Analyze IQ Lab
Version 3.0

Document Version: 3.0-2014-02-28

Copyright © Analyze IQ Limited, 2008-2014.
All Rights Reserved.

Table of Contents

1 Introduction 4
1.1 Overview of the Analyze IQ Product Suite 4
1.2 Overview of this Manual 5
2 Installation 7
2.1 USB Key Installation 7
2.2 Website 7
2.3 Getting a License 7
2.4 Upgrading Your Existing Installation 8
2.5 Help Menu 8
2.5.1 Help Menu - *About Analyze IQ Lab* 9
2.5.2 Help Menu - *Get Support* 9
2.5.3 Help Menu – *Getting Started* 10
2.5.4 Help Menu – *User Manual* 10
2.5.5 Help Menu – *Update License* 10
3 Principles of Operation 11
4 Perform an Analysis 13
4.1 Getting Started with Analyze IQ Lab 13
4.2 Using a Model for Analysis 14
4.2.1 Open a Spectrum File 14
4.2.2 File Formats 15
4.2.3 Acquire from Instrument 17
4.2.4 Save a Spectrum 18
4.2.5 Apply Model 19
4.2.6 View Model Results 19
4.3 Pre-Processing Opened Spectra 23
4.3.1 Pre-Processing Methods Available 23
4.3.2 Settings for Smooth and Derivative Pre-Processing 24
4.3.3 Example of Applying Pre-Processing 25
4.4 Viewing Spectrum Details: SPC Format 26
4.5 Viewing Spectrum Details: AIQ or SpectroML Format 27
5 Build a New Model (Classification Example) 29
5.1 Switch To Build New Model Page 29
5.2 Preparing the Dataset 31
5.2.1 Spectra Manager 31
5.2.2 Import from Single File 43
5.2.3 Import from Multiple Spectrum Files 49
5.2.4 Importing Data From Unscrambler 50
5.3 Apply Pre-processing to Dataset 54
5.4 Plotting and Exporting Spectra 57
5.4.1 Selecting Spectra for Plotting or Exporting 57
5.4.2 Plotting Spectra 57
5.4.3 Exporting Spectra 60
5.5 Select Classification/Quantification Method 60
5.6 Evaluate Selected Method 61
5.6.1 Evaluation Procedure 61
5.6.2 Plot Classification Model Evaluation Results 64
5.6.3 Export Model Evaluation Results 66
5.7 Understanding and Reducing Classification Error Rates 66
5.7.1 Choosing an Evaluation Procedure 66
5.7.2 Understanding Results of Evaluation 67
5.7.3 Reducing Evaluation Error Rates 68
5.8 Model Details 69
5.9 Generate Final Model 70
5.10 Deleting a Model 72

6 Build and Apply a Quantification Model 73
6.1 Building a Quantification Model 73
6.2 Plot and Export Quantification Results 78
6.2.1 Plot Quantitative Model Evaluation Results 78
6.2.2 Export Model Evaluation Results 79
6.3 Completing the Model 79
6.4 Applying a Quantification Model 80
6.5 Understanding and Reducing Quantification Error Values 81
7 Batch Analyze 83
7.1 Performing a Batch of Analyses 83
7.1.1 Running Analyses Without Providing Correct Answers 84
7.1.2 Viewing the Results 85
7.1.3 Running Analyses With Correct Answers Provided 87
7.1.4 Viewing the Results 88
8 Model Manager 89
8.1 Backing up and Deleting Models 89
8.1.1 Overview of the Backup/ Delete Window 89
8.1.2 Selecting Models to Backup or Delete 91
8.1.3 Adding a Note 91
8.1.4 Backing up selected models 92
8.1.5 Deleting selected models 93
8.2 Restoring Models 94
8.2.1 Overview of the Restore Window 94
8.2.2 Restoring the models 95
9 Classification & Quantification Methods 96
9.1 Classification Methods 96
9.1.1 Support Vector Machine (SVM) 96
9.1.2 Principal Component Analysis (PCA) Classifiers 97
9.1.3 k-Nearest Neighbors 99
9.1.4 Spectral Attribute Voting (SAV) 99
9.1.5 Linear Regression for Classification 99
9.2 Quantification Methods 100
9.2.1 Linear Regression 100
9.2.2 Support Vector Machine (SVM) Regression 100
9.2.3 k-Nearest Neighbours Regression 100
9.2.4 Principal Component Regression (PCR) 101
9.3 Choosing Algorithms and Selecting Parameters 101
9.3.1 Linear Regression 101
9.3.2 Support Vector Machine (SVM) 102
9.3.3 k-Nearest Neighbors (k-NN) 103
9.3.4 Spectral Attribute Voting (SAV) 104
9.3.5 Principal Component Analysis (PCA) 104
10 Technical Support, Sales and Services 105
11 References 106
12 Patent, Copyright & Trademark Notices 106

1 Introduction

Analyze IQ Lab is a software package that uses state-of-the-art model-based techniques, drawn from chemometrics and data mining, for the rapid analysis of spectral data. The types of analysis that can be undertaken fall into two general categories:

1. Classification (also known as Qualitative Analysis or Discrimination): use models to predict whether a sample belongs to a specific category of materials; for example, predicting whether or not a sample of an unknown liquid contains a specific solvent.
2. Quantification (also known as Regression or Numerical Estimation): use models to predict a numeric value associated with a target substance; for example, predicting the concentration of a substance in a mixture.

Analyze IQ Lab also allows users to build their own models that can be deployed for the analysis of spectra. **Analyze IQ Lab** can be used for the following two core activities:

1. Application of models to the analysis of unknown spectral data.
2. The building of classification or quantification models.

These two core activities are carried on separate pages[1] of **Analyze IQ Lab**: the *Apply Model* page and the *Build Model* page. Spectral analysis can be carried out by using models built by the user or by applying pre-built models that have been included in **Analyze IQ Lab**. As part of the **Analyze IQ Lab** installation, sample models may be installed; if selected for the install, these models are ready for use when **Analyze IQ Lab** is first started.

1.1 Overview of the Analyze IQ Product Suite

Analyze IQ Lab is part of the Analyze IQ product suite.

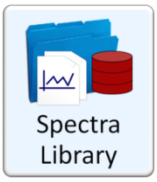
Analyze IQ Lab: Provides rapid & accurate chemometric analysis methods to turn complex spectral data into decisions. Its model building wizard allows you to build, evaluate and refine sophisticated models, quickly and reliably.

Analyze IQ RealTime: Enables you to package chemometric models that have been constructed and validated by expert analysts and deploy them in fielded applications, where they can be applied automatically for real time predictions. Available in Portable, Desktop, or Cloud Editions.

[1] A *page* is a group of windows that function together for performing a task.

Analyze IQ SmartSearch: Allows you to search rapidly and accurately through a library of spectra, to identify the best matches for unknown spectra, enabling you to answer "what is this?" questions easily and reliably, and providing you with useful supplementary data.

Analyze IQ Spectra Manager: Makes it easy for you to **store** and **organize** your valuable spectra and related meta-data. An ideal alternative to the high cost & effort of implementing a LIMS.



Analyze IQ Spectra Library: A carefully-curated collection of 1868 Raman spectra of common organic materials, including laboratory chemicals, hazardous materials, research chemicals, and solvent mixtures.

1.2 Overview of this Manual

This manual provides detailed instructions on the use of **Analyze IQ Lab** for the analysis of spectral data and for the generation of spectral analysis models, using the variety of methods provided in **Analyze IQ Lab**. It is laid out as described next.

Section 2 describes the **Analyze IQ Lab** installation process.

Section 3 provides an overview on the principles of operation of **Analyze IQ Lab**.

Section 4 details how to use **Analyze IQ Lab** for the analysis of spectral data. It gives a detailed example of the application of a model for the analysis of a spectrum stored in one of the many formats supported (including SPC, AIQ, SpectroML JCAMP-DX or simple text files). It also details how to view the detailed contents of a spectrum file and how to apply pre-processing methods to a sample spectrum.

Section 5 explains how to build your own model in **Analyze IQ Lab**. An important step in the building of new models is dataset preparation. This section tells you how to import a dataset from Excel or Unscrambler, and how to generate a dataset from an **Analyze IQ Spectra Manager** database. The examples in Sections 4 and 5 consider a classification task, in which the goal is identify whether a specific target substance is present in the sample corresponding to the spectrum under analysis.

Section 6 presents a step-by-step example of the building of a new quantification model and also demonstrates the use of a new quantification model for the analysis of sample spectrum.

Section 7 describes how to perform a batch analysis of a set of spectra with a set of models, using the Batch Analyze feature.

Section 8 explains how to backup, restore and delete models, and transfer models between computers, using the *Model Manager*.

Section 9 provides brief descriptions of all of the classification and quantification methods that are available in **Analyze IQ Lab** for the building new analysis models. It also describes the parameters for each of these methods.

Section 10 contains some information on technical support and sales, and finally **Section 11** provides a list of references.

Note that Section 2 and parts of Section 4 are also contained in the **Analyze IQ Lab Getting Started Guide.**

2 Installation

Analyze IQ Lab can be installed and run under all recent versions of the Windows desktop operating system, including Windows XP, Windows Vista, Windows 7 and Windows 8. It is 32-bit software and works fully on both 32-bit and 64-bit operating systems.

2.1 USB Key Installation

To install from USB key, insert the USB key into a USB slot and follow the instructions that appear after you open it.

If the installation screen does not appear automatically when the USB key is inserted, navigate to the top level of the drive and run the **StartHere.exe** program to start the Analyze IQ software installation.

During installation, the Setup program installs a Java Runtime Environment. If you already have the same one or a newer one installed, you can safely skip this step.

2.2 Website

If you wish to download **Analyze IQ Lab** from the Analyze IQ website, www.AnalyzeIQ.com, you must first register on it. To do so, go to the website and click on *Register* under the login form. A confirmation email is sent to the address that you entered into the registration form. This confirmation email includes a link for activating your new Analyze IQ account.

After you have registered as a user, you can download the **Analyze IQ Lab** software package at the following address: http://www.AnalyzeIQ.com/Download.html. This page allows you to download the software setup file. After downloading the setup file, start the installation process by double-clicking on it.

2.3 Getting a License

At the end of the installation process, when you first run **Analyze IQ Lab**, you will be prompted to 'Get a License', which requires an internet connection. This final step must be carried out in order to be able to run **Analyze IQ Lab**.

You can decide to defer getting a license until a later time. Please see 2.5.5 for further details.

If you are installing an evaluation copy of the software, you will enter your registered username to get a license after you install the software. Otherwise, you will enter a Commercial License Code that is given to you by Analyze IQ Ltd, or your software supplier, when you purchase the software.

2.4 Upgrading Your Existing Installation

If you have a license for an older version of **Analyze IQ Lab**, you may be entitled to install new releases. If so, you are recommended to back up data on your computer before proceeding.

Depending on your specific computer configuration and operating system, if you have built your own models with an older version of **Analyze IQ Lab**, you may find that you cannot see them in the newer version of **Analyze IQ Lab** when you install it. The models are easily transferred into the newer version; please contact Technical Support for instructions specific to your computer's configuration.

Please note that you should also take care when upgrading to a new release of **Analyze IQ Spectra Manager** that you do not overwrite spectra that you have added to the database. Refer to the ***Spectra Manager User Manual*** or contact Technical Support for further information.

2.5 Help Menu

As shown in Figure 1, you can access details about your installed version of the program, access contact details for Technical Support, Open the User Manual and Getting Started Guide in PDF format, and update your license.

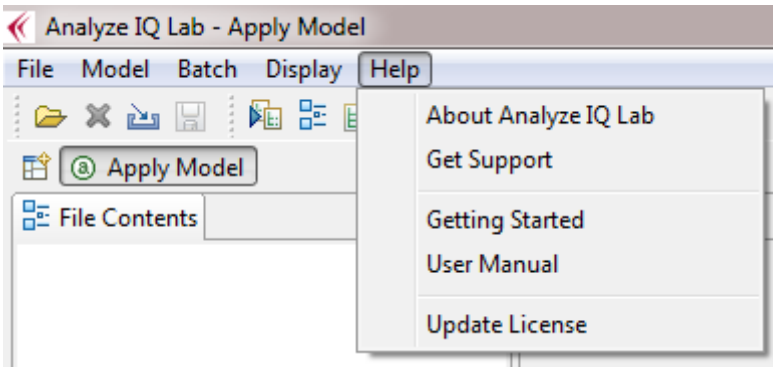


Figure 1: Help Menu

2.5.1 Help Menu - *About Analyze IQ Lab*

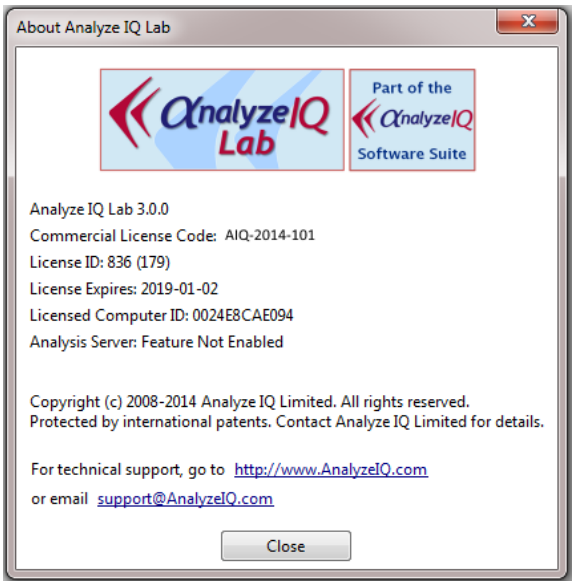


Figure 2: About Analyze IQ Lab

As shown above in Figure 2, the About **Analyze IQ Lab** menu shows you details regarding the version currently installed, including the software version number, your licence code, license ID, the license expiry date and the MAC address of the computer.

2.5.2 Help Menu - *Get Support*

In this menu, you can access contact details for Technical Support as can be seen in Figure 3.

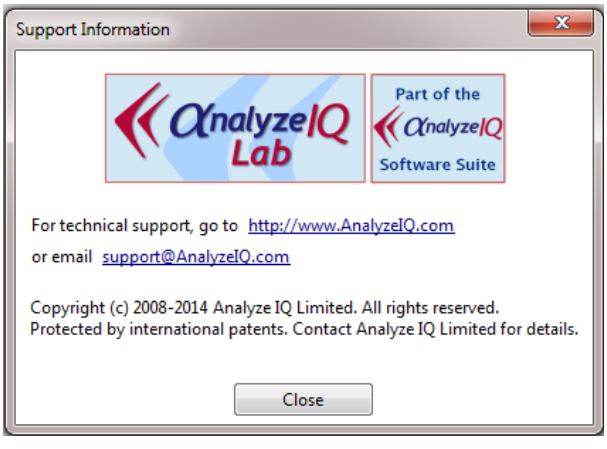


Figure 3: Get Support

2.5.3 Help Menu – *Getting Started*

From this menu, you can open the **Analyze IQ Lab** Getting Started Guide in PDF format.

2.5.4 Help Menu – *User Manual*

From this menu, you can open the **Analyze IQ Lab** User Manual (this document) in PDF format.

2.5.5 Help Menu – *Update License*

If you deferred getting a licence after first installing the software, or if you are extending an Evaluation Licence or upgrading to a Commercial Licence, you can update your licence from this menu. Enter either your user name for an Evaluation Licence or your Commercial License Code to acquire a license from our server (an internet connection is necessary). See Figure 4.

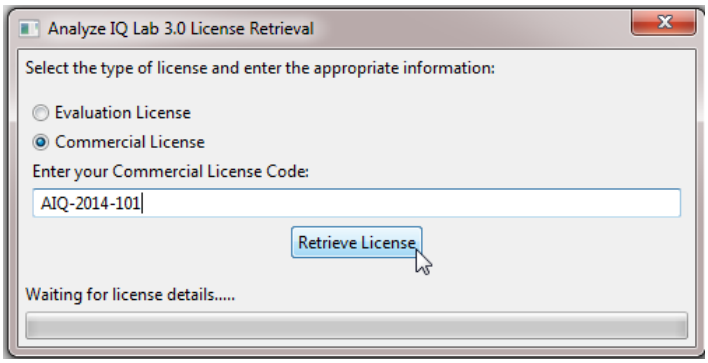


Figure 4: Update License

3 Principles of Operation

Before starting to use **Analyze IQ Lab**, it is helpful to understand its principles of operation.

Analyze IQ Lab is designed for accurate analysis of spectroscopic data. Molecular spectroscopic techniques such as infra-red (IR), near infra-red (NIR), and Raman spectroscopy are widely used in analytical chemistry to characterize the molecular structure of materials, by measuring the radiant energy absorbed or scattered in response to excitation by an external light source.

In practical applications, a given target substance to be identified is often not in pure form, but mixed with other constituents that contribute peaks to the spectrum. These peaks may mask or overlap each other, thereby causing difficulty for standard identification procedures.

Analyze IQ Lab employs a model-driven approach to spectral data analysis, with sophisticated techniques for constructing new models from data. Its architecture is shown in Figure 5.

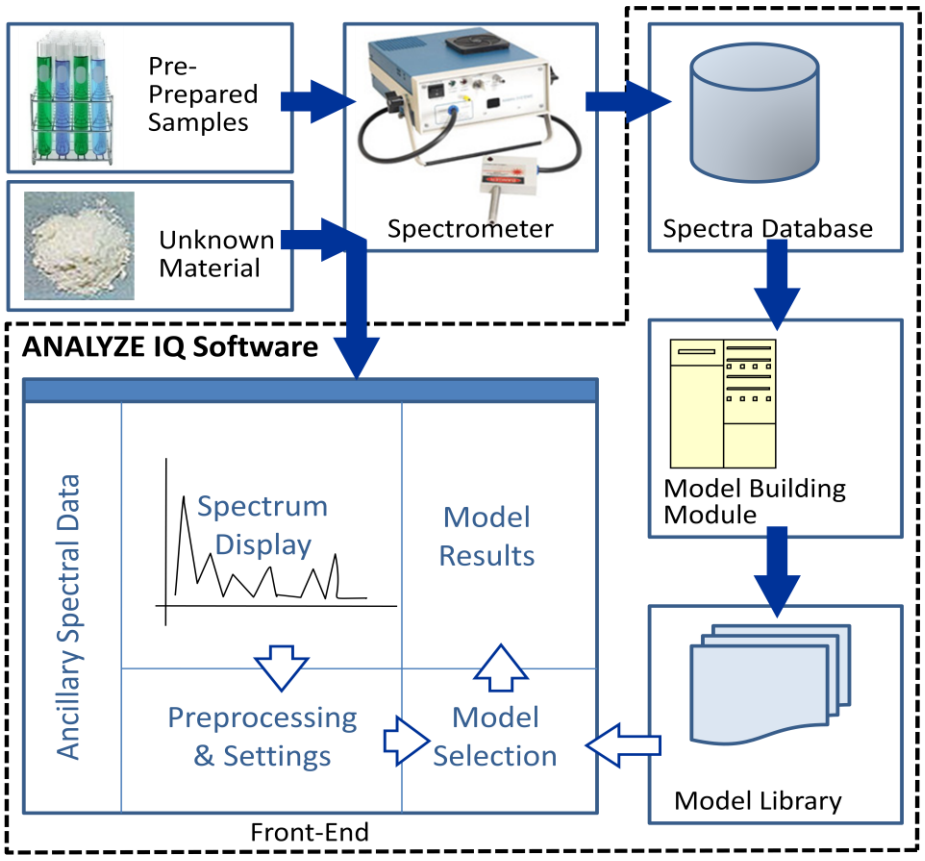


Figure 5: Analyze IQ Lab System Architecture

A key feature of **Analyze IQ Lab** it its library of models, some of which are used for qualitative analysis (e.g. identification of the presence of a target substance in a sample) and others of which are used for quantitative analysis (e.g. determination of the concentration of a substance in a sample). You can think of an Analyze IQ model as an abstraction of a large amount of data about what constitutes the essence of a target substance, so that substance can be identified or its concentration determined reliably under a wide range of conditions, including the presence of contaminants and noise. In general, **Analyze IQ Lab**'s models operate more reliably than analysis methods based on direct matching against spectral libraries, spectrum subtraction, and standard statistical techniques such as PCR and SIMCA.

When you wish to perform qualitative or quantitative analysis of a material, you open its spectrum in **Analyze IQ Lab**, choose an appropriate model from the library available to you, and apply it. The model results are then displayed. For example, you will be informed whether a target substance has been detected or what the concentration of a target substance is determined to be. **Section 4** describes in detail how to perform an analysis using an existing model.

Users are not restricted to using the models provided with **Analyze IQ Lab**. You can add to the model library by building new models from your own spectral data. As explained in **Section 5**, **Analyze IQ Lab** provides you with a step-by-step procedure to make this procedure easy to follow.

To build a new model, data is required, so a range of substances of known composition are prepared in advance; these may be stored using the ***Analyze IQ Spectra Manager*** software package for spectral data management, a spreadsheet or a text file (see Section 5.2 for details). Then, analytical methods are used to construct analytical models that compactly summarize all of the spectral data. These models are embedded in the **Analyze IQ Lab** software, for future use. When the spectrum of an unknown material is collected, its composition can be determined instantaneously using the models that have been tuned to identify specific target materials.

