears in only one sample, is absent from the replicate samples, and does not provide any utility for statistical analysis. Entities that are one-hit wonders may be filtered using Filter by Flags.

Organism

A group of biochemical systems that function together as a whole thereby creating an individual living entity, such as an animal, plant, or microorganism. Individual living entities may be multicellular or unicellular. See also specimen.

Overfit

A prediction model that is so complex that it has poor predictive performance and can exaggerate meaning of minor variations in the sample data. For an independent sample of validation data taken from the same population as the training data, the prediction model typically does not fit the validation data as well as it fits the training data. Overfitting occurs more often when the size of the training data set is small, or when the number of parameters in the model is large.

p-value

The probability of obtaining a statistical result that is comparable to or greater in magnitude than the result that was actually observed, assuming that the null hypothesis is true. The null hypothesis is stated that no correlation exists between the independent variables and the measurements taken from the samples. Rejection of the null hypothesis is typically made when the p-value is less than 0.05 or 0.01. A p-value of 0.05 or 0.01 may be restated as a 5% or 1% chance of rejecting the null hypothesis when it is true. When the null hypothesis is rejected, the result is said to be statistically significant meaning that a correlation exists between the independent variables and the measurements as specified in the hypothesis.

Parameter

Another term for an independent variable. Referred to as a parameter or parameter name and is assigned a parameter name during the various steps of the metabolomic data analysis. See also condition and attribute.

Parameter value

Another term for one of several values within a parameter for which exist correlating samples. Parameter value may also be referred to as a condition during the various steps of the metabolomic data analysis. See also attribute value.

Parent

The original set of information that is processed by an algorithm to create one or more subsets of information. A subset entity list is referred to as the child of a parent entity list.

Peptide

Linear chain of amino acids that is shorter than a protein. The length of a peptide is sufficiently short that it is easily made synthetically from the constituent amino acids.

Peptide bond

The covalent bond formed by the reaction of a carboxyl group with an amine group between two molecules, e.g., between amino acids.

Permutation

Any of the total number of subsets that may be formed by the combination of individual parameters among the independent variables. For example the number of permutations of A and B in variable Φ in combination with X, Y, and Z in variable θ equals six (6 = 2 x 3) and may be represented as AX, AY, AZ, BX, BY, and BZ. Note that the permutations within a variable are not relevant such as AB, XY, XZ, and YZ.

Phenotype

Measurable and observable characteristics of a sample from an organism as a result of its genotypic interaction with the environment.

Polarity

The condition of an effect as being positive or negative, additive or subtractive, with respect to some point of reference, such as with respect to the concentration of a metabolite.

Pooled sample

When the amount of available biological material in very small samples may be combined into a single sample (pooled) and then split into different aliquots for multiple analyses. By pooling the sample, sufficient material exists to obtain replicate analyses of each sample where formerly there was insufficient material to obtain replicate analytical results. The trade-off loss of information about the biological variation that was formerly present in each unique sample is offset by a gain in statistical significance of the results.

Prediction

A statistically significant conclusion made regarding the identity and/or characteristics of a sample by using knowledge from the prior evaluation of samples with known identities and/or characteristics.

Principal component

Transformed data into axes, or principal components, so that the patterns between the axes most closely describe the relationships between the data. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The principal components often may be viewed, and interpreted, most readily in graphical axes with additional dimensions represented by color and/or shape representing the key elements (independent variables) of the hypothesis. This is part of the principal component analysis process employed by Mass Profiler Professional.

Principal component analysis

The mathematical process by which data containing a number of potentially correlated variables is transformed into a data set in relation to a smaller number of variables called principal components which account for the most variability in the data. The result of the data transformation leads to the identification of the best explanation of the variance in the data, e.g. identification of the meaningful information. Also known as PCA.

Proteomics

The study of the structure and function of proteins occurring in living organisms.

Quality

A feature, attribute, and/or characteristic element whose presence, absence, or inability to be properly ascertained because of instrumental factors, contributes to the assessment of whether a sample is or is not representative of the larger specimen.

Recursive

Reapplying the same algorithm to a subset of a previous result in order to generate an improved result.

Recursive feature finding

A multi-step process in the class prediction and metabolomics workflows that improves the accuracy of finding statistically significant features in sample data files. Step 1: Find untargeted compounds by molecular feature in MassHunter Qualitative Analysis. Step 2: Filter the molecular features in Mass Profiler Professional. Step 3: Find targeted compounds by formula in MassHunter Qualitative Analysis. Importing the most significant features identified using Mass Profiler Professional back into MassHunter Qualitative Analysis as targeted features improves the accuracy in finding these features from the original sample data files.