Analyze IQ Lab provides a wide range of analytical techniques for building new models, ranging from conventional statistical methods such as linear regression to innovative proprietary algorithms including Weighted Spectral Linear SVM and Spectral Attribute Voting. Details on these are provided in **Section 9**.

Finally, users may be interested to note that Analyze IQ Limited provides a range of services including training, model building, custom software solutions, and independent validation of analyses. Please refer to **Section 10** for contact information.

4 Perform an Analysis

This section details how to use **Analyze IQ Lab** for the analysis of spectral data. It gives a detailed example of the application of a model for the analysis of a spectrum. It also details how to view the detailed contents of a file and how to apply pre-processing methods to a sample spectrum.

4.1 Getting Started with Analyze IQ Lab

When the **Analyze IQ Lab** application is first started, you are presented with the *Apply Model* page as shown in Figure 6. (Note that this figure assumes that the sample models were selected in the installation of **Analyze IQ Lab**.) This figure highlights the five areas of this page, which are as follows:

1. ***Spectrum Details****:* shows details of the spectrum of the currently opened file, such as date recorded, instrument manufacturer etc.
2. ***View Spectrum****:* displays a plot of the spectrum of the currently opened file.
3. ***Models****:* shows the current set of models that are available for the analysis of the spectrum. Figure 6 shows three pre-built models packaged with **Analyze IQ Lab**.
4. ***View Results****:* when a model has been applied to a spectrum, the results of the analysis are shown in this area.
5. ***Pre-processing****:* shows the pre-processing techniques that can be applied to the currently opened spectrum. It also shows a history of what pre-processing steps have already been applied.

When **Analyze IQ Lab** is run for the first time, all areas of the *Apply Model* page are empty, unless pre-built models have been shipped with your **Analyze IQ Lab** package, in which case the *Available Models* area lists these models. In the example of Figure 6, three models are available:

1. PCR Acetonitrile
2. SAV Acetonitrile
3. WS Linear SVM Acetonitrile

All of these three sample models have been designed for the identification of acetonitrile; the difference between them is that each model has been generated using a different analysis technique; see the descriptions for each model.

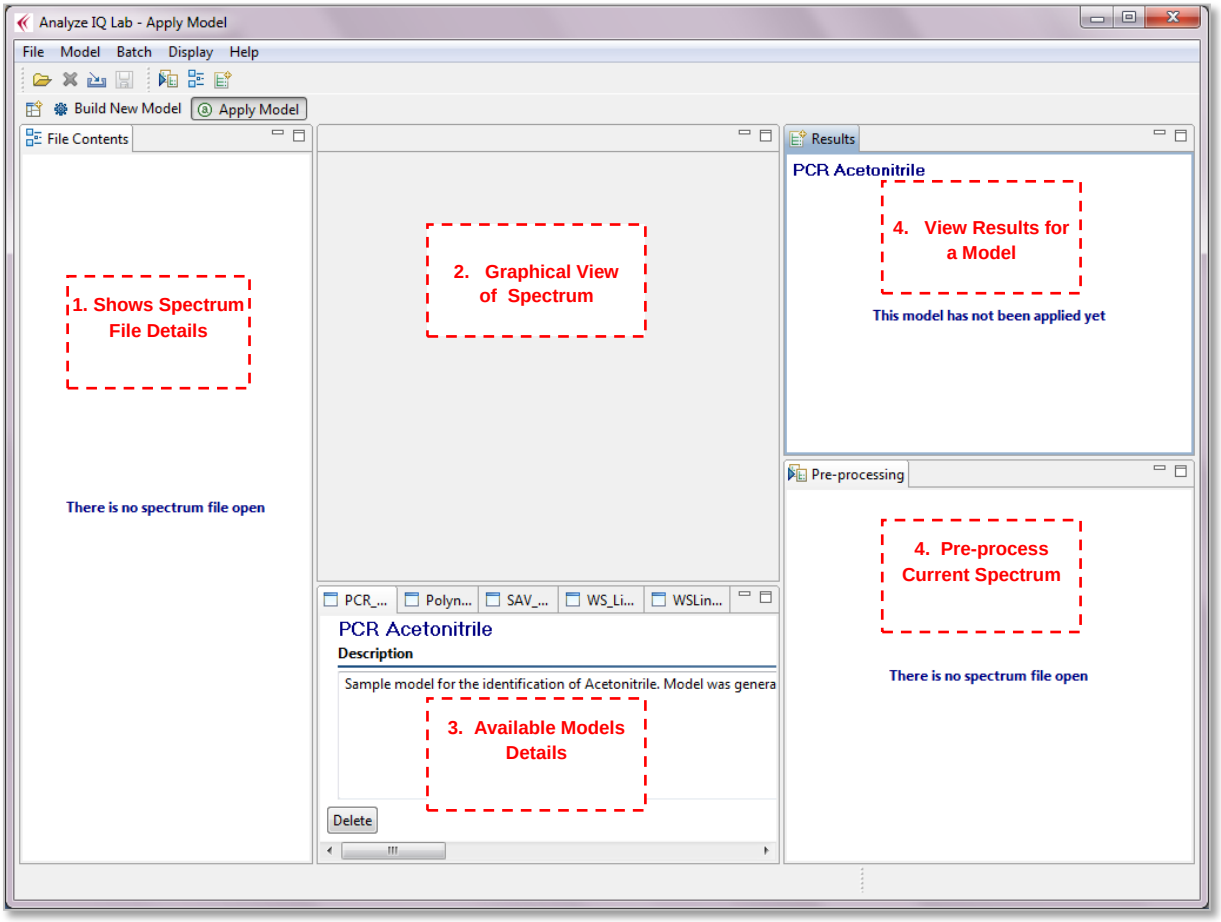


Figure 6: Starting page of Analyze IQ Lab

4.2 *Using a Model for Analysis*

To perform an analysis of a spectrum with a model, such as the *PCR Acetonitrile* model shown in Figure 6, you must carry out the following steps:

1. Open a spectrum file (see Section 4.2.1), or acquire a spectrum from an instrument (see Section 4.2.3)
2. Select and apply a model (see Section 4.2.5).
3. View model results (see Section 4.2.6).

The following sections explain these steps.

4.2.1 Open a Spectrum File

A spectrum file can be opened either via the menu or the shortcut (see Figure 7). In either case, you are presented with an *Open File* window, in which you browse to the appropriate directory, select a spectrum file and click on *Open*, as shown in Figure 8. The spectrum file opened in Figure 8 happens to be in SPC format and stores the

spectrum of a mixture made up of 10% acetonitrile, 70% methanol and 20% water. This spectrum file is one of the sample SPC files provided with **Analyze IQ Lab**. The sample spectrum files can be found in a sub-folder of the main **Analyze IQ Lab** installation folder.

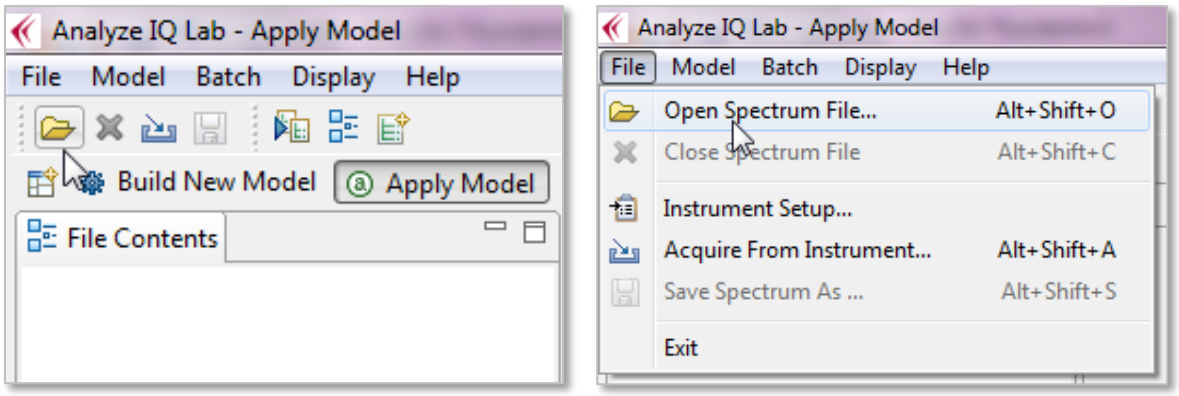


Figure 7: Open Spectrum file using: (left) the menu; (right) the shortcut

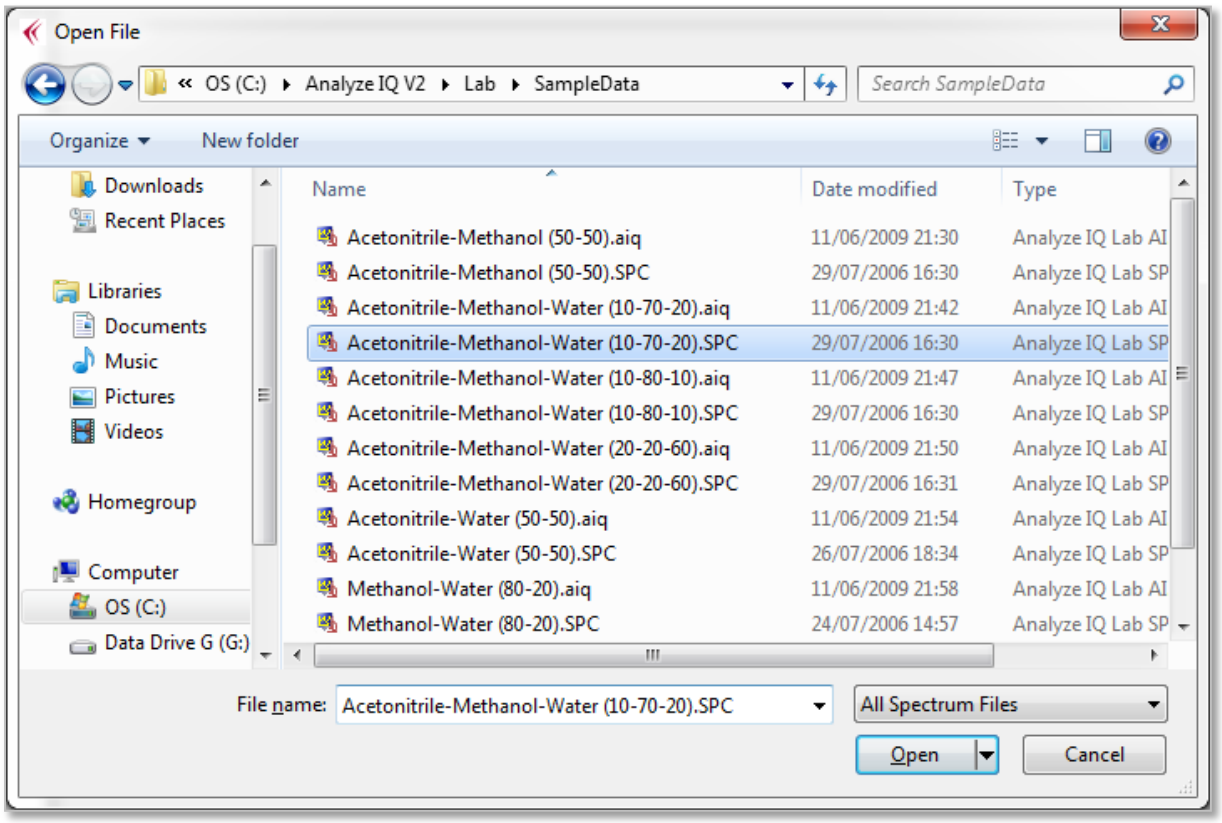


Figure 8: Open Spectrum file window

4.2.2 File Formats

Analyze IQ Lab supports several popular file formats; see Figure 9:

1. ***SPC:*** Thermo-Fisher's GRAMS SPC file format**.**

2. ***SpectroML:*** NIST's XML-based format for spectroscopy and chromatography data.
3. ***AIQ:*** The .AIQ file format is used to store data in a suitable format for Analyze IQ. It is also used in Instrument Interfaces. AIQ files use an XML format that is almost a subset of the NIST SpectroML format, **except** that it has an added property, instrumentSetting.excitationLine, that does not occur in SpectroML.
4. ***JCAMP-DX:*** IUPAC's popular JCAMP-DX file format. (Note that there are many variations of JCAMP-DX format; the Analyze IQ software supports most but may not support all. Contact Technical Support if you have problems.)
5. ***XY Text Files:*** These are simple files with either two rows or two columns of data, the first with X values and the second with Y values. See below for further details.

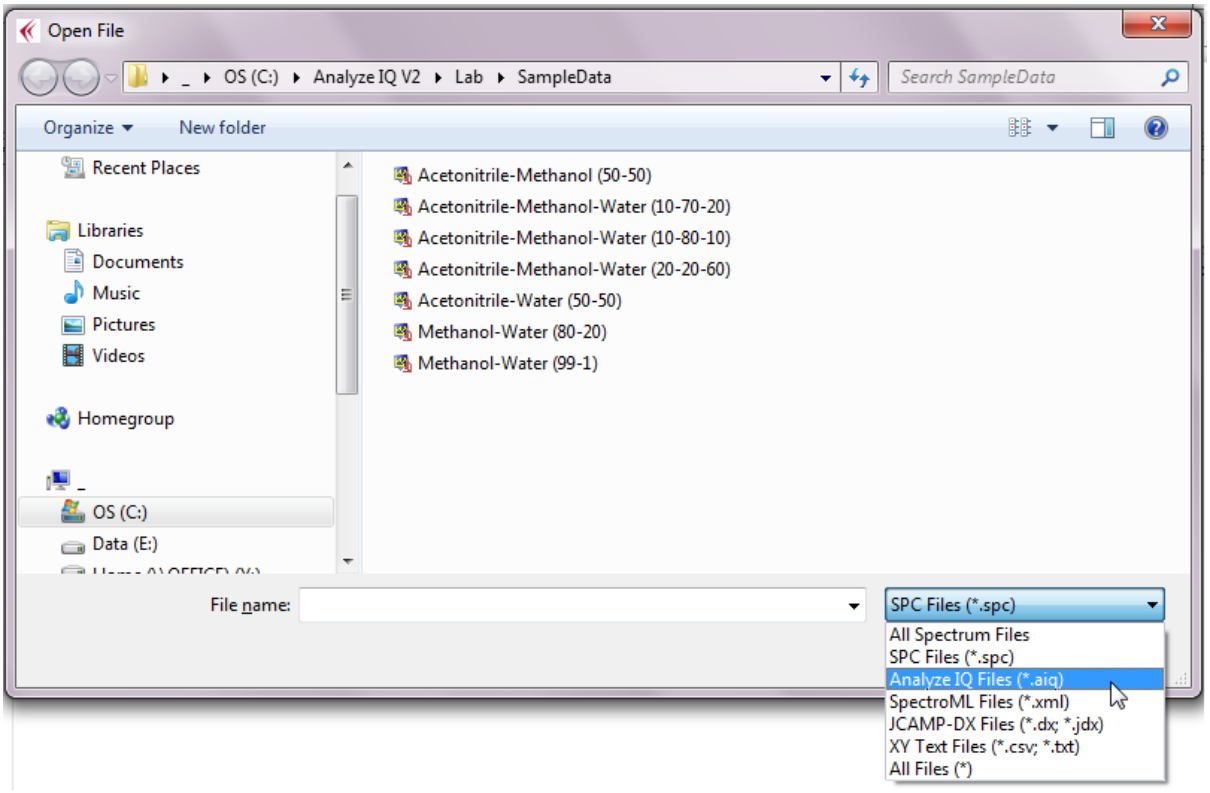


Figure 9: File Formats

As noted above, an *XY Text File* format simply consists of two rows or two columns of data, the first with X values and the second with Y values. Values may be separated by tabs or commas; the latter are known as Comma Separated Values (CSV) files. You may optionally include X and Y axis labels by putting them at the start of the two rows/columns.

The XY Text File format is particularly useful when you need to import data into Analyze IQ Lab from a software product that does not support any of our standard formats. You can easily generate files in this format with a spreadsheet program (such as Microsoft Excel or OpenOffice Calc) or a text editor.

An example of the start of an XY Text File is shown below; this one is in two rows and includes axis labels:

Wavenumber (cm-1)	250.0	252.0	254.0	256.0	...
Intensity	2454.0	2439.0	2307.0	2202.0	...

Figure 10: Example of the start of an XY Text File in row format with axis labels

When the spectrum file in any format is opened, details of the file contents are shown in the left-hand pane and a graph of the spectrum is shown in the middle of the *Apply Model* page (see Figure 11 below).

Note also that an *Apply* button is added to the *PCR Acetonitrile* pane; **Analyze IQ Lab** is now ready to apply this model (and any other models that are loaded) to the opened spectrum. For details on applying a model, see the next section.

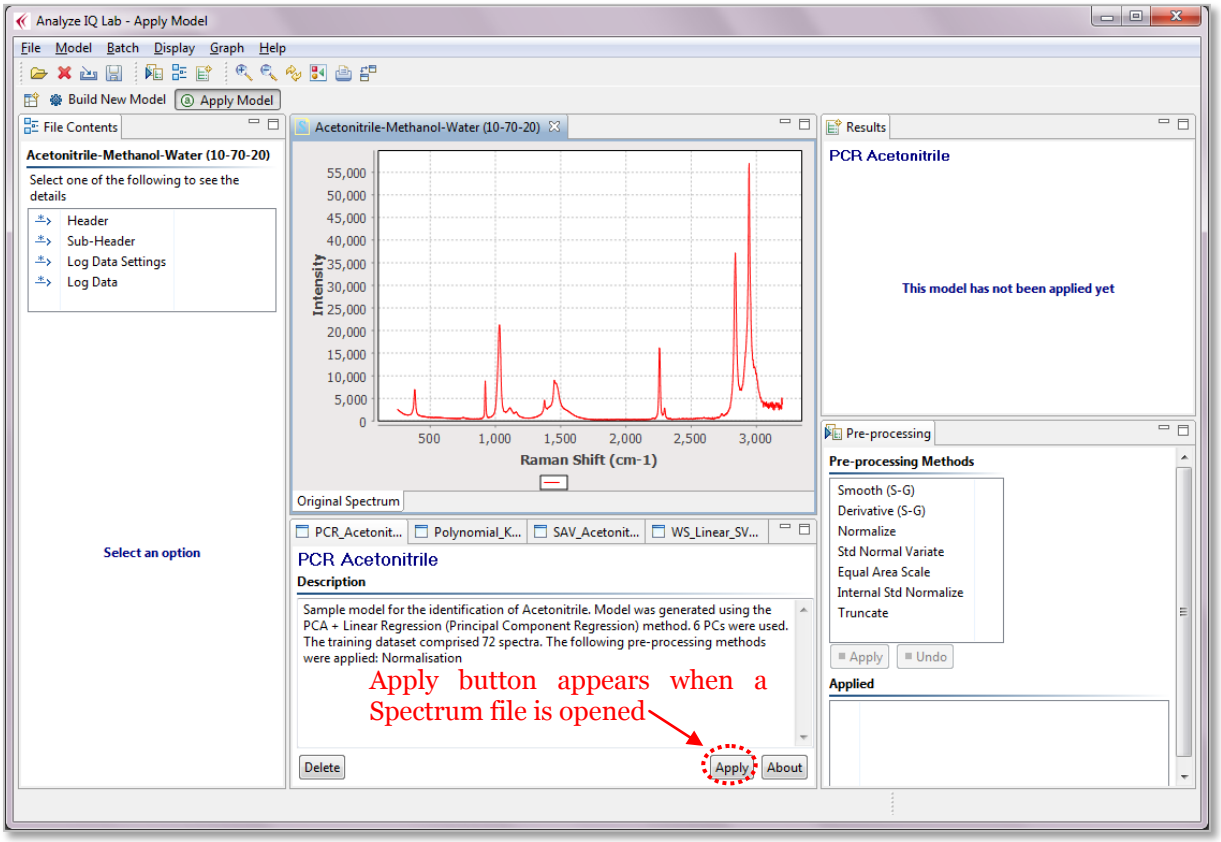


Figure 11: *Apply Model* page after Spectrum file is opened

4.2.3 Acquire from Instrument

With **Analyze IQ Lab**, you can acquire a spectrum directly from an attached spectrometer, provided that appropriate drivers have been supplied with your installation.

You can select which instrument to use by using the *Instrument Setup* menu option as shown in Figure 7 below. This opens an Instrument Setup dialog box (not shown) which shows the status of all installed instruments and allows you to select one.

After you have selected the instrument, you then can acquire a spectrum. To do this, select the *Acquire From Instrument…* option from the *File* menu, as shown in the figure below. An Instrument dialog box will appear, allowing you to specify settings appropriate to the specific instrument. (This dialog box is not shown, and its appearance varies from instrument to instrument.)

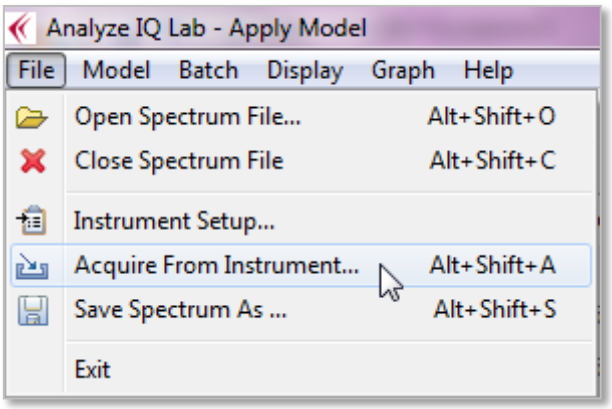


Figure 12: Menu options to select an instrument and acquire a spectrum directly from the attached instrument

Press the *Acquire* button on the Instrument dialog box to begin data acquisition. When data acquisition is finished, the Instrument dialog box will close and the data will be displayed, just as if you had opened a file.

If you are satisfied with the spectrum that you have acquired from the instrument, you will want to save it, as described next.

4.2.4 Save a Spectrum

The *File – Save Spectrum As…* menu function to save your data in either AIQ or SpectroML format. All ancillary information, as displayed on in the Contents view, is also saved.

Note that the spectrum data is saved in its 'raw' form, **without** any pre-processing you may have performed, such as normalisation or computing a derivative (see Section 4.3.1 for a description of pre-processing operations). This is to avoid problems subsequently arising with spectra being treated as 'raw' when they may have been pre-processed in a way that greatly changes their appearance, such as with a first derivative transform.

Note also that **Analyze IQ Lab** does not currently support saving SPC files; they are "read-only".

4.2.5 Apply Model

To analyze the current spectrum with a model, click on the *Apply* button of the model that you wish to apply (see Figure 13). In this example, clicking on the *Apply* button of the *PCR Acetonitrile* model will determine if the material for which the current spectrum was recorded contains acetonitrile or not (according to this particular model).

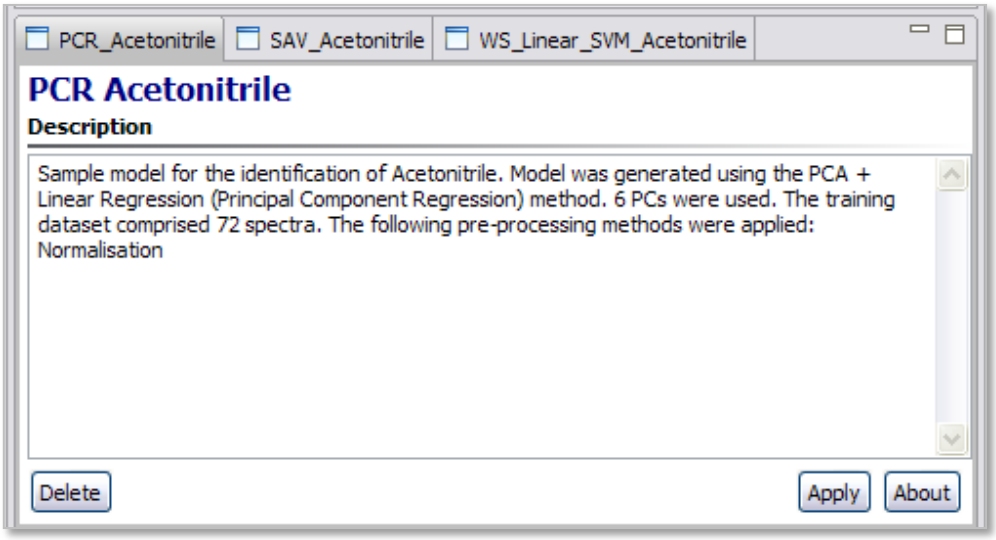


Figure 13: Applying the *PCR Acetonitrile* model

The next section describes how the results of applying a model are displayed.

4.2.6 View Model Results

When a model has been applied to a spectrum, the results are displayed in the top right-hand area of the *Apply Model* page. Figure 14 shows the results of applying the *PCR Acetonitrile* model on the spectrum opened in Figure 11. There are two parts to the displaying of results.

Firstly, the *Results* pane in the top right-hand area of the page lists the substances that the model has identified to be in the material associated with the test spectrum. In this example, the *PCR Acetonitrile* model has predicted the presence of acetonitrile in the material. This *Results* pane also lists details of any pre-processing steps that are applied by the model.

Secondly, a new spectrum is shown in the spectrum viewing area of the page; this illustrates the test spectrum after it has been pre-processed according to the requirements of the model (which can be determined by clicking on the *About* button for that model). Note that the pre-processing steps required for the model are automatically applied to the original test spectrum as it appears when the spectrum file is opened, i.e. any pre-processing steps that may have been applied in the bottom right *Pre-processing* pane do not have any bearing on the application of a model. In

this example, the *PCR Acetonitrile* model normalizes the sample spectrum, as can be seen in the *Results for PCR_Acetonitrile* graph in Figure 14 (i.e. all intensity values lie in the range 0 to 1).

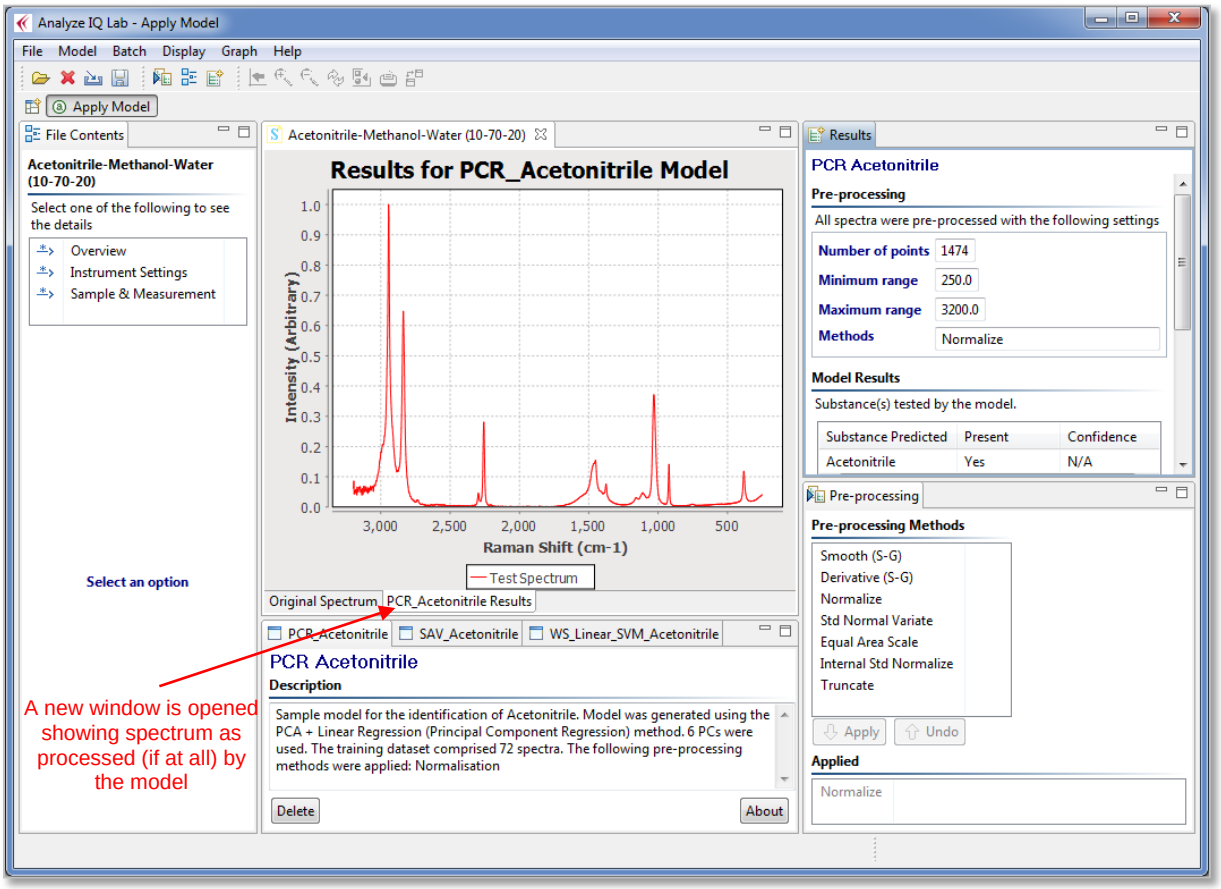


Figure 14: The results of applying the *PCR Acetonitrile* model

The *Results* pane of the *Apply Model* page (see Figure 15) also lists the target substances that the current model was designed to test for. One target substance, acetonitrile, is listed for the *PCR Acetonitrile* model.

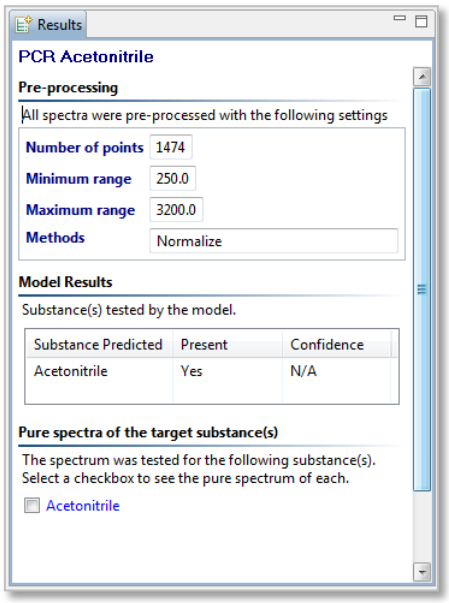


Figure 15: A closer view of the Results pane

Checking the box next to a listed target substance results in the spectrum of the target substance in its pure form being plotted in the same window containing the test spectrum (this pure target spectrum is also automatically pre-processed according to the requirements of the model). Figure 16 illustrates this feature, which is useful for visually comparing a test spectrum with the spectrum of a target substance of a model. A blue line is used for the plot of the pure target spectrum.

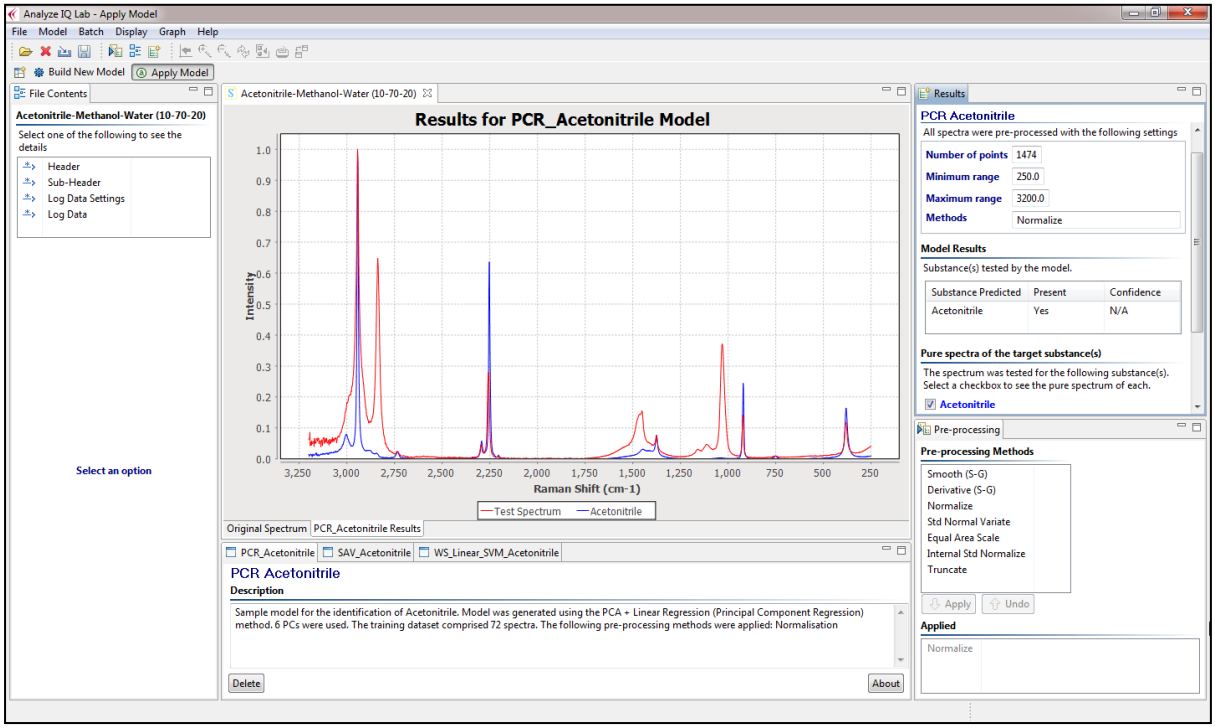


Figure 16: Displaying the spectrum of the pure target substance in the same graph that contains the spectrum of the test sample

The example of applying a model described above demonstrated a scenario in which the model returned a positive result, i.e. predicted the presence of the target substance. Figure 17 shows the result of applying the same *PCR Acetonitrile* model to another test spectrum, which is the spectrum of a mixture of methanol and water, i.e. not containing acetonitrile. The *Results* pane in this example states that the model's target substance (acetonitrile) has not been detected.

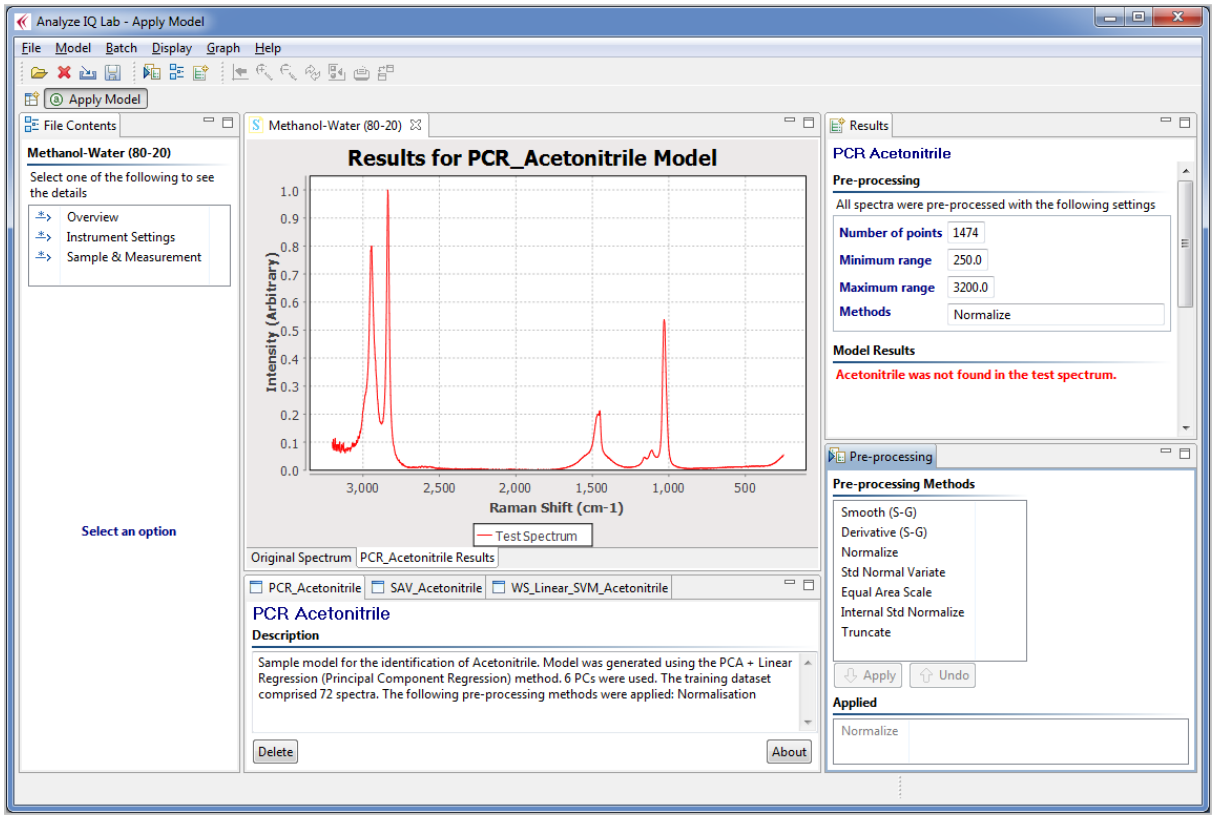


Figure 17: The PCR Acetonitrile model does not find its target substance in the test sample; this is correct since the test sample is a mixture of methanol and water

4.3 Pre-Processing Opened Spectra

4.3.1 Pre-Processing Methods Available

The pre-processing methods listed in the bottom right-hand corner of the *Apply Model* page (see Figure 18) can be applied to the currently opened spectrum. Different methods have different settings associated with them; when you select a method, the associated settings are displayed beside it. For example, Figure 18 shows the settings associated with the *Smooth (S-G)* pre-processing method.

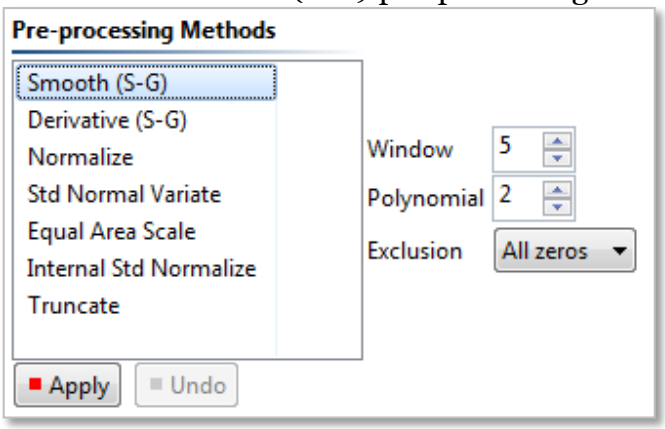


Figure 18: Pre-processing methods in Analyze IQ Lab

The details of the pre-processing methods are as follows:

1. ***Smooth (Savitzky-Golay):*** This is an method for smoothing curves, to remove small, high-frequency fluctuations. It involves performing a local polynomial regression (of degree k) over a sliding window of at least k+1 equally spaced points, to compute a moving average. See *Savitzky & Golay* [1] for more details on this pre-processing technique.
2. ***Derivative (Savitzky-Golay):*** This is an method to compute a derivative of a curve. It is strongly related to the Savitzky-Golay Smoothing method; it involves performing a local polynomial regression (of degree k) over a sliding window of at least k+1 equally spaced points to determine smoothed value for each point, and then fitting computing the derivative based on the local curvature of the polynomial. Note that selecting a zero-order derivative is equivalent to smoothing. See *Savitzky & Golay* [1] for more details on this pre-processing technique.
3. ***Normalization:*** This performs 0/1 Normalization of spectra. Each sample is re-scaled, so that the each one's lowest point corresponds to 0 and its highest point corresponds to 1.
4. ***Standard Normal Variate:*** The SNV transform re-scales each spectrum, so that the points on the spectrum end up having an average of 0 and a standard deviation of 1. This is known to produce results that are very similar to **Multiplicative Scatter Correction** (MSC), and in practice results in spectra with similar properties for chemometric modeling [4].

5. ***Equal Area Scale:*** This transformation rescales all spectra so that they all have equal area under their curves. You can specify a **scale factor**, which defaults to 1000. The larger the scale factor, the larger the overall height of the scaled spectra.
6. ***Internal Standard Normalize:*** This is like the standard Normalization, except that instead of the spectrum being scaled such that its overall highest point is rescaled to 1, with Internal Standard Normalization, a point at the specified X value is rescaled to 1. This is typically used where a specific peak in the spectrum corresponds to a component that is known and constant (the internal standard), and all other spectra are to be normalized relative to it. In using this, you must specify the **X Location** of the internal standard. To allow for minor variations in measurements between spectra, you can also specify a **Plus/Minus** value (which defaults to 0), and the spectrum is then normalized relative to the highest peak found in the range of the X ± this value.
7. ***Truncate:*** There are two forms of truncation supported, **Crop** and **Notch**:
    a. ***Crop:*** The start and end of the spectrum are cropped; the sections of the spectrum before the specified starting position and after the specified ending position on the X axis are replaced with the replacement value (which defaults to 0)
    b. ***Notch:*** A notch is removed within the spectrum; a section of the spectrum going from the specified starting position to the specified ending position on the X axis is replaced with the replacement value (which defaults to 0).

4.3.2 Settings for Smooth and Derivative Pre-Processing

The following is a description of the parameter settings used in the *Smooth (S-G)* and *Derivative (S-G)* methods:

1. ***Window:*** the length, specified in number of spectral data points, of the window over which smoothing is applied. (Default is 5.)
2. ***Polynomial Order:*** the order of the polynomial fitted to the spectrum for smoothing. (Default is 2, i.e. second order.)
3. ***Derivative:*** The order of the derivative: applies to *Derivative* only. (Default is 1, i.e. first derivative.)
4. ***Exclusion:*** determines whether wrapping is used (*Wrap* option) in the smoothing calculation for the beginning and the end of the spectrum or not (*All zeros* option). For example, if *Wrap* is selected, a window of length 5 centered on the last data point includes the previous two spectral data points and the first two data points of the spectrum. Similarly, a window of length 5 centered on the first data point of the spectrum includes the second and third data points and the last two data points of the spectrum. If *All Zeros* is selected with a window length of 5, the first two data points and the last two data points of the spectrum are set to zero as a result of the pre-processing.