Reduction

The process whereby the number of variables in a data set is decreased to improve computation time and information quality, such as an extracted ion chromatogram obtained from GC/MS and LC/MS data files. Reduction provides smaller, viewable and interpretable data sets by employing feature selection and feature extraction. Also known as dimension reduction and data reduction. This is part of most of the advanced processes employed by Mass Profiler Professional.

Regression analysis

Mathematical techniques for analyzing data to identify the relationship between dependent and independent variables present in the data. Information is gained from the estimation, regression, or the sign and proportionality of the effects of the independent variables on the dependent variables. This is part of the principal component analysis process employed by Mass Profiler Professional. Also known as regression.

Replicate

Multiple identical samples collected from a population so that the sample evaluation results in a value that more closely approximates the true value.

Sample

A part, piece, or item that is taken from a specimen and understood as being representative of the larger specimen (e.g., blood sample, cell culture, body fluid, aliquot) or population. An analysis may be derived from samples taken at a particular geographical location, taken at a specific period of time during an experiment, or taken before or after a specific treatment. A small number of specimens used to represent a whole class or group.

Sample class prediction

Assigning class membership of a sample based on the results of applying an algorithm that was developed through the prior analysis of samples with known class membership. Also known as a workflow used to build a model and classify samples from mass spectrometry data. Also see Class prediction.

Sampling

The process of taking samples that have a statistical representation of the population under evaluation.

Signature

An identifying characteristic of a sample that is distinctive of all samples from the same class.

Specimen

An individual organism, e.g., a person, animal, plant, or other organism, of a class or group that is used as a representative of a whole class or group.

Spike

The specific and quantitative addition of one or more compounds to a sample.

Standard

A chemical or mixture of chemicals used to compare the quality of analytical results or to measure and compensate the precise offset or drift incurred over a set of analyses.

Standard deviation

A measure of variability among a set of data that is equal to the square root of the arithmetic average of the squares of the deviations from the mean. A low standard deviation value indicates that the individual data tend to be very close to the mean, whereas a high standard deviation indicates that the data is spread out over a larger range of values from the mean.

State

A set of circumstances or attributes characterizing a biological organism at a given time. A few sample attributes may include temperature, time, pH, nutrition, geography, stress, disease, and controlled exposure.

Statistics

The mathematical process employed in manipulating numerical data from scientific experiments to derive meaningful information. This is part of the principal component analysis, t-test, and ANOVA processes employed by Mass Profiler Professional.

Subject

A chemical or biological sample taken from a specimen, or a whole specimen, that undergoes a treatment, experiment, or an analysis for the purposes of further understanding.

Supervised learning

A process employed in data analysis that uses knowledge of the phenotype to simplify the data (i.e., reduce the number of entities) to retain the entities that provide the best correlation to the characteristics (conditions) involved in the particular analysis. The goal of supervised learning is to identify a mathematical relationship that accurately associates the entities in your samples to the conditions in your interpretation; for example, when you are evaluating qualitative samples representing diseased versus healthy samples, or when you are evaluating quantitative samples representing degree of disease progression or response to therapy. Unsupervised learning techniques are applied to your data when the experiment involves samples with unknown relationships to the conditions.

Survey

Collection of samples from less than the entire population in order to estimate the population attributes.

t-Test

A statistical test to determine whether the mean of the data differs significantly from that expected if the samples followed a normal distribution in the population. The test may also be used to assess statistical significance between the means of two normally distributed data sets. See also ANOVA.

Training data set

Samples that have known functional or sample classification and are used to create a class prediction model.

Unidentified compound

Chromatographic components that are only uniquely denoted by their mass and retention times and which have not been assigned an exact identity, such as compound name and molecular formula. Unidentified compounds are typically produced by feature finding and deconvolution algorithms. See also Identified Compound.

Validate

To prove the correctness of the sample classification resulting from the application of an algorithm to known samples that were not part of the training data set, and judging the results based on the desired statistical significance.

Validation

The process of evaluating a class prediction algorithm by comparing the sample classification results to the known class membership.

Validation data set

Samples that have known functional or sample classification, that were not part of the training data set, and that are used to validate a class prediction model.

Variable

An element in a data set that assumes changing values, e.g., values that are not constant over the entire data set. The two types of variables are independent and dependent.

Volume

The area of the extracted compound chromatogram (ECC). The ECC is formed from the sum of the individual ion abundances within the compound spectrum at each retention time in the specified time window. The compound volume generated by MFE is used by Mass Profiler Professional to make quantitative comparisons.