4.3.3 Example of Applying Pre-Processing

For example, to apply a First Derivative, to the current spectrum (see Figure 19):

1. Select *Derivative (S-G)* in the list under the *Pre-processing Methods* title.
2. Set the appropriate parameters. In this example, first-derivative (*Derivative* = 1, the default) using a second-order polynomial (*Polynomial* = 2) with a window size of 5 (*Window* = 5, the default) is being applied.
3. Click on the *Apply* button.

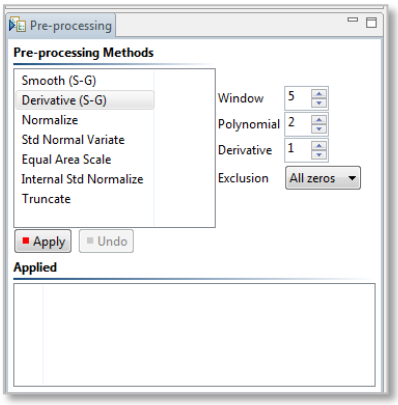


Figure 19: Applying a Derivative (Savitzky-Golay) to current spectrum

After applying the First Derivative, the spectrum displayed in the middle of the Apply Model page is modified accordingly (see Figure 20).

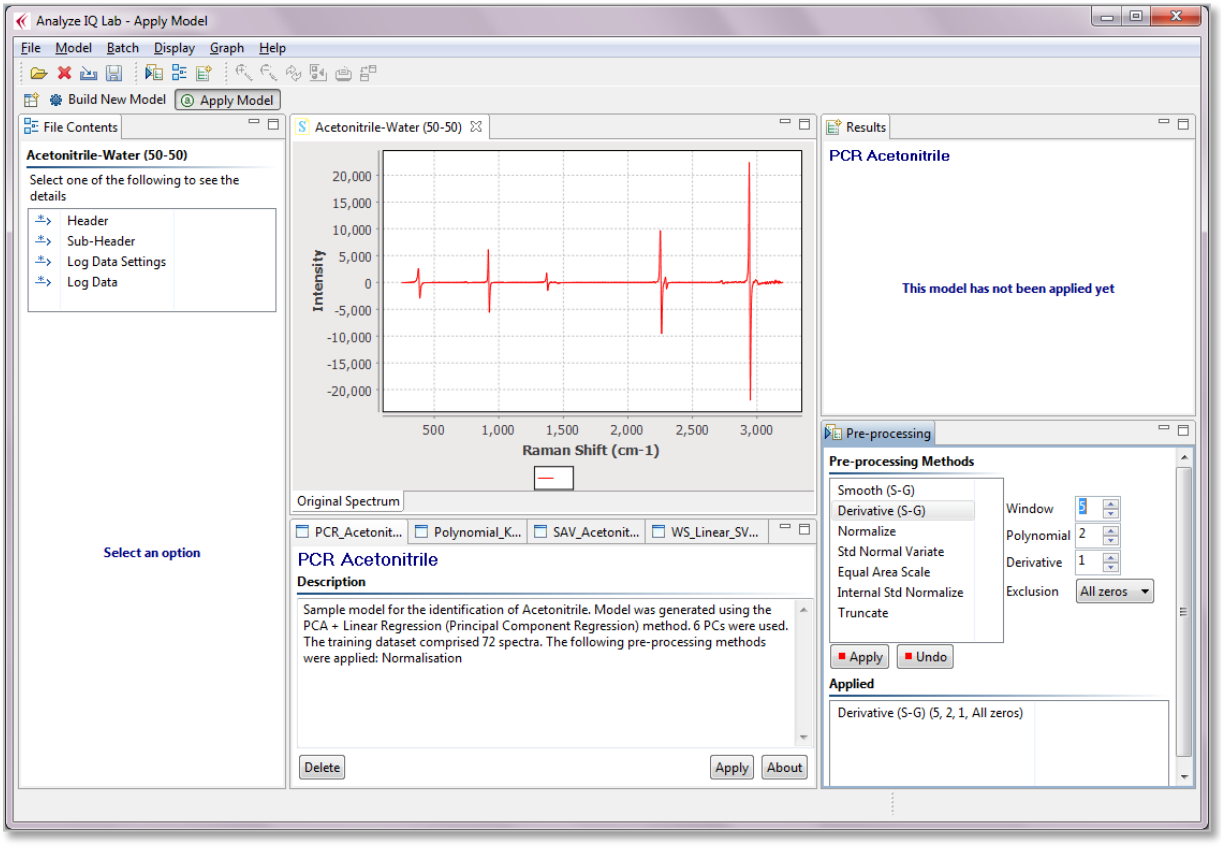


Figure 20: Result of applying first-order derivative to a spectrum

Note that a history of all previously applied pre-processing steps is recorded under the *Applied* section in the *Pre-processing* pane. Each of these previously applied pre-processing steps can be undone by successive clicks of the *Undo* button.

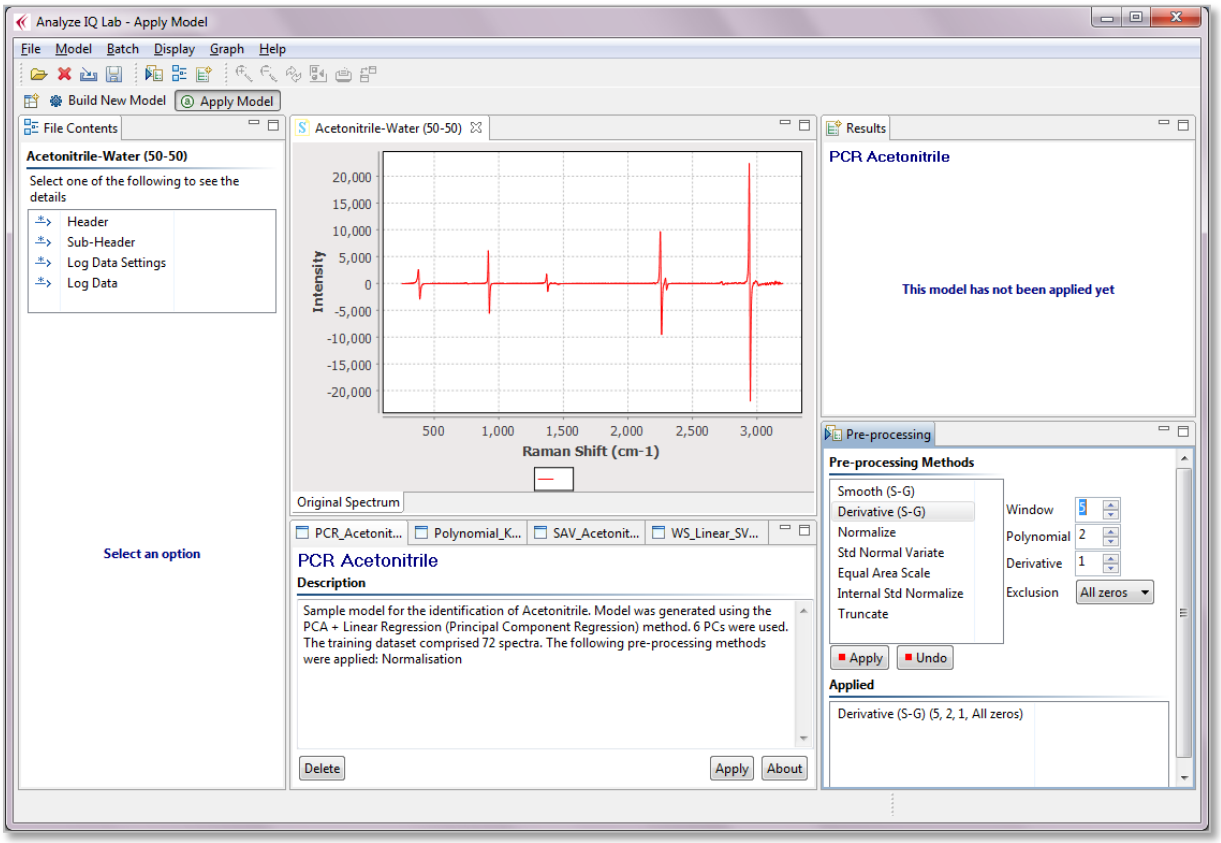


Figure 20: Result of applying the Savitzky-Golay derivative pre-processing method; note that the graph in the centre of the screen has been redrawn

4.4 Viewing Spectrum Details: SPC Format

When a spectrum file is opened (see Section 4.2.1 on how to open a spectrum file), you can view details of the file, by selecting one of the categories listed in the *Spectrum File Contents* pane on the left-hand side of the *Apply Model* page. The categories that are available depend on what kind of file you have opened.

Figure 21 demonstrates the details that are available when an SPC is open. In the figure, the *Log Data* section of the currently opened spectrum file is displayed. *Log Data* contains details such as the date the spectrum was recorded on, the detector temperature, whether background was acquired, etc. Clicking on *Header*, *Sub-Header* and *Log Data Settings* displays other details of the file.

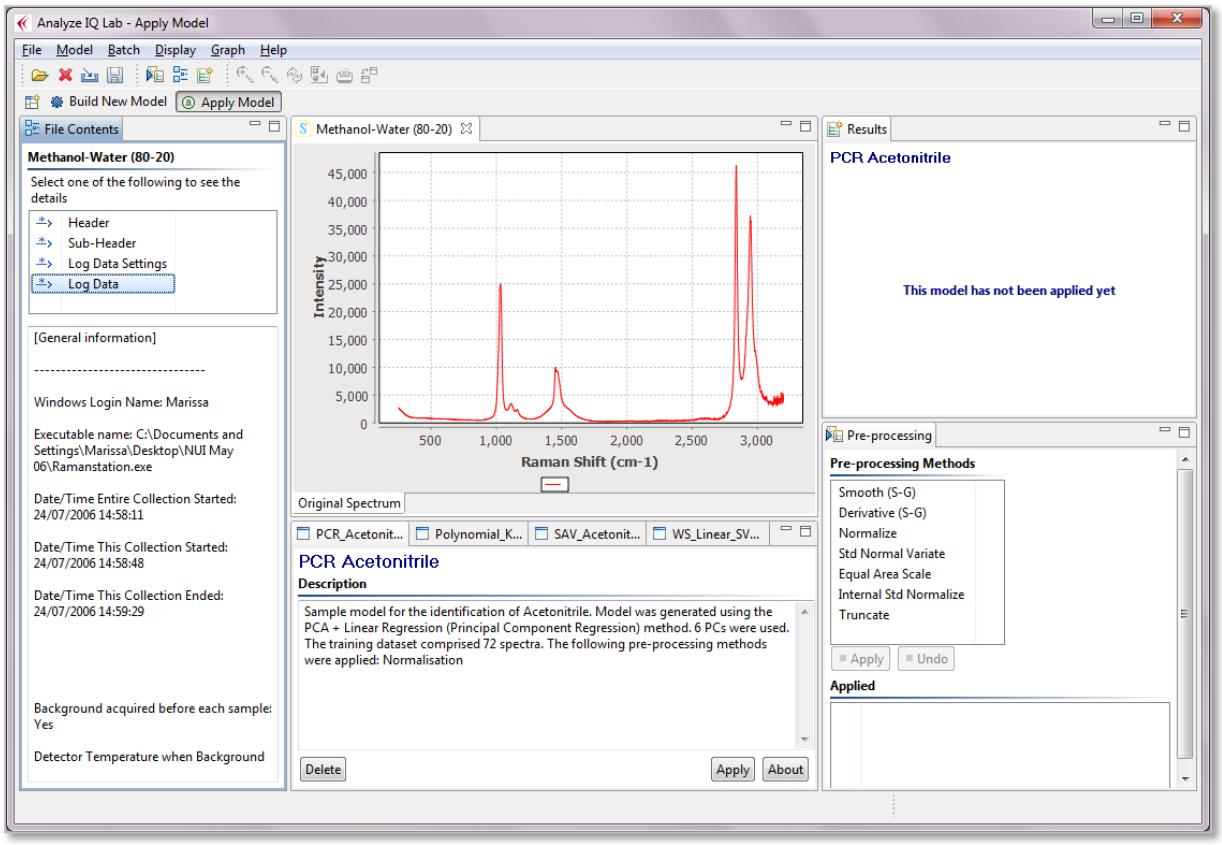


Figure 21: View SPC file contents: Log Data section (left side of screen)

4.5 *Viewing Spectrum Details: AIQ or SpectroML Format*

The data contained in AIQ or SpectroML files is in general more accessible to interpretation than that in SPC files.

Figure 22 shows an example of what is displayed when an AIQ or SpectroML formatted file is opened. There are three headings, *Overview, Instrument Settings,* and *Sample & Measurement*. Click on a heading to display the relevant details. If a specific detail has not been stored, the text "(none)" is displayed.

In the figure, the *Instrument Settings* heading has been selected, and details such as the experiment title, instrument manufacturer, instrument model and resolution are visible.

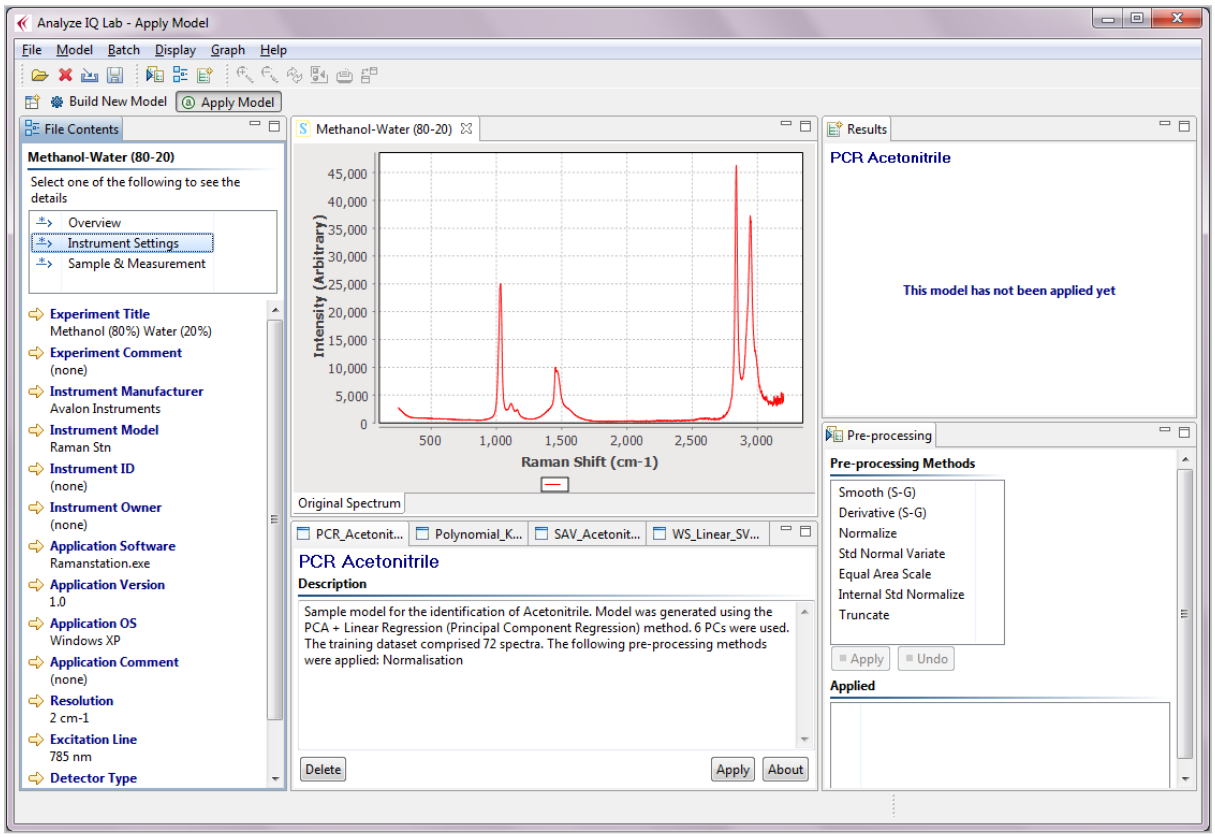


Figure 22: View AIQ file contents: Instrument Settings section (left side of screen)

Note that files in other formats, such as XY Text Files or JCAMP-DX, are converted internally to AIQ format. They are therefore displayed in the same way as the above, although many of the details may be unspecified.

5 Build a New Model (Classification Example)

After applying various models to spectrum files of different test spectra, the next step is to build your own models using the chemometric and innovative machine learning techniques provided in **Analyze IQ Lab**. This section details how to build a new qualitative or quantitative model in **Analyze IQ Lab**. For the purposes of illustration, this section includes an example of building a classification model. However, you would follow the same procedure in building a quantification model. Refer to Section 6 for an example of building a quantification model.

The main steps in the building of a new model are as follows:

1. If the *Apply Model* page is currently open, switch to the *Build New Model* page (see Section 5.1).
2. Prepare a training dataset (see Section 5.2).
3. Apply pre-processing to the training dataset, if required (see Section 5.3).
4. Select a Classification/Quantification method (see Section 5.5).
5. Evaluate the selected method (see Section 5.6).
6. Enter model details (see Section 5.8)
7. Generate the final model (see Section 5.9).

The above steps are described in the following sections. As will be demonstrated, you carry out each of the main steps in building a model by clicking on the *Next* button in the top right-hand corner of the *Build New Model* page. To move back to a previous step in the model building process, click on the *Back* button.

Note that the examples given in the following sub-sections are based on the building of a model for the classification of acetonitrile. If you are interested in building quantification models, you are recommended to read this section and then read the example of building a quantification model given in Section 6.

5.1 Switch To Build New Model Page

You can switch to the *Build New Model* page by either selecting the shortcut on the left-hand side of the application window (see Figure 23) or by selecting *Build New Model* under the *Model - Tasks* menu (see Figure 24).

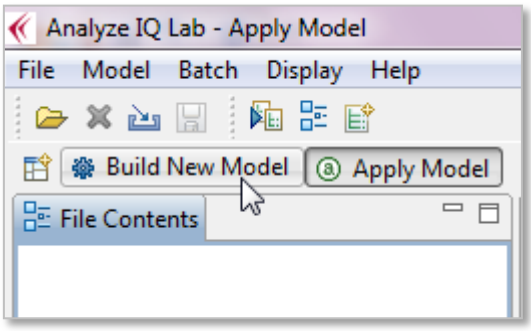


Figure 23: Switching to the *Build New Model* Page using shortcut

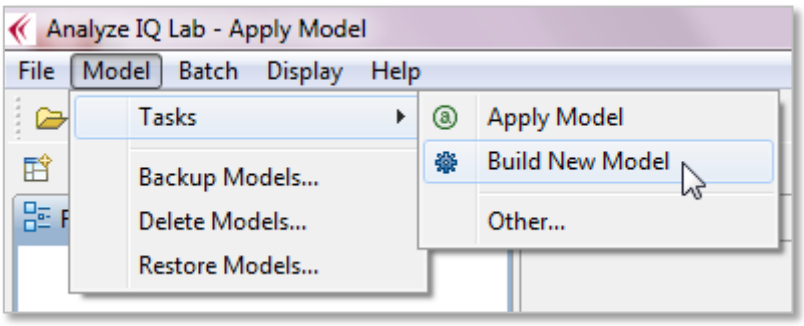


Figure 24: Switching to the Build New Model Page using the menu

After switching to the *Build New Model* page, the user is presented with the page shown in Figure 25.

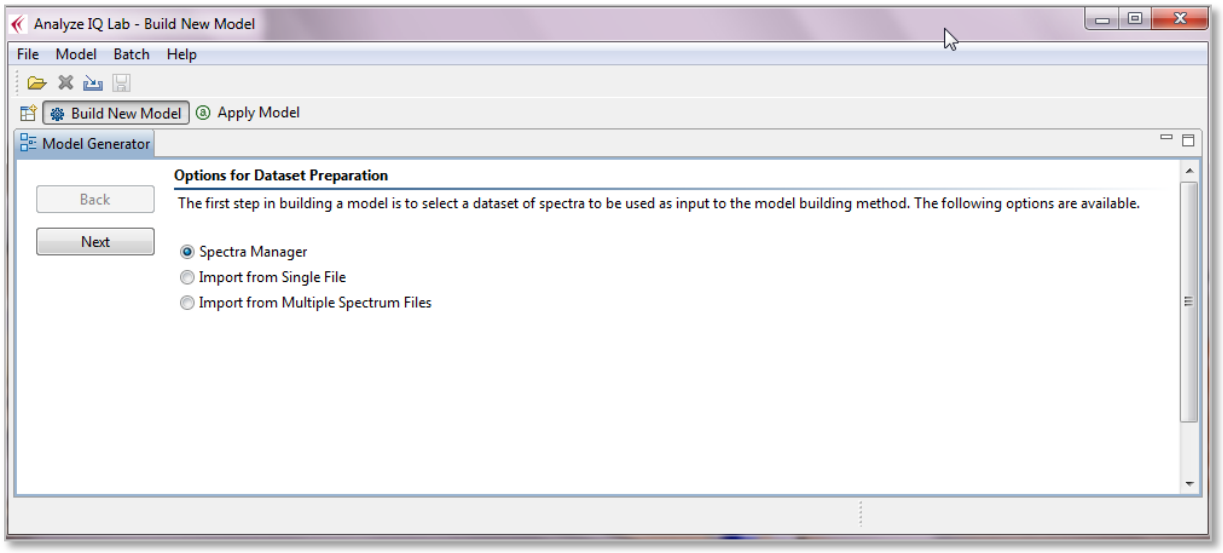


Figure 25: Build New Model Page

5.2 Preparing the Dataset

The first step to building a model is to prepare a dataset of sample spectra, on which the new model is to be based. Typically, samples of pure materials and mixtures are prepared with specific compositions under laboratory conditions and their spectra are collected prior to building a model.

As shown in Figure 25, three options are available for supplying a dataset to **Analyze IQ Lab**:

1. Spectra Manager (described in Section 5.2.1)
2. Import from Single File (described in Section 5.2.2)
3. Import from Multiple Spectrum Files (described in Section 5.2.3)

As will be described in Section 5.2.4, the 'Import from Single File' feature can be used to transfer data from other software packages, such as the Unscrambler chemometric package.

5.2.1 Spectra Manager

The *Spectra Manager* option can be only used if ***Analyze IQ Spectra Manager*** (the spectral data management software package) has been installed on your machine. The alternatives are to import data from a single text file containing data from multiple spectra, as described later in Section 5.2.2, or to import from multiple spectrum files in a folder, as described later in Section 5.2.3.

To generate a spectral dataset from data stored in Spectra Manager, select the *Spectra Manager* option and then click on the *Next* button. You are then presented with the details of the connection, as shown in Figure 26.

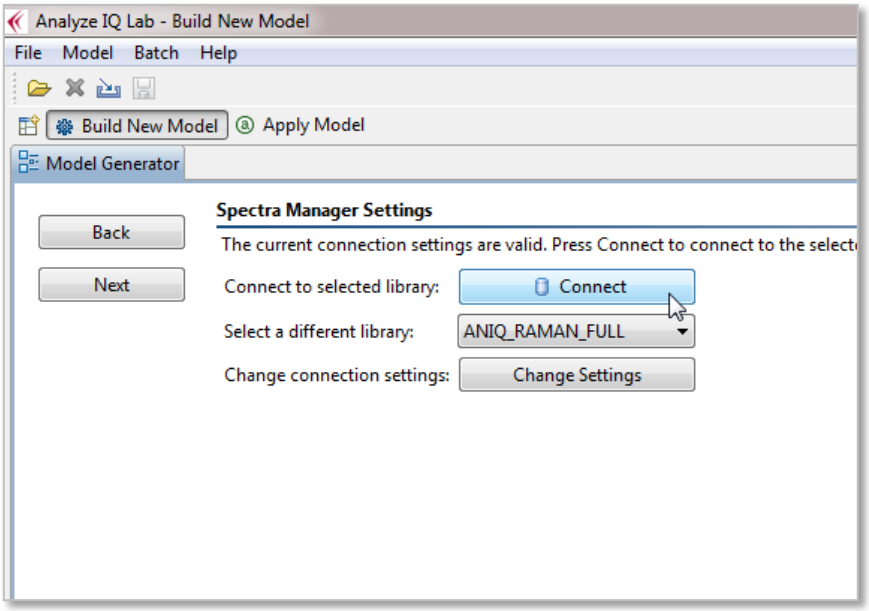


Figure 26: After selecting Spectra Manager option for spectral dataset preparation

The connection settings can be changed, but this is not advised, unless you are an advanced database user. See Section 0 for more details on changing the database settings.

5.2.1.1 Connect to Spectra Manager/View Spectral Data

The first step in generating a dataset from ***Spectra Manager*** is to connect by clicking on the *Connect* button (see **Error! Reference source not found.**), which results in the contents of its database being displayed, as shown in Figure 27.

5.2.1.2 Terminology Used for Data Stored in Spectra Manager

Spectra Manager stores a list of spectra, where each spectrum is assigned a unique ID within the database. **Spectra Manager** stores its spectra grouped according to *substances* and *constituents*:

- It stores a list of *substances*, where a substance is a compound that is defined by a CAS number and an IUPAC name.
- It stores a list of *constituents*, where a constituent refers to a particular specimen of a substance and is described by the following properties: *Substance*, *Manufacturer*, *Lot Number*, *Catalog Number*, *Purity*, *Location* and *Date Opened*. For example, a substance with IUPAC name Acetonitrile and CAS Number 000075-05-8 is stored in the **Spectra Manager** list of *Substances*. A sample of Acetonitrile manufactured by Aldrich with 99% purity and Lot Number 3725 is an example of one *constituent*. For each substance recorded in **Spectra Manager**, one or more constituents are also recorded.
- It stores a set of *spectra*, where each spectrum represents either a *pure* sample, which comprises a single constituent at 100% concentration, or a *mixture* of two or more constituents, where the sum of the constituent concentrations is equal to 100%.

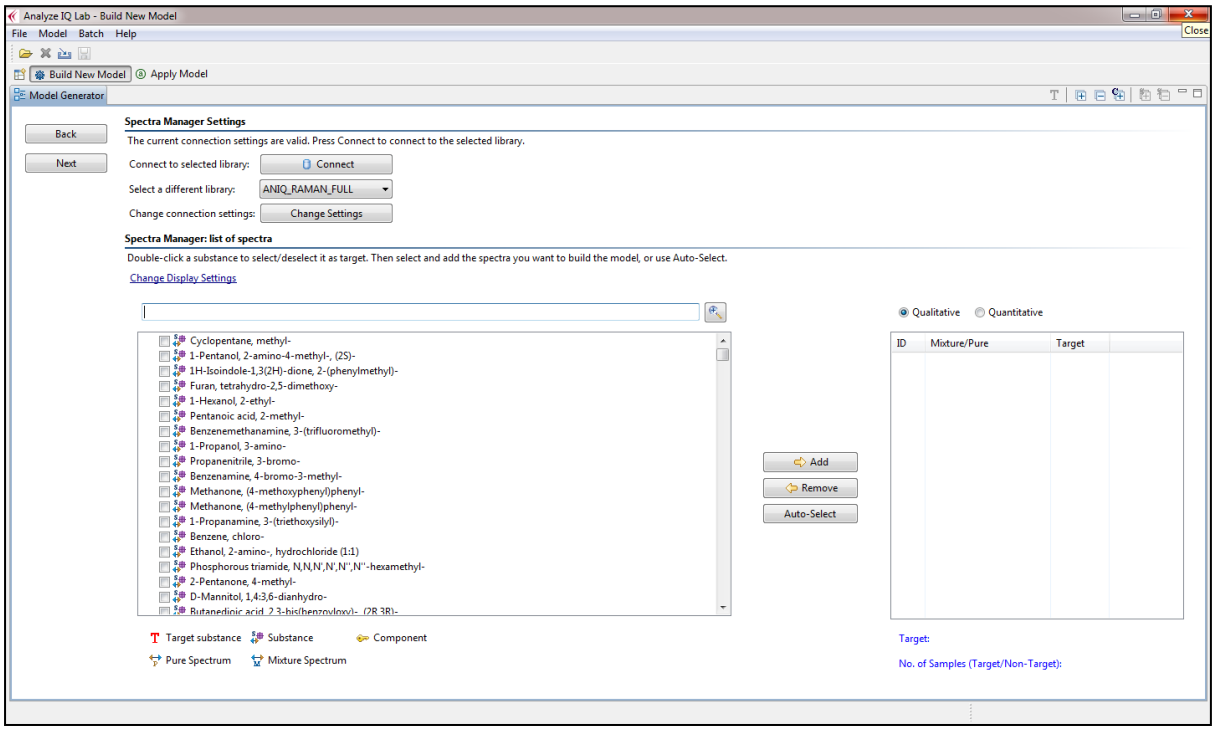


Figure 27: Displaying contents of *Analyze IQ Spectra Manager* database

As can be seen in Figure 27, the contents of a spectral database are displayed on the left-hand side of the *Build* Model page in a tree structure. At the top level of this tree, all *Substances* are shown. These group together all spectra that are defined by the same CAS number and IUPAC name. For each substance listed, the database contains one or more constituents, either as sole constituents in pure samples, or as constituent within mixtures, or both.

After first connecting to ***Spectra Manager***, a list of all substances is shown. In the example shown in Figure 27, both the name and CAS (Chemical Abstracts Service) number are listed for each substance. To change the details shown for each substance, see Section 5.2.1.7.

Under the *Substance* level, the spectra are grouped by *individual constituents,* where a constituent is distinguished by the following properties:

1. Manufacturer
2. Catalogue Number (from manufacturer)
3. Lot Number
4. Date Opened
5. Purity
6. Location

To view all the constituents contained in a substance group, expand the branch associated with that group (see Figure 28). As can be seen, ***Analyze IQ Spectra Manager*** is designed to allow users to include multiple entries for a single material, for the purposes of quality control/assurance and analysis model development.

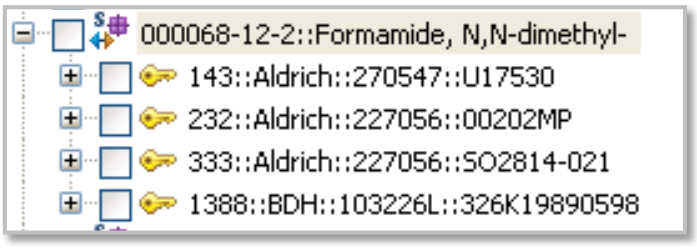


Figure 28: Displaying all members of a Substance Group in the database

In the example shown in Figure 28, all constituents (individual specimens) that correspond to the N,N-Dimethylformamide (DMF) substance in the database are being displayed. Three of these constitiuents are from the same manufacturer, but differ in either catalogue or Lot number. To view the list of spectra for an individual constituent, expand the branch associated with that substance.

Under the *Constituent* level, individual spectra are listed. In Figure 29, the four constituent lists for the DMF substance have been expanded, to show all spectra for DMF in the database. The first three constituent lists contain on spectrum each, but thee last constituent list contains several individual spectra: one corresponding to the substance in pure form (100% concentration) and the others corresponding to spectra of mixtures in which that substance is present in various concentrations; these vary from 50% to 99%, as can be seen in the figure.

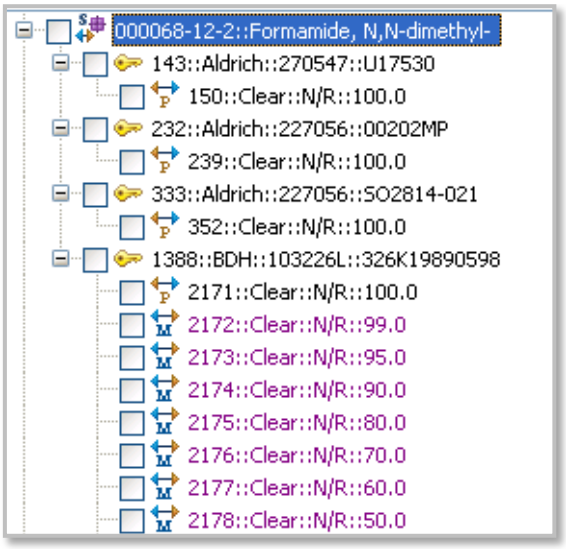


Figure 29: Listing spectra for a substance

The icon beside each spectrum listed indicates whether it is a spectrum of the substance in its pure form or as a constituent of a mixture. The legend below the list of spectra explains each icon. In addition, pure spectra are listed in black text whereas mixture spectra are listed in purple text.

In the example of Figure 29, the Spectrum ID (each spectrum is uniquely identified by this ID in the database), the color, consistency (N/R = not recorded) and concentration of the constituent in the mixture (100% when the spectrum is of a

substance in its pure form) is shown for each spectrum. For example, one of the spectra listed in Figure 29 is Spectrum 2175, which is a spectrum of a mixture that contains 80% DMF (manufactured by BDH) and is clear in color; consistency is not recorded.

For information on changing the details listed for each spectrum, see Section 5.2.1.7.

5.2.1.3 Select a Target Substance for Building a New Model

The ***Spectra Manager*** contents viewer described above allows the user to select a subset of spectra to be used as the spectral dataset for building a model. The right-hand side of the *Build Model* page shows the contents of the currently selected spectral dataset; it is empty initially (see Figure 30). Before selection of spectra, it is recommended to select the target substance. This is done by double-clicking on the substance that is to be the target for either the classification or the quantification model. Once selected, a red 'T' symbol appears beside the branch for the target substance. In Figure 30, DMF has been selected as the target substance; the name of the current target is also indicated under dataset window on the right-hand side.

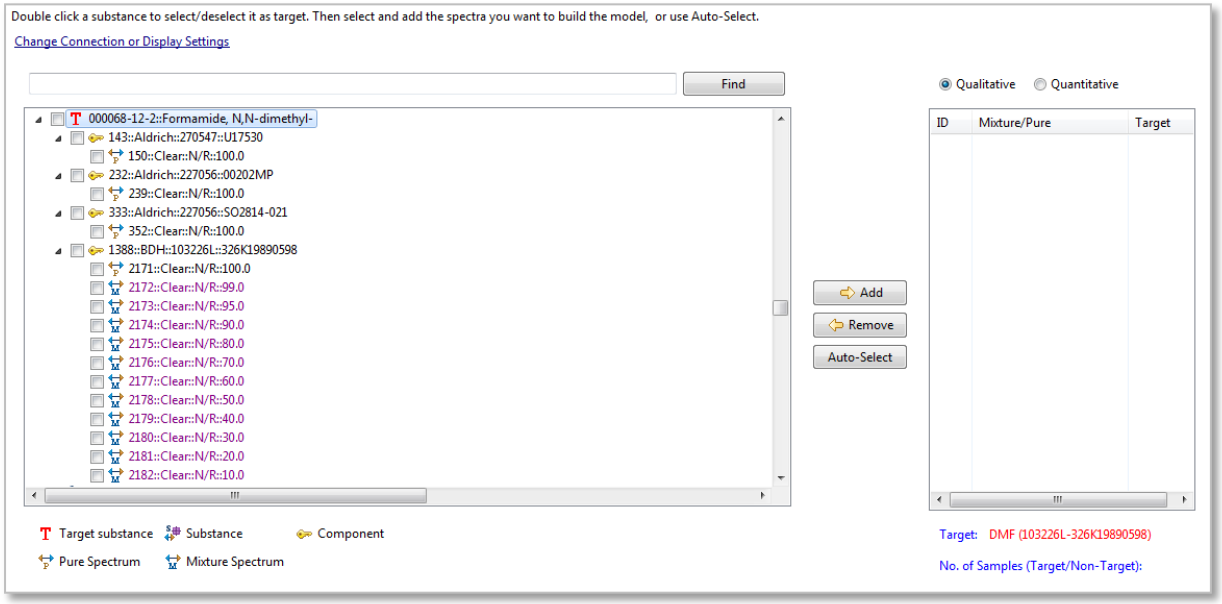


Figure 30: Selection of the Target substance

Note that only one target substance may be set at any one time; you may change the current target substance by double-clicking on a different substance branch. If no target substance is selected, an error message will appear when you try to move to the next step in model building (i.e. when the *Next* button is clicked).

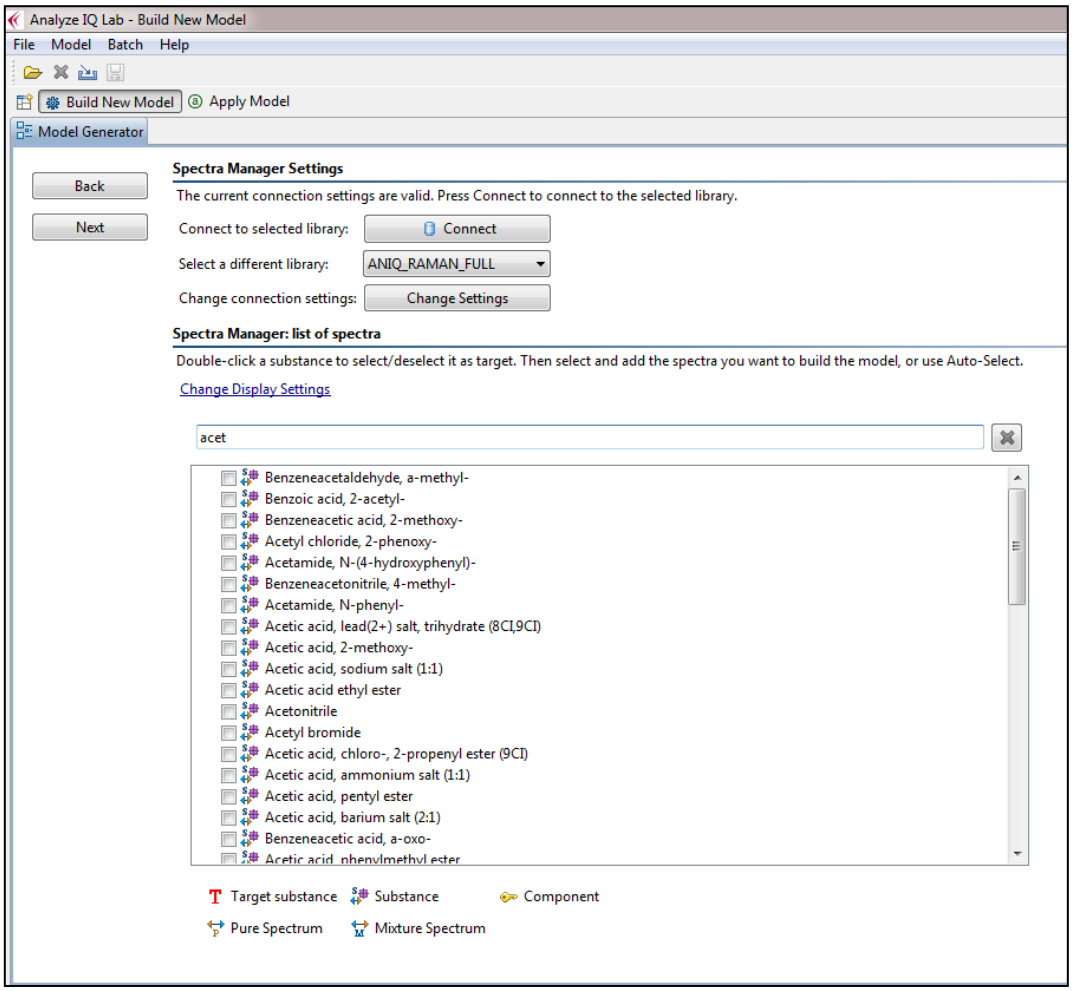


Figure 31: Searching for spectra by substance name

The *Find* button above the list of spectral contents can be used to search for spectra by substance name. Once you start typing in the search box as shown in Figure 31, the list of Spectrum will adjust according as you increase the number of letters in your search as shown in Figure 32.

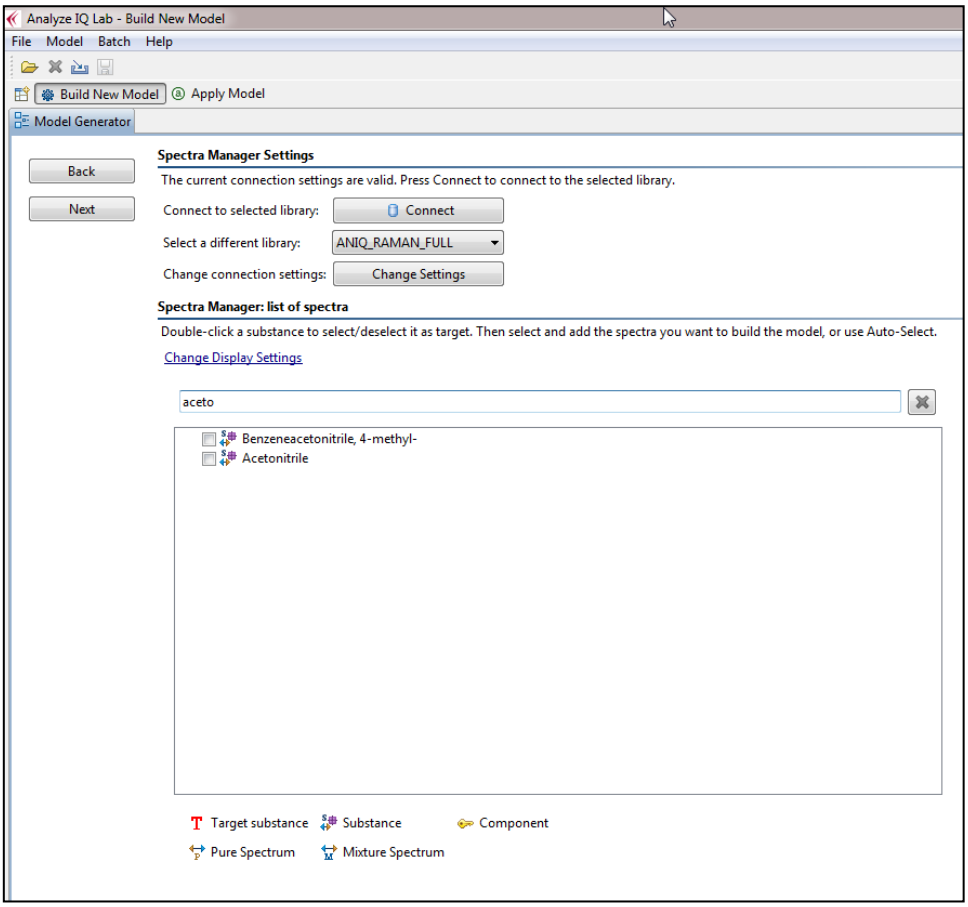


Figure 32 Searching for spectra by substance name

5.2.1.4 Select a Training Dataset of Spectra

To add an individual spectrum to the dataset, click on the box at the start of the branch for that spectrum and then click the *Add* button (see Figure 33). After each addition or removal of spectra from the current dataset, the details under the dataset window are updated to show the current number of samples in the dataset and the breakdown between those containing the target substance and those not containing the target substance (provided the target has been selected). Figure 33 shows that the current dataset contains a single spectrum, which is of a sample that contains the target, DMF.