Wizard

A sequence of dialog boxes presented by Mass Profiler Professional that guides you through well-defined steps to enter information, organize data, and perform analyses.

References

This section consists of citations to Agilent manuals, primers, application notes, presentations, product brochures, technical overviews, videos, and software that help you use Agilent products and perform your class prediction and metabolomics analyses.

Manuals

- Agilent G3835AA MassHunter Profinder Software - Quick Start Guide (G3835-90014, Revision A, December 2013)
- Integrated Biology with Agilent Mass Profiler Professional - Workflow Guide (5991-1909EN, Revision A, June 2013)
- Integrated Biology with Agilent Mass Profiler Professional - Workflow Guide Overview (5991-1910EN, Revision A, June 2013)
- Agilent MassHunter Workstation Software Qualitative Analysis - Familiarization Guide (G3335-90156, Revision A, April 2013)
- Agilent MassHunter Workstation Software Quantitative Analysis - Familiarization Guide (G3335-90152, February 2013)
- Agilent G3835AA MassHunter Mass Profiler Professional - Quick Start Guide (G3835-90009, Revision A, November 2012)
- Agilent G3835AA MassHunter Mass Profiler Professional - Familiarization Guide (G3835-90010, Revision A, November 2012)
- Agilent G3835AA MassHunter Mass Profiler Professional - Application Guide (G3835-90011, Revision A, November 2012)
- Agilent Metabolomics Workflow - Discovery Workflow Guide (5990-7067EN, Revision B, October 2012)
- Agilent Metabolomics Workflow - Discovery Workflow Overview (5990-7069EN, Revision B, October 2012)
- Agilent Mass Profiler Professional - Manual (January 2012)

Primers

- Proteomics: Biomarker Discovery and Validation (5990-5357EN, February 11, 2010)
- Metabolomics: Approaches Using Mass Spectrometry (5990-4314EN, October 27, 2009)

Application Notes

- Detecting Contamination in Shochu Using the Agilent GC/MSD, Mass Profiler Professional, and Sample Class Prediction Models (5991-0975EN, August 2, 2012)
- Metabolomic Profiling of Wines using LC/QTOF MS and MassHunter Data Mining and Statistical Tools (5990-8451EN, June 22, 2011)
- Multi-omic Analysis with Agilent's GeneSpring 11.5 Analysis Platform (5990-7505EN, March 25, 2011)

- An LC/MS Metabolomics Discovery Workflow for Malaria-Infected Red Blood Cells Using Mass Profiler Professional Software and LC-Triple Quadrupole MRM Confirmation (5990-6790EN, November 19, 2010)
- Profiling Approach for Biomarker Discovery using an Agilent HPLC-Chip Coupled with an Accurate-Mass Q-TOF LC/MS (5990-4404EN, October 20, 2009)
- Metabolite Identification in Blood Plasma Using GC/MS and the Agilent Fiehn GC/MS Metabolomics RTL Library (5990-3638EN, April 1, 2009)
- Metabolomic Profiling of Bacterial Leaf Blight in Rice (5989-6234EN, February 14, 2007)

Presentations

- Advances in Instrumentation and Software for Metabolomics Research (Advances in Instrumentation and Software for Metabolomics.pdf, September 18, 2012)
- Workflows to Support Automated Class Prediction with Complex Samples (WP20_405__Workflows_Support_Automated_Class_Prediction.pdf, June 25, 2012)
- Multi-omics Analysis Software for Targeted Identification of Key Biological Pathways (May 3, 2012)
- Predictive Classification of Contaminants Encountered During the Distillation of Shochu, a Distilled Beverage Native to Japan (ASMS_2011_ThP_316.pdf, June 23, 2011)
- Metabolomics LCMS Approach to: Identifying Red Wines according to their variety and Investigating Malaria infected red blood cells (November 3, 2010)
- Small Molecule Metabolomics (November 3, 2010)
- Presentation: Metabolome Analysis from Sample Prep through Data Analysis (November 3, 2010)

Product Brochures

- Assess Food Quality and Point of Origin (5991-0900EN, October 5, 2012)
- Emerging Insights: Agilent Solutions for Metabolomics (5990-6048EN, April 30, 2012)
- Agilent Mass Profiler Professional Software - Discover the Difference in your Data (5990-4164EN, April 27, 2012)
- Pathways to Insight - Integrated Biology at Agilent (5991-0222EN, March 30, 2012)
- Confidently Better Bioinformatics Solutions (5990-9905EN, February 2, 2012)
- Integrated Biology from Agilent: The Future is Emerging (5990-6047EN, September 1, 2010)

www.agilent.com

Revision A, August 2014

5991-1911EN

Agilent Technologies