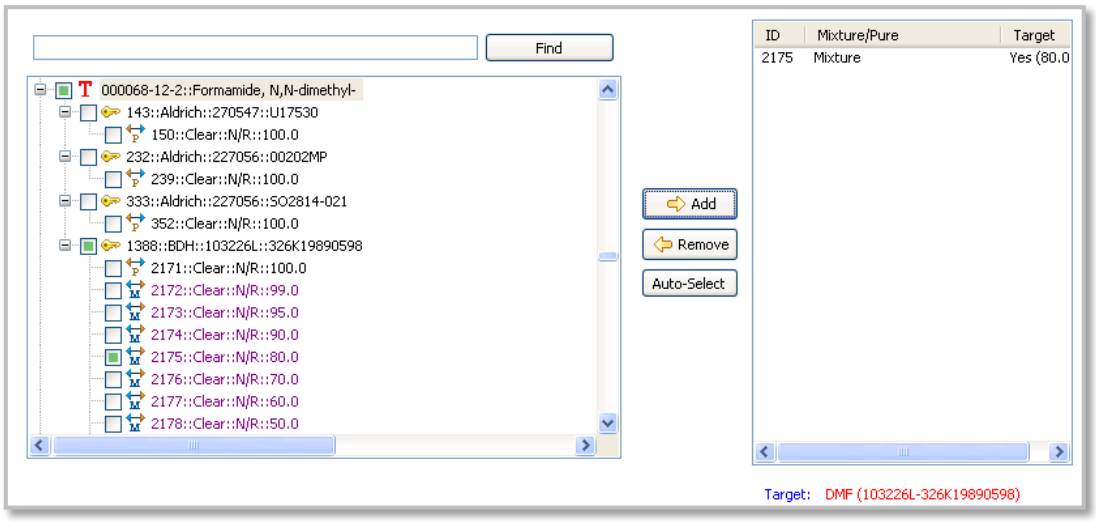


Figure 33: Adding a single spectrum to the training dataset

All of the spectra associated with one constituent can be added to the dataset by selecting the box at the start of branch for that constituent, followed by clicking the *Add* button (see Figure 34).

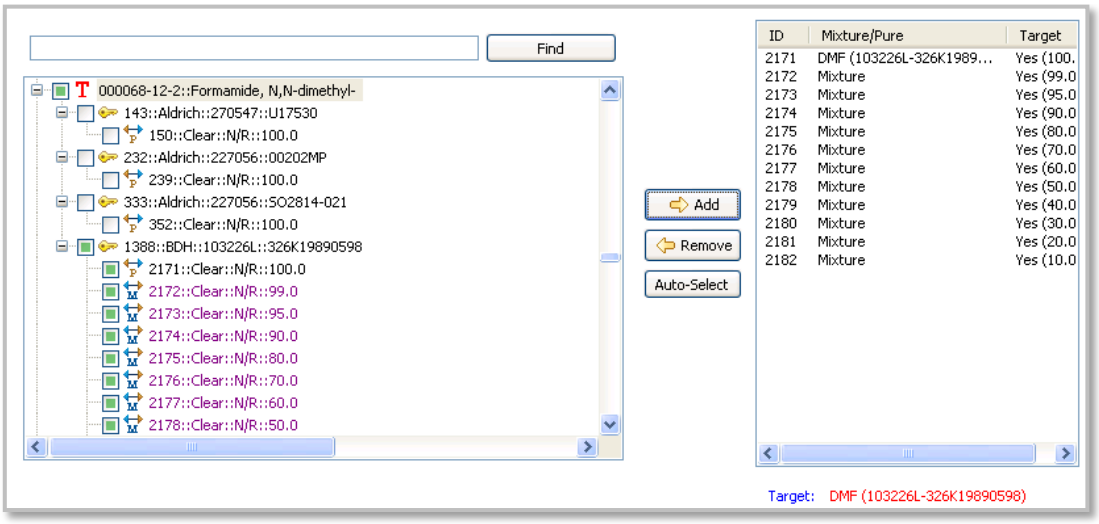


Figure 34: Adding all spectra associated with an individual constituent to the dataset

Finally, all of the spectra for a substance can be added to the current dataset, by first selecting the box at the start of the branch for that substance, and then clicking the *Add* button. This is shown in Figure 35.

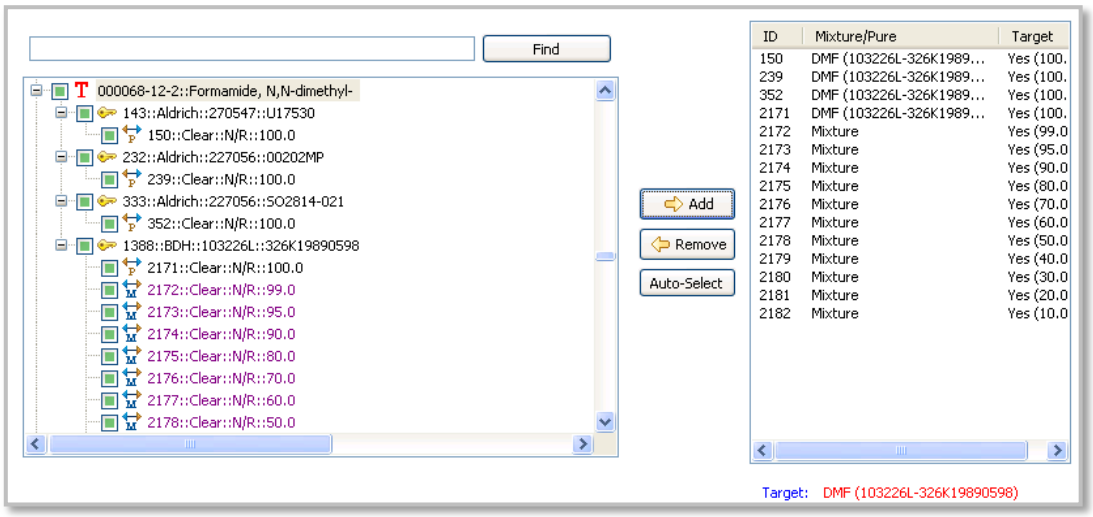


Figure 35: Adding all spectra for one substance to the training dataset

5.2.1.5 Auto-Select Feature

A convenient and powerful alternative to selecting spectra individually or in groups is to use the *Auto-Select* feature. This automatically constructs a training dataset based on the target substance that has been selected. To use this feature, select a target substance and click on the *Auto-Select* button. A dataset is constructed by automatically selecting a range of mixtures from the database, some of which contain the target substance and others that do not contain the target substance. The dataset is also constructed to contain pure samples of the target substance and pure samples of the diluents, if present in the database. The automatic dataset construction feature is a proprietary Analyze IQ technique that is useful when comparing the performance of different classification or quantification methods over a single dataset.

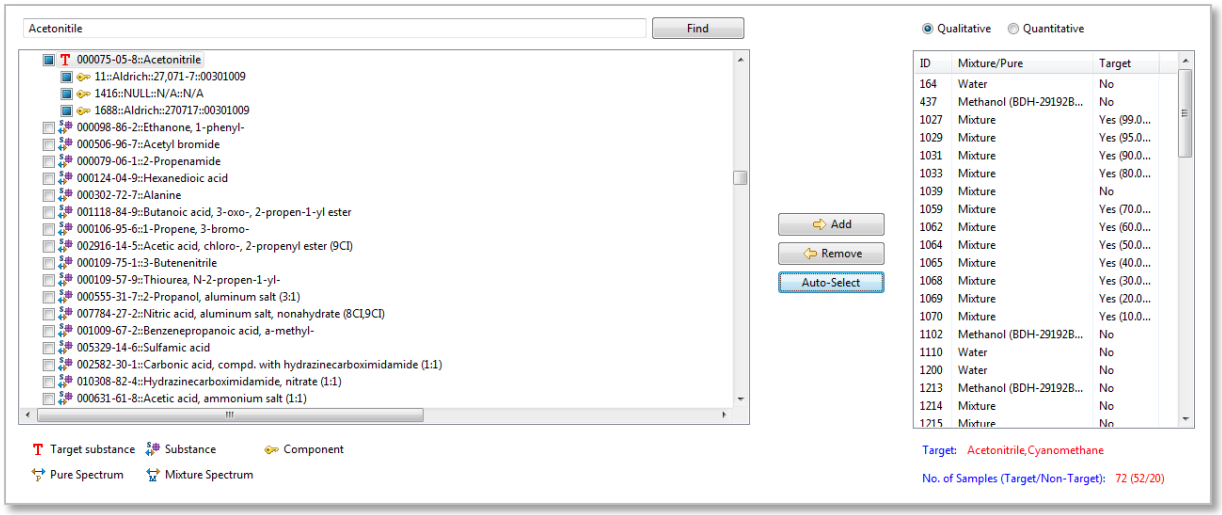


Figure 36: Using the *Auto-Select* feature

An example of the operation of the *Auto-Select* feature is shown in Figure 36. In this case, acetonitrile is selected as the target substance. After clicking the *Auto-Select* button, a dataset is constructed, which comprises 72 samples in total: 52 containing acetonitrile and 20 without the target substance. The *Auto-Select* feature is the most convenient way to quickly create a dataset.

After selecting a dataset, either manually or using the *Auto-Select* feature, press the *Next* button to move on to the next step, which is to apply pre-processing to the dataset; see Section 5.3.

5.2.1.6 Removing Spectra from the Selected List

To remove an individual spectrum from the selected list, click on it in the list on the right-hand side of the screen, and press the *Remove* button, highlighted in Figure 37.

To remove *all* spectra from the selected list, type *Ctrl-A* to *select all*, and then press the *Remove* button.

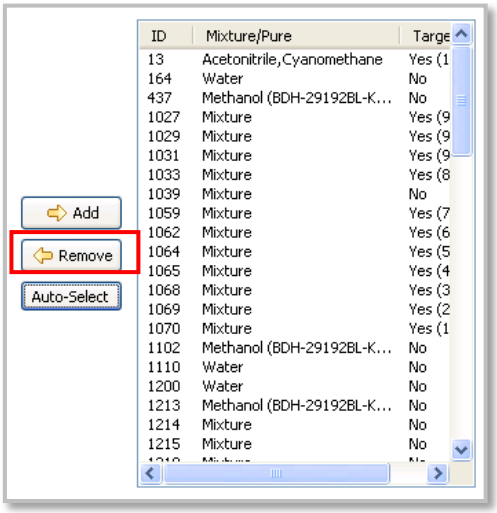


Figure 37: The *Remove* button

5.2.1.7 Change Spectra Manager Display Settings

To change the display settings for data from a ***Spectra Manager*** library, click on the blue underlined *Change Display Settings* link shown previously in Figure 27. This opens up the window shown in Figure 38. This window shows the different fields that can be displayed at each level of the library data view: *Substances*, *Components* (which is an alternative term for *Constituents*) and *Spectra*. To have a particular field displayed in the list of data, click on the checkbox beside the field name. For example, you can list the CAS Number and common for each substance in the library.

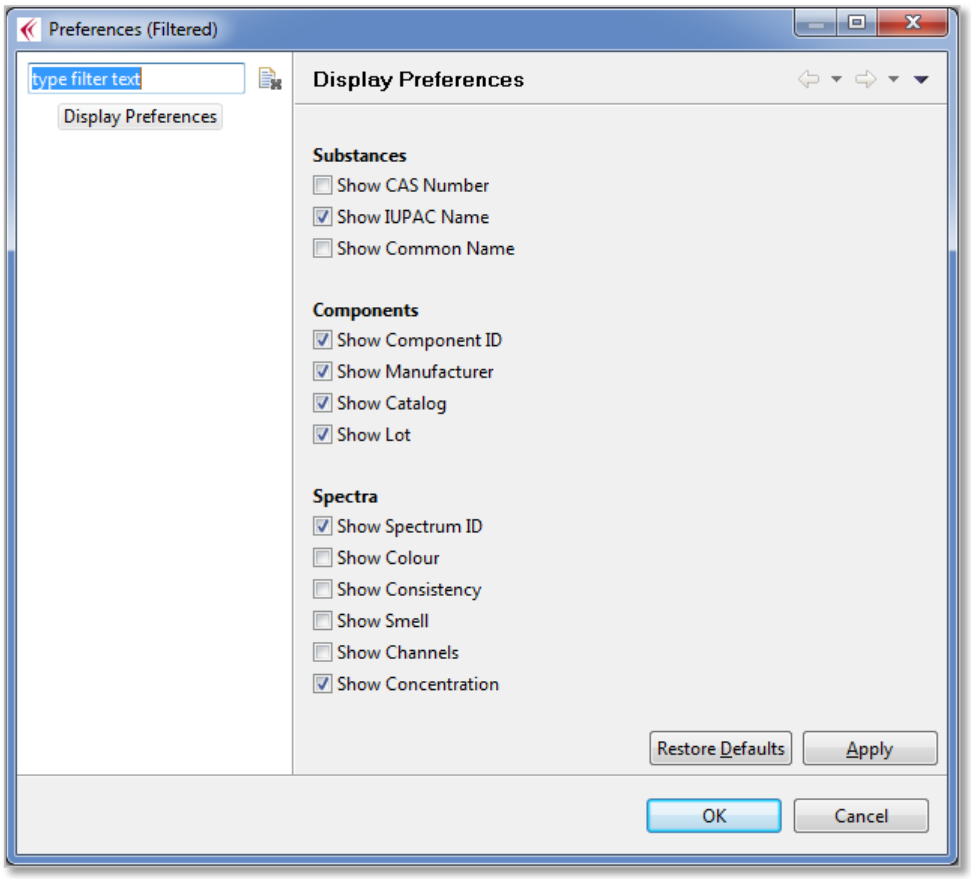


Figure 38: Display Settings for Data from Spectra Manager

5.2.1.8 Change Spectra Manager Connection Settings

Note that it is not normally advisable to change the Spectra Manager connection settings. However, this feature can be used if an organization or department wishes to maintain central libraries that multiple users can work with.

The ***Spectra Manager*** server connection settings can be changed by clicking on the *Change Settings* button, as shown in Figure 39.

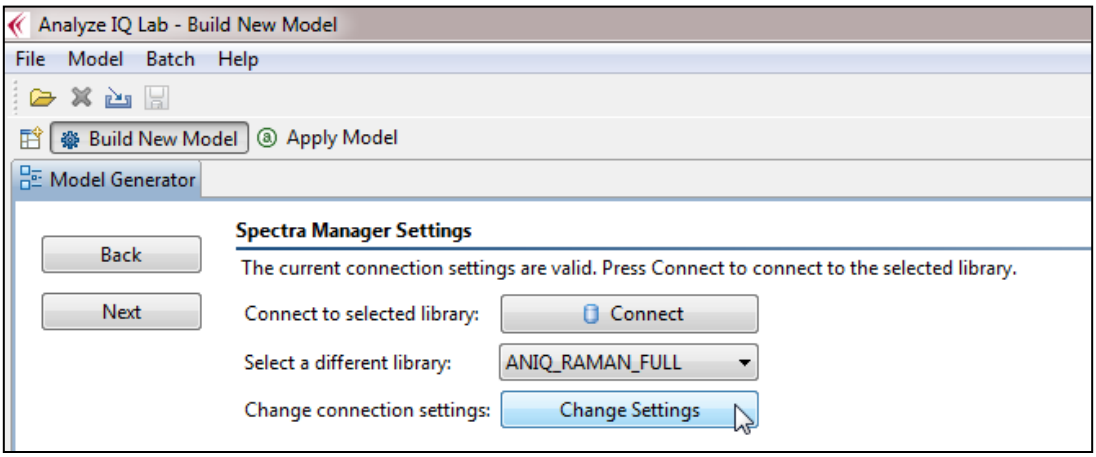


Figure 39: Change settings for connecting to a Spectra Manager Server

Clicking on the connection link opens up the window shown in Figure 40. The following settings must be specified:

1. ***Host Computer:*** The hostname or IP address of the spectral database server. If the spectral database is located on the same computer that Analyze IQ Lab is running on, enter 'localhost', as shown in Figure 40.
2. ***Instance Name:*** You must specify either an instance name or a port number (see next). The default instance name is ANALYZEIQ. You should not change this unless advised by Technical Support.
3. ***Port Number:*** You must specify either a port number or instance name. By default, the instance name is used and the port number is left blank.
4. ***Main Database:*** The primary database for your installation. You should not change this setting from its default value of ANIQ_MAIN, as shown in Figure 40, unless advised by Technical Support.
5. ***Username/Password:*** Enter your spectral database server username and password. Note that this username and password is different to the Windows username and password. The administrator of the spectral database server may provide your username and password.

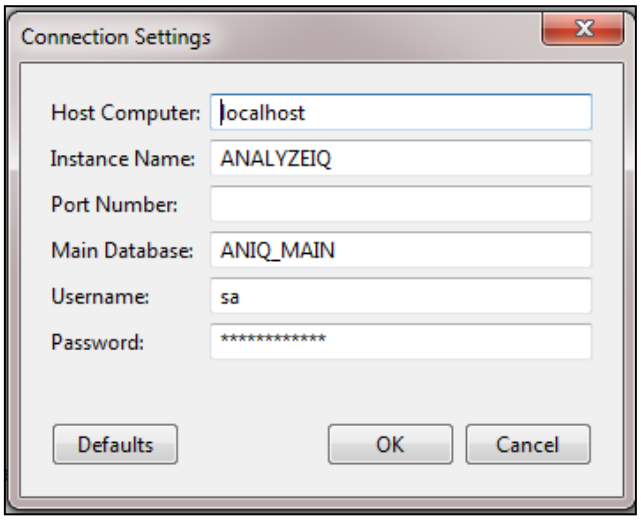


Figure 40: Database connection settings

5.2.2 Import from Single File

Section 5.2.1 has described how to construct a spectral dataset from data stored in ***Analyze IQ Spectra Manager***, the spectral data management software package. The alternatives are to import spectral data from a single file containing a table of data, as described now, or to import spectra from a folder containing multiple spectra, as will be described in Section 5.2.3.

Users commonly store their spectral data in spreadsheets using programs such as Microsoft Excel or OpenOffice Calc. Therefore, this section describes how to export data from Microsoft Excel in a format suitable to be imported into **Analyze IQ Lab**. Another common option is to transfer data into **Analyze IQ Lab** from a chemometric package such as Unscrambler; that procedure is described later in Section 5.2.4.

To import a spectral dataset from a text file, select the *Import* option and then click on the *Next* button. The user is then presented with two import options (as shown in Figure 41): *Sample Per Row* and *Sample Per Column*. In either case, the file may be a .txt file or .csv (comma separated variables) file. **Note that when importing data, the first spectrum in the file must be the pure target, or "best" example of the target material.**

To explain the format of the file for the *Sample Per Row* and *Sample Per Column* options, the procedure for importing a spectral dataset from an Excel spreadsheet in both scenarios will be outlined.

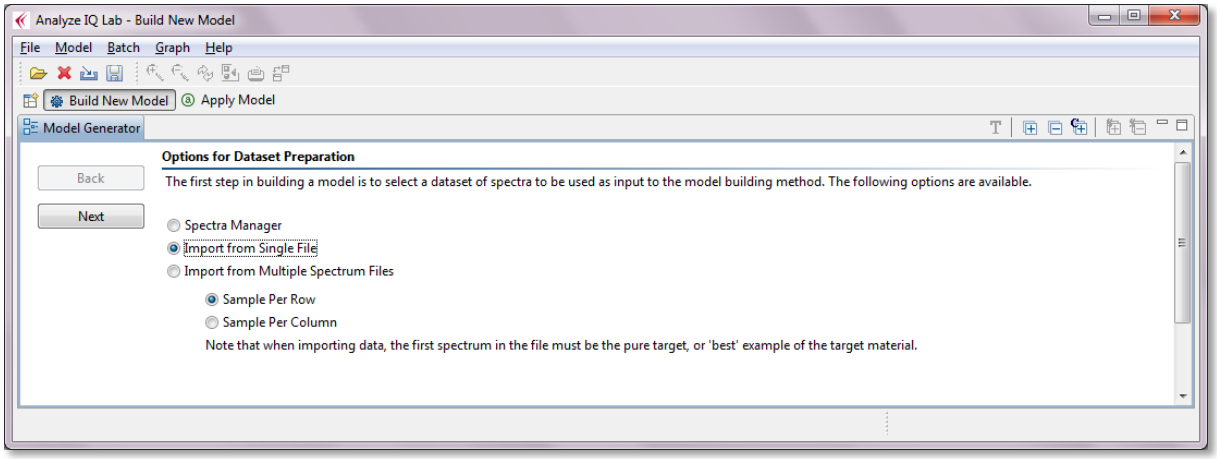


Figure 41: Import from Single File Options

5.2.2.1 Sample Per Row Spectral Dataset File

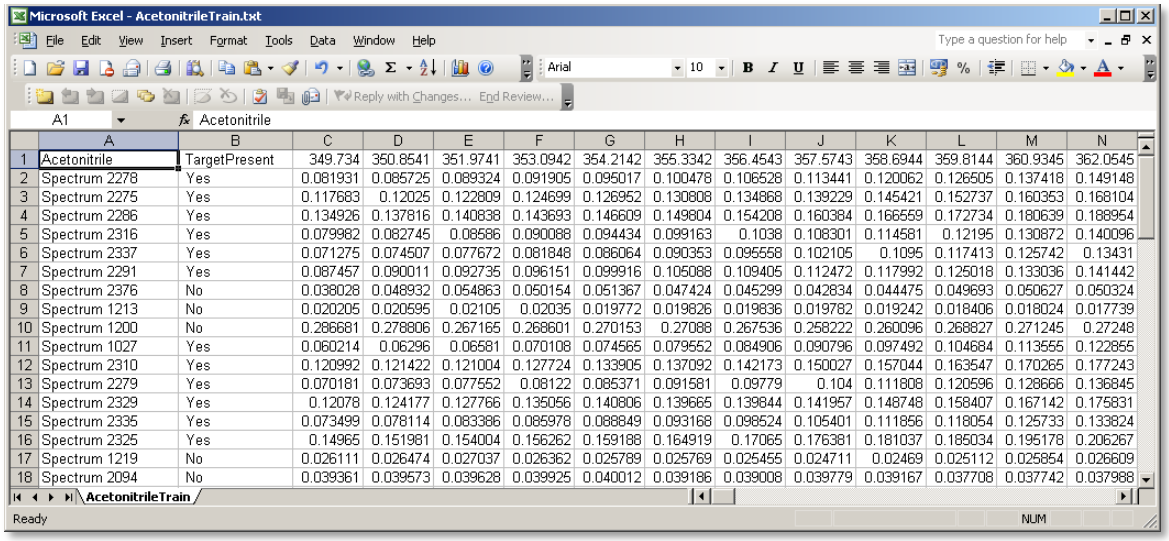
Sample Per Row is the default option for importing a spectral dataset. In this format, each row of the text file stores the information for one spectral sample, with the first row containing information pertaining to the entire dataset. A sample dataset in *Sample Per Row* format, as seen in an Excel spreadsheet, is shown in Figure 42. The details of the *Sample Per Row* format are as follows:

- The first entry in the first row of the dataset file (Cell A1 in Figure 42) contains the name of the target substance (for classification or quantification). The example in Figure 42 shows that acetonitrile is the target substance.
- The second entry in the first row of the dataset file (Cell B1 in Figure 42) is a descriptive label that should indicate whether this dataset is to be used for a classification or a quantification task, e.g. *Target Present* for a classification dataset or *Concentration* for a quantification task. Note that this label is descriptive only.
- The remaining entries in the first row of the dataset file (Cell C1 to end of Row 1 in Figure 42) indicate the recorded wavelengths or wavenumbers for each spectrum in the dataset. Note that each spectrum in the dataset must have been recorded over an identical set of wavelengths or wavenumbers.
- The second and subsequent rows contain data for individual spectra, as described next. Note that the first spectrum (i.e. the one in Row 2) must be the pure target, or "best" example of the target material.
- The first entry in all rows of the dataset file, with the exception of the first row (Cell A2 to the end of Column 1 in Figure 42) stores the ID of each sample in the dataset.
- The second entry in all rows of the dataset file, with the exception of the first row, (Cell B2 to the end of Column 2 in Figure 42) stores the target label, which differs depending on whether the dataset is to be used for classification or quantification:

- In classification, this entry contains the text 'Yes' or 'No', indicating the presence or absence of the target substance in the sample corresponding to that row.
- In quantification, this entry contains a value from 0 to 100 (inclusive), which indicates the percentage concentration of the target substance in the sample corresponding to that row.

The sample dataset shown in Figure 42 is a dataset prepared for the classification of acetonitrile.

- The third entry in all rows of the dataset file (with the exception of the first row) stores the intensity recorded at the first wavenumber or wavelength for each spectral sample. The fourth entry in all rows of the dataset file (with the exception of the first row) stores the intensity recorded at the second wavenumber or wavelength for each spectral sample, and so on.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Acetonitrile	TargetPresent	349.734	350.8541	351.9741	353.0942	354.2142	355.3342	356.4543	357.5743	358.6944	359.8144	360.9345	362.0545
2	Spectrum 2278	Yes	0.081931	0.085725	0.089324	0.091905	0.095017	0.100478	0.106528	0.113441	0.120062	0.126505	0.137418	0.149148
3	Spectrum 2275	Yes	0.117683	0.12025	0.122809	0.124699	0.126952	0.130808	0.134868	0.139229	0.145421	0.152737	0.160353	0.168104
4	Spectrum 2286	Yes	0.134926	0.137816	0.140838	0.143693	0.146609	0.149804	0.154208	0.160384	0.166559	0.172734	0.180639	0.188954
5	Spectrum 2316	Yes	0.079982	0.082745	0.08586	0.090088	0.094434	0.099163	0.1038	0.108301	0.114581	0.12195	0.130872	0.140096
6	Spectrum 2337	Yes	0.071275	0.074507	0.077672	0.081848	0.086064	0.090353	0.095558	0.102105	0.1095	0.117413	0.125742	0.13431
7	Spectrum 2291	Yes	0.087457	0.090011	0.092735	0.096151	0.099916	0.105088	0.109405	0.112472	0.117992	0.125018	0.133036	0.141442
8	Spectrum 2376	No	0.038028	0.048932	0.054863	0.050154	0.051367	0.047424	0.045299	0.042834	0.044475	0.049693	0.050627	0.050324
9	Spectrum 1213	No	0.020205	0.020595	0.02105	0.02035	0.019772	0.019826	0.019836	0.019782	0.019242	0.018406	0.018024	0.017739
10	Spectrum 1200	No	0.286681	0.278806	0.267165	0.268601	0.270153	0.27088	0.267536	0.258222	0.260096	0.268827	0.271245	0.27248
11	Spectrum 1027	Yes	0.060214	0.06296	0.06581	0.070108	0.074565	0.079552	0.084906	0.090796	0.097492	0.104684	0.113555	0.122855
12	Spectrum 2310	Yes	0.120992	0.121422	0.121004	0.127724	0.133905	0.137092	0.142173	0.150027	0.157044	0.163547	0.170265	0.177243
13	Spectrum 2279	Yes	0.070181	0.073693	0.077552	0.08122	0.085371	0.091581	0.09779	0.104	0.111808	0.120596	0.128666	0.136845
14	Spectrum 2329	Yes	0.12078	0.124177	0.127766	0.135056	0.140806	0.139665	0.139844	0.141957	0.148748	0.158407	0.167142	0.175831
15	Spectrum 2335	Yes	0.073499	0.078114	0.083386	0.085978	0.088849	0.093168	0.098524	0.105401	0.111856	0.118054	0.125733	0.133824
16	Spectrum 2325	Yes	0.14965	0.151981	0.154004	0.156262	0.159188	0.164919	0.17065	0.176381	0.181037	0.185034	0.195178	0.206267
17	Spectrum 1219	No	0.026111	0.026474	0.027037	0.026362	0.025789	0.025769	0.025455	0.024711	0.02469	0.025112	0.025854	0.026609
18	Spectrum 2094	No	0.039361	0.039573	0.039628	0.039925	0.040012	0.039186	0.039008	0.039779	0.039167	0.037708	0.037742	0.037988

Figure 42: Sample dataset in Excel in *Sample Per Row* format

When a set of spectral data samples has been arranged according to the above rules in an Excel spreadsheet, it can then be exported into a .txt file (remember to select the comma as the separator) or a .csv file, ready for importing into **Analyze IQ Lab** under the *Sample Per Row* option. For importing Excel data into **Analyze IQ Lab**, it is recommended to export the data to a .csv file.

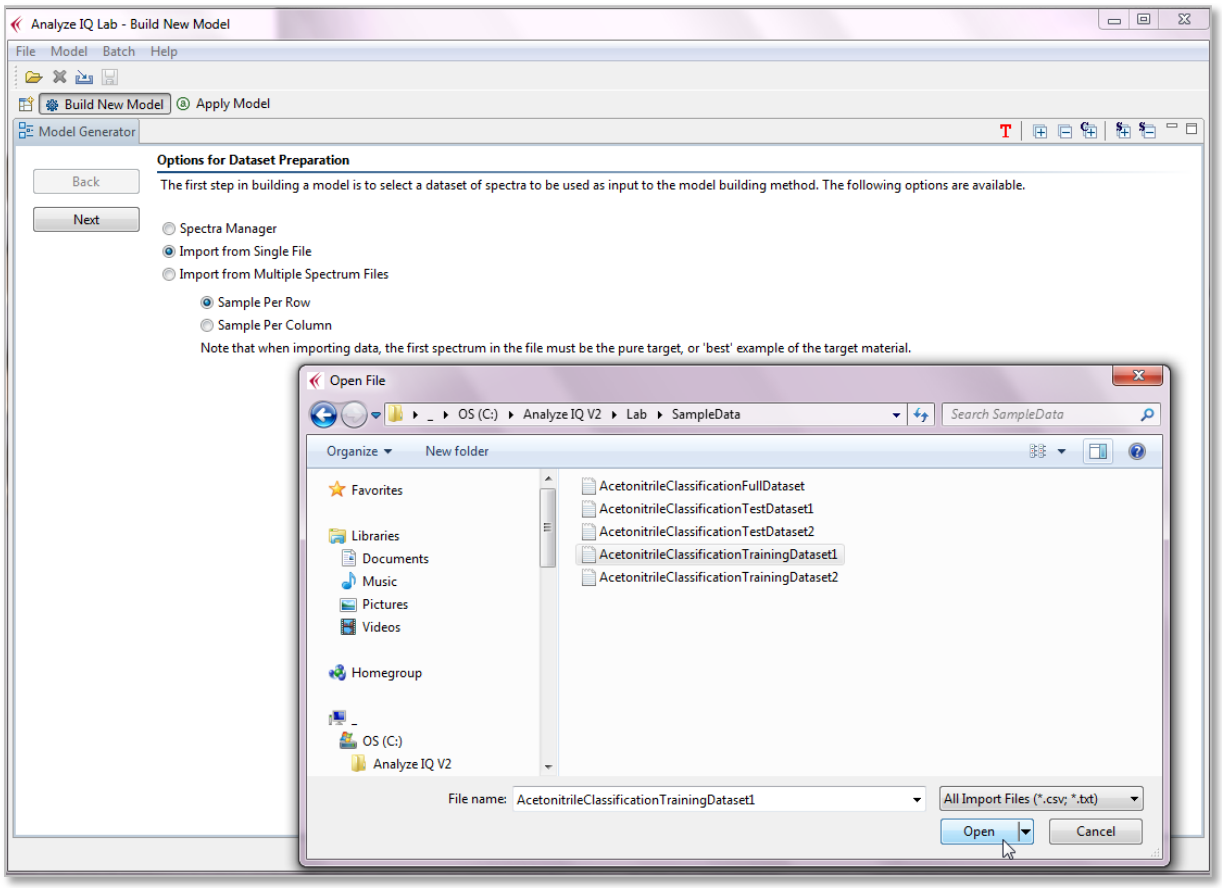


Figure 43: Importing a dataset using Sample Per Row option

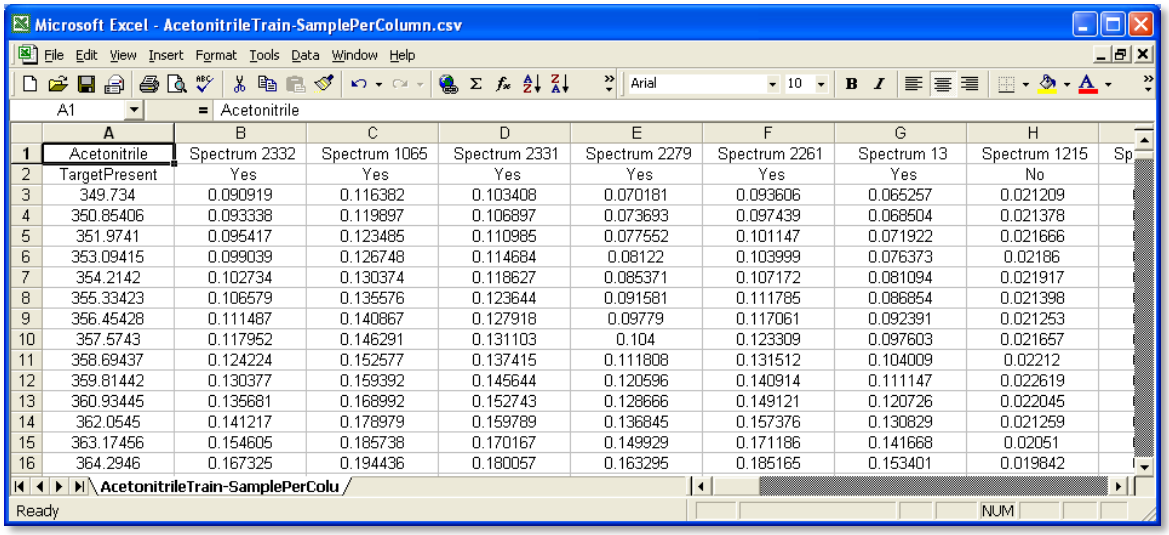
To import a spectral dataset file saved in *Sample Per Row* format, select the *Sample Per Row* option (as shown earlier in Figure 41) and click on *Next*. The user is then prompted to select the dataset file (.txt or .csv), as shown in Figure 43. After selecting the dataset and clicking the *OK* button, the contents of the dataset are presented in tabular form (see Figure 54 in Section 5.3).

Refer to Section 5.3 for the next steps in the model building process.

5.2.2.2 Sample Per Column Spectral Dataset File

For the case where the spectral samples are too large for Excel[2] to store on a sample per row basis, a spectral dataset may be stored in *Sample Per Column* format. In the *Sample Per Column* format each sample occupies a separate column of the spreadsheet, as shown in Figure 44.

[2] Excel 2003 and prior versions have limit of 256 columns, which would only allow for the storage of spectral samples recorded over 254 wavelengths or wavenumbers.



	A	B	C	D	E	F	G	H
1	Acetonitrile	Spectrum 2332	Spectrum 1065	Spectrum 2331	Spectrum 2279	Spectrum 2261	Spectrum 13	Spectrum 1215
2	TargetPresent	Yes	Yes	Yes	Yes	Yes	Yes	No
3	349.734	0.090919	0.116382	0.103408	0.070181	0.093606	0.065257	0.021209
4	350.85406	0.093338	0.119897	0.106897	0.073693	0.097439	0.068504	0.021378
5	351.9741	0.095417	0.123485	0.110985	0.077552	0.101147	0.071922	0.021666
6	353.09415	0.099039	0.126748	0.114684	0.08122	0.103999	0.076373	0.02186
7	354.2142	0.102734	0.130374	0.118627	0.085371	0.107172	0.081094	0.021917
8	355.33423	0.106579	0.135576	0.123644	0.091581	0.111785	0.086854	0.021398
9	356.45428	0.111487	0.140867	0.127918	0.09779	0.117061	0.092391	0.021253
10	357.5743	0.117952	0.146291	0.131103	0.104	0.123309	0.097603	0.021657
11	358.69437	0.124224	0.152577	0.137415	0.111808	0.131512	0.104009	0.02212
12	359.81442	0.130377	0.159392	0.145644	0.120596	0.140914	0.111147	0.022619
13	360.93445	0.135681	0.168992	0.152743	0.128666	0.149121	0.120726	0.022045
14	362.0545	0.141217	0.178979	0.159789	0.136845	0.157376	0.130829	0.021259
15	363.17456	0.154605	0.185738	0.170167	0.149929	0.171186	0.141668	0.02051
16	364.2946	0.167325	0.194436	0.180057	0.163295	0.185165	0.153401	0.019842

Figure 44: Sample dataset in Excel in Sample Per Column format

The *Sample Per Column* format is the same as the *Sample Per Row* format that was described above in Section 5.2.2.1, except that the rows and columns are interchanged. The details of the *Sample Per Column* format are as follows:

- The first entry in the first column of the dataset file (Cell A1 in Figure 44) contains the name of the target substance (for classification or quantification). The example in Figure 44 shows that acetonitrile is the target substance.
- The second entry in the first column of the dataset file (Cell A2 in Figure 44) is a descriptive label that should indicate whether this dataset is to be used for a classification or a quantification task, e.g. *Target Present* for a classification dataset or *Concentration* for a quantification task. Note that this label is descriptive only.
- The remaining entries in the first column of the dataset file (Cell A3 to end of Column 1 in Figure 44) indicate the recorded wavelengths or wavenumbers for each spectrum in the dataset. Note that each spectrum in the dataset must have been recorded over an identical set of wavelengths or wavenumbers.
- The second and subsequent columns contain data for individual spectra, as described next. Note that the first spectrum (i.e. the one in Column 2) must be the pure target, or "best" example of the target material.
- The first entry in all columns of the dataset file, with the exception of the first column (Cell B1 to the end of Row 1 in Figure 44) stores the ID of each sample in the dataset.
- The second entry in all columns of the dataset file, with the exception of the first row, (Cell B2 to the end of Row 2 in Figure 44) stores the target label, which differs depending on whether the dataset is to be used for classification or quantification:

    - In classification, this entry contains the label 'Yes' or 'No', indicating the presence or absence of the target substance in the sample corresponding to that row.

- In quantification, this entry contains a value from 0 to 100 (inclusive), which indicates the percentage concentration of the target substance in the sample corresponding to that row.

The sample dataset shown in Figure 44 is a dataset prepared for the classification of acetonitrile.

- The third entry in all columns of the dataset file (with the exception of the first column) stores the intensity recorded at the first wavenumber or wavelength for each spectral sample. The fourth entry in all columns of the dataset file (with the exception of the first column) stores the intensity recorded at the second wavenumber or wavelength for each spectral sample, and so on.

When a set of spectral data samples has been arranged according to the above rules in an Excel spreadsheet, it can then be exported into a .txt file (remember to select the comma as the separator) or a .csv file, ready for importing into **Analyze IQ Lab** under the *Sample Per Column* option. For importing Excel data into **Analyze IQ Lab**, it is recommended to export the data to a .csv file.

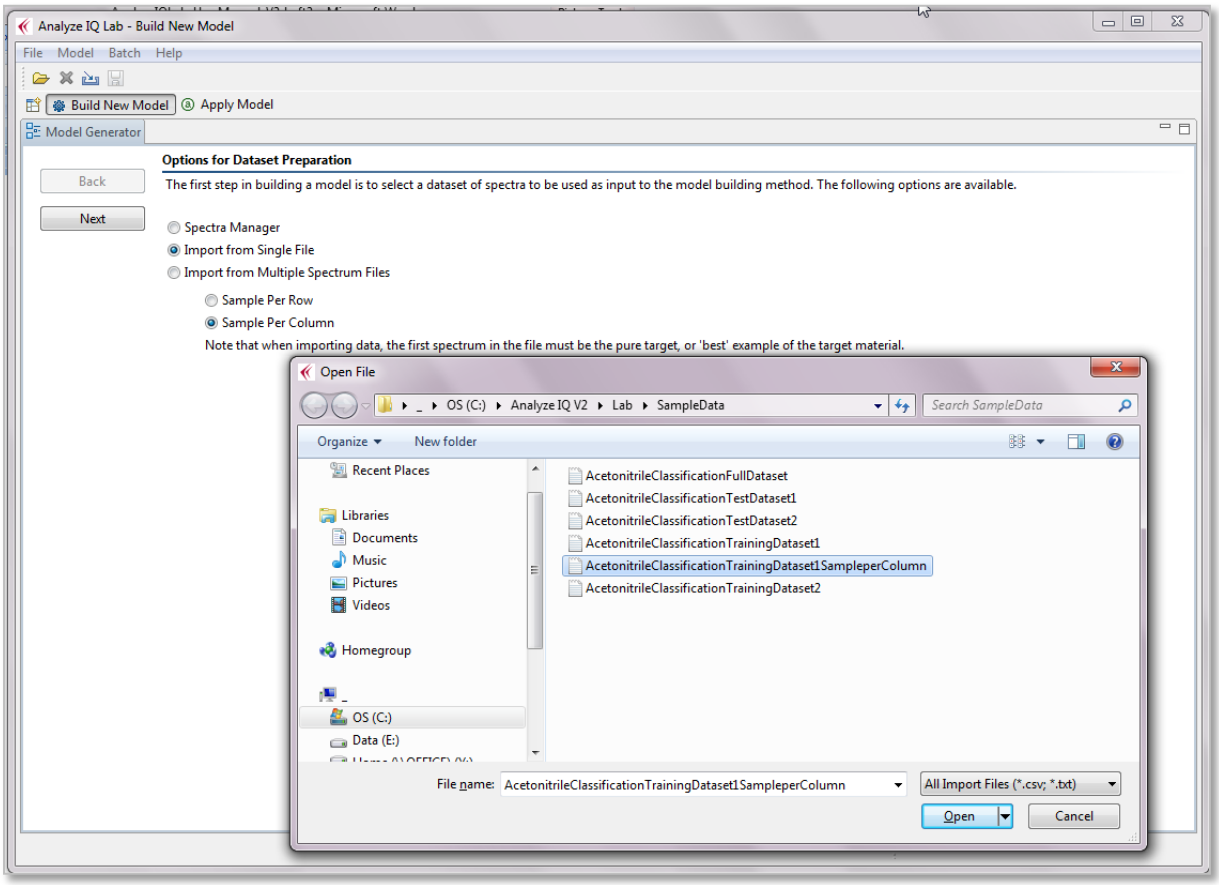


Figure 45: Importing a dataset using Sample Per Column option

To import a spectral dataset file saved in *Sample Per Column* format, select the *Sample Per Column* option (shown earlier in Figure 41) and click on *Next*. The user is then prompted to select the dataset file (.txt or .csv), as shown in Figure 45. After

selecting the dataset and clicking the *OK* button, the contents of the dataset are presented in tabular form (see Figure 54 in Section 5.3).

Note that when a file in *Sample Per Column* format is imported, its data is rotated automatically so that it its contents are presented in rows on screen.

Refer to Section 5.3 for the next steps in the model building process.

5.2.3 Import from Multiple Spectrum Files

This option allows you to build a model from data in the form of a folder containing multiple spectrum files. The spectrum files may be in any of the formats supported by Analyze IQ Lab; refer back to Section 4.2.2 for a list of them.

When you select this option, you must to provide two pieces of information, as shown in Figure 46:

1. The name of the folder containing the spectra; use the *Browse* button to select this
2. A list of all spectrum file names to be imported, along with the target value corresponding to each one; use the *Select* button to select this.

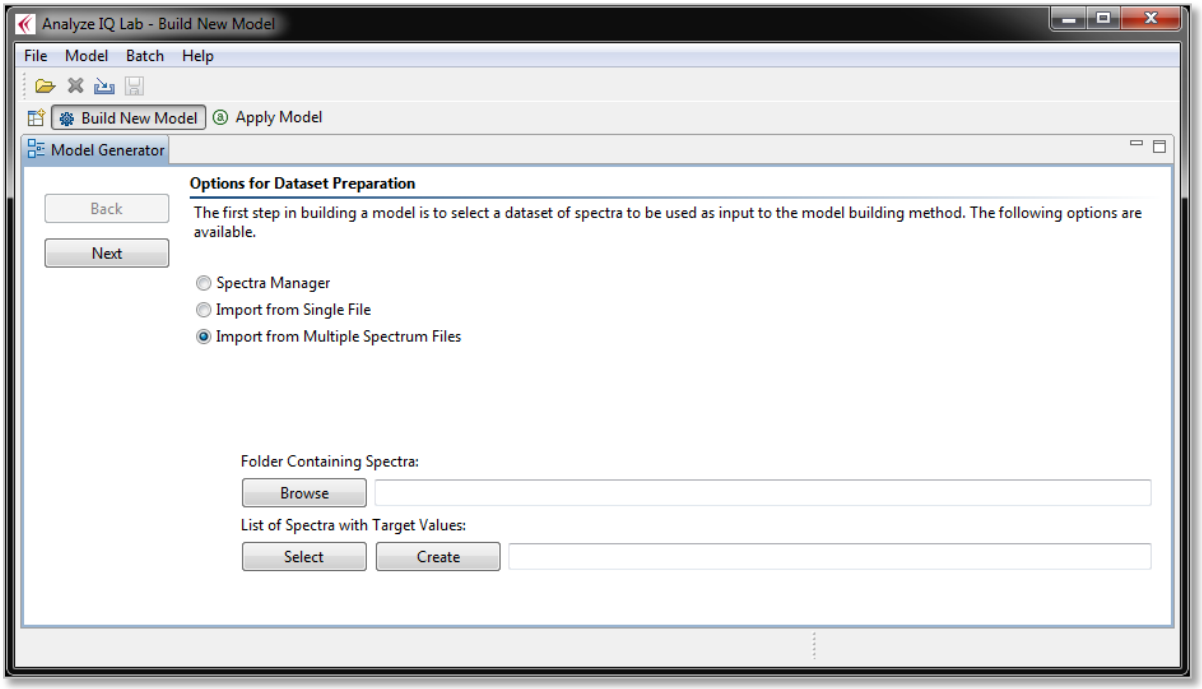


Figure 46: Preparing a Dataset from Multiple Spectrum Files

The list of spectra with target values must be a text file with comma-separated values, with either a .csv or .txt extension. Each line of the file must contain two items, separated by commas:

```
Spectrum file name, target value for model
```

Note that the first line of the file can optionally be a header. Figure 47 shows an example of a file like this, as viewed in Microsoft Excel:

	A	B
1	File name	Value
2	Acetonitrile-Methanol (50-50).aiq	50
3	Acetonitrile-Methanol-Water (10-70-20).aiq	10
4	Acetonitrile-Methanol-Water (10-80-10).aiq	10
5	Acetonitrile-Methanol-Water (20-20-60).aiq	20
6	Acetonitrile-Water (50-50).aiq	50
7	Methanol-Water (80-20).aiq	0
8	Methanol-Water (99-1).aiq	0
9		

Figure 47: Example of a List of Spectra with Target Values

If you do not have a file like this prepared, you can use the *Create* button to create one. This operates as follows:

1. It prompts you for a file extension (for example, SPC or AIQ)
2. It scans the folder you selected for all files with that extension
3. It creates a CSV file containing all of the names

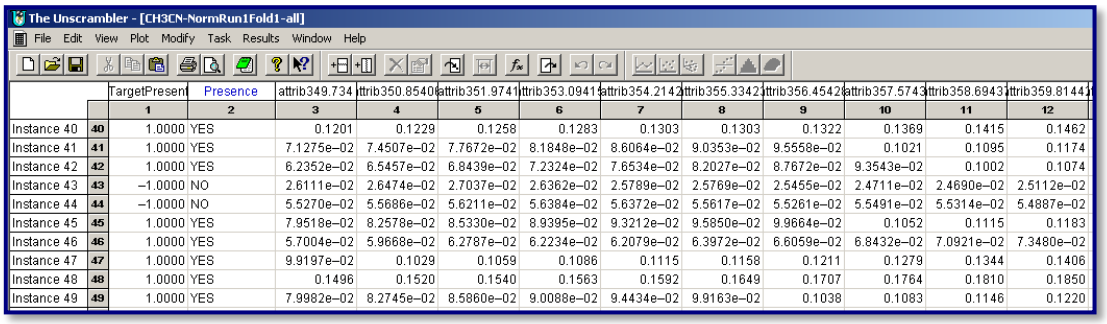
One thing that *Create* cannot do is specify the target values; this will depend on your model's target. Therefore, after creating the file, you must edit it. The button label changes to *Edit,* and when you click on Edit, the file is opened in whatever program you normally use to edit CSV files (typically Microsoft Excel). After entering the target values corresponding to each spectrum, please take care to re-save the file in CSV format, not any other format such as XLSX or XLS.

After having done that, press the Next button to proceed to the next step in the model building process, as will be described in Section 5.3.

5.2.4 Importing Data From Unscrambler

To import data from Unscrambler, take the following steps (the example of a classification dataset is used):

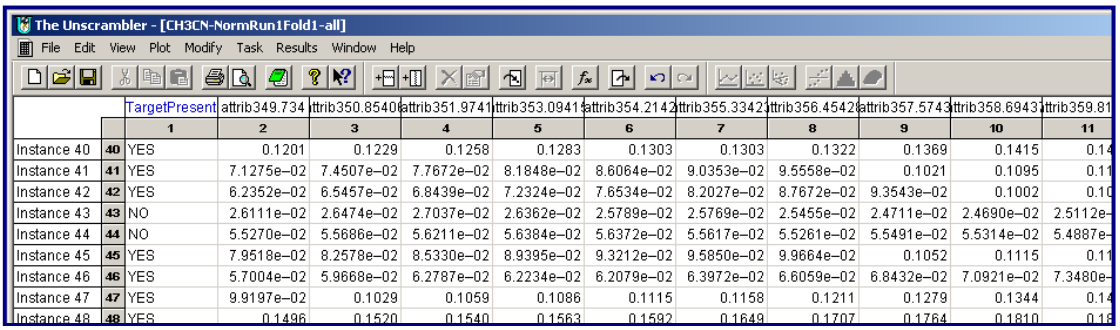
1. Open the data file in Unscrambler, as shown in Figure 48.



		TargetPresent	Presence	attrib349.734	attrib350.8540	attrib351.9741	attrib353.0941	attrib354.2142	attrib355.3342	attrib356.4542	attrib357.5743	attrib358.6943	attrib359.8144
		1	2	3	4	5	6	7	8	9	10	11	12
Instance 40	40	1.0000	YES	0.1201	0.1229	0.1258	0.1283	0.1303	0.1303	0.1322	0.1369	0.1415	0.1462
Instance 41	41	1.0000	YES	7.1275e–02	7.4507e–02	7.7672e–02	8.1848e–02	8.6064e–02	9.0353e–02	9.5558e–02	0.1021	0.1095	0.1174
Instance 42	42	1.0000	YES	6.2352e–02	6.5457e–02	6.8439e–02	7.2324e–02	7.6534e–02	8.2027e–02	8.7672e–02	9.3543e–02	0.1002	0.1074
Instance 43	43	–1.0000	NO	2.6111e–02	2.6474e–02	2.7037e–02	2.6362e–02	2.5789e–02	2.5769e–02	2.5455e–02	2.4711e–02	2.4690e–02	2.5112e–02
Instance 44	44	–1.0000	NO	5.5270e–02	5.5686e–02	5.6211e–02	5.6384e–02	5.6372e–02	5.5617e–02	5.5261e–02	5.5491e–02	5.5314e–02	5.4887e–02
Instance 45	45	1.0000	YES	7.9518e–02	8.2578e–02	8.5330e–02	8.9395e–02	9.3212e–02	9.5850e–02	9.9664e–02	0.1052	0.1115	0.1183
Instance 46	46	1.0000	YES	5.7004e–02	5.9668e–02	6.2787e–02	6.2234e–02	6.2079e–02	6.3972e–02	6.6059e–02	6.8432e–02	7.0921e–02	7.3480e–02
Instance 47	47	1.0000	YES	9.9197e–02	0.1029	0.1059	0.1086	0.1115	0.1158	0.1211	0.1279	0.1344	0.1406
Instance 48	48	1.0000	YES	0.1496	0.1520	0.1540	0.1563	0.1592	0.1649	0.1707	0.1764	0.1810	0.1850
Instance 49	49	1.0000	YES	7.9982e–02	8.2745e–02	8.5860e–02	9.0088e–02	9.4434e–02	9.9163e–02	0.1038	0.1083	0.1146	0.1220

Figure 48: Spectral dataset in Unscrambler

2. Delete the column of numerical values (which correspond to Yes/No) and delete any spectra not to be used for building the model. You should then be presented with the dataset shown in Figure 49.



		TargetPresent	attrib349.734	attrib350.8540	attrib351.9741	attrib353.0941	attrib354.2142	attrib355.3342	attrib356.4542	attrib357.5743	attrib358.6943	attrib359.81
		1	2	3	4	5	6	7	8	9	10	11
Instance 40	40	YES	0.1201	0.1229	0.1258	0.1283	0.1303	0.1303	0.1322	0.1369	0.1415	0.14
Instance 41	41	YES	7.1275e–02	7.4507e–02	7.7672e–02	8.1848e–02	8.6064e–02	9.0353e–02	9.5558e–02	0.1021	0.1095	0.11
Instance 42	42	YES	6.2352e–02	6.5457e–02	6.8439e–02	7.2324e–02	7.6534e–02	8.2027e–02	8.7672e–02	9.3543e–02	0.1002	0.10
Instance 43	43	NO	2.6111e–02	2.6474e–02	2.7037e–02	2.6362e–02	2.5789e–02	2.5769e–02	2.5455e–02	2.4711e–02	2.4690e–02	2.5112e–
Instance 44	44	NO	5.5270e–02	5.5686e–02	5.6211e–02	5.6384e–02	5.6372e–02	5.5617e–02	5.5261e–02	5.5491e–02	5.5314e–02	5.4887e–
Instance 45	45	YES	7.9518e–02	8.2578e–02	8.5330e–02	8.9395e–02	9.3212e–02	9.5850e–02	9.9664e–02	0.1052	0.1115	0.11
Instance 46	46	YES	5.7004e–02	5.9668e–02	6.2787e–02	6.2234e–02	6.2079e–02	6.3972e–02	6.6059e–02	6.8432e–02	7.0921e–02	7.3480e–
Instance 47	47	YES	9.9197e–02	0.1029	0.1059	0.1086	0.1115	0.1158	0.1211	0.1279	0.1344	0.14
Instance 48	48	YES	0.1496	0.1520	0.1540	0.1563	0.1592	0.1649	0.1707	0.1764	0.1810	0.18

Figure 49: Spectral dataset in Unscrambler, with first column showing TargetPresent as a category variable (Yes/No)

3. Under the *File* menu, select *Export*, to open the export dialog shown in Figure 50. Select *ASCII Files* as the format, enter the filename and click on *Export*.

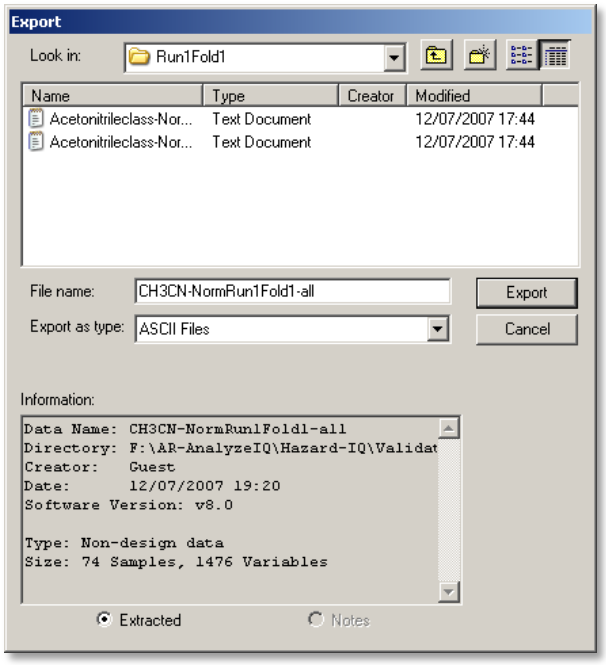


Figure 50: Unscrambler Export window

4. The Export ASCII window shown in Figure 51 is then opened. Select the settings shown in this figure: File format = Wide ASCII (Include variable names, Include sample names); Name qualifier = Double Quote; Numeric qualifier = None; and Item delimiter = Comma.

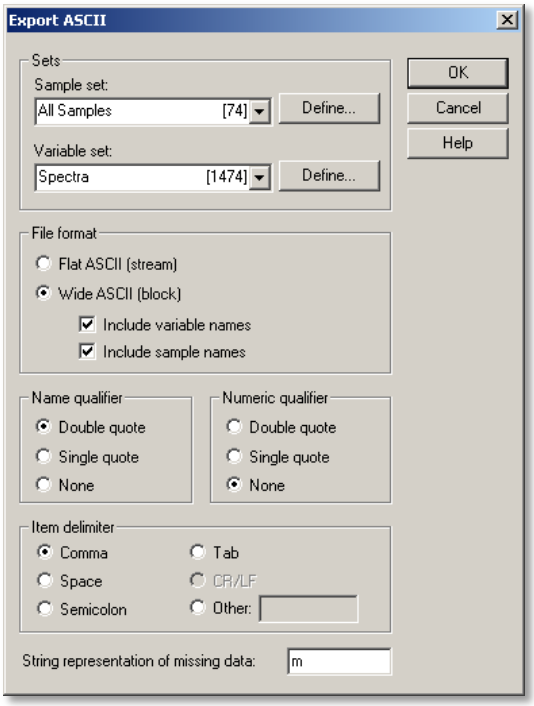


Figure 51: Unscrambler Export ASCII window

5. Open the file exported by Unscrambler in a text editor; Figure 52 shows a text file exported by Unscrambler as viewed in Wordpad.

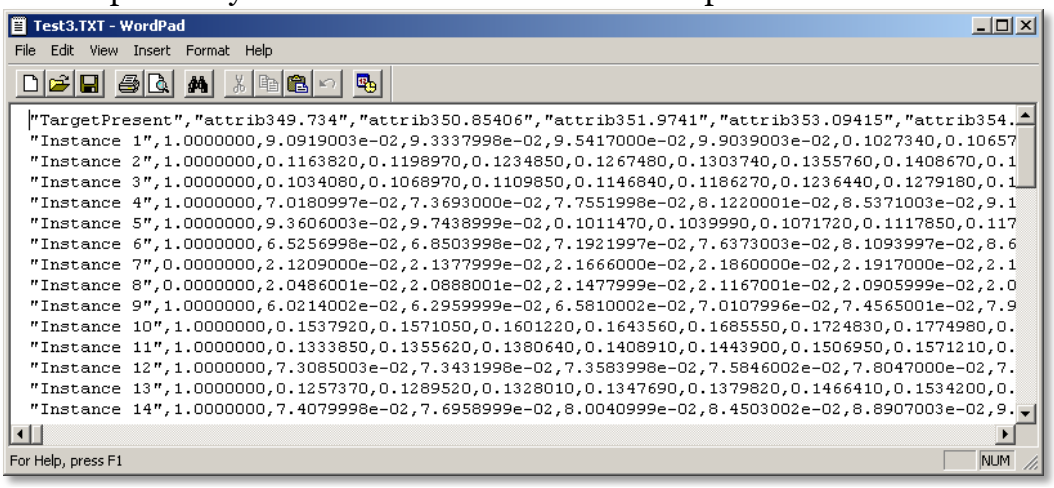


Figure 52: Text file exported by Unscrambler

6. Make the following modifications to the dataset text file:

    a. Enter the target name followed by a comma, e.g. “Acetonitrile,” at start of the first line.
    b. In the second column (under “Target Present”), replace each 1.000000 with the word “Yes” and each 0.0000000 with the word “No”. It is *not* recommended to use an automatic Search/Replace utility to carry out this, as there is a danger that spectral intensity values may be inadvertently altered (especially in the case of spectral datasets that have already been normalised or baseline-corrected).

    Note the above modifications could also be carried out in Excel and then exported to a .txt or .csv file (as described in Section 5.2.2.1). The step of replacing each 1.000000 with “Yes” and each 0.0000000 with “No” is more easily done in Excel without errors, since you can do a Search & Replace on a single column of data.

    The final text should look similar to that shown in Figure 53. This text file is ready for import under the *Sample Per Row* format option (see Section 5.2.2.1).

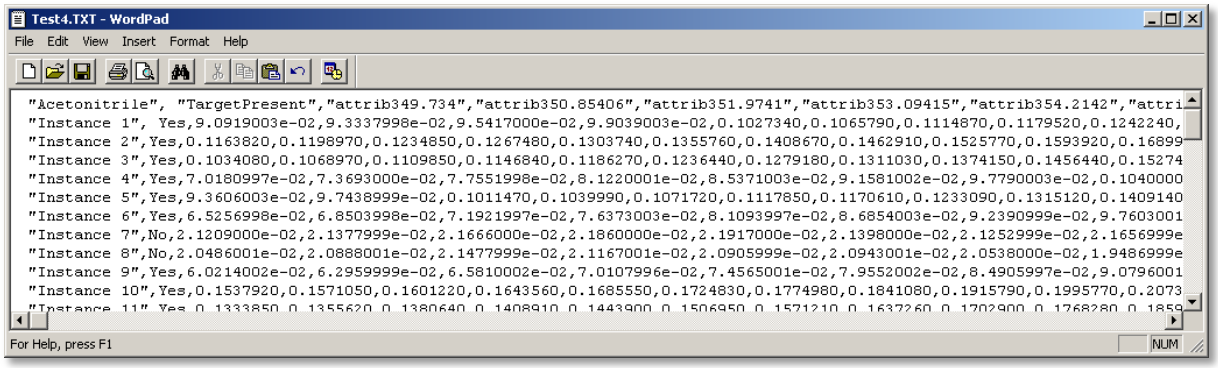


Figure 53: Text file ready for import into Analyze IQ Lab

5.3 Apply Pre-processing to Dataset

After importing a dataset or generating it from ***Spectra Manager***, the contents of the dataset are presented to the user, as shown in Figure 54. For the purposes of continuing the example of building a classification model, we assume that the dataset was drawn from ***Spectra Manager***, with acetontrile as the target and auto-selecting 72 spectra.

As part of the model building, the user can specify any pre-processing steps that will be carried out on the dataset before passing it to the classification or quantification algorithm.

When building a new model, all of the same pre-processing methods are available as when viewing a spectrum. Refer back to Section 4.3.1 for details.

Note that pre-processing steps can be applied repeatedly if desired; **Analyze IQ Lab** will keep track of the sequence of pre-processing steps applied.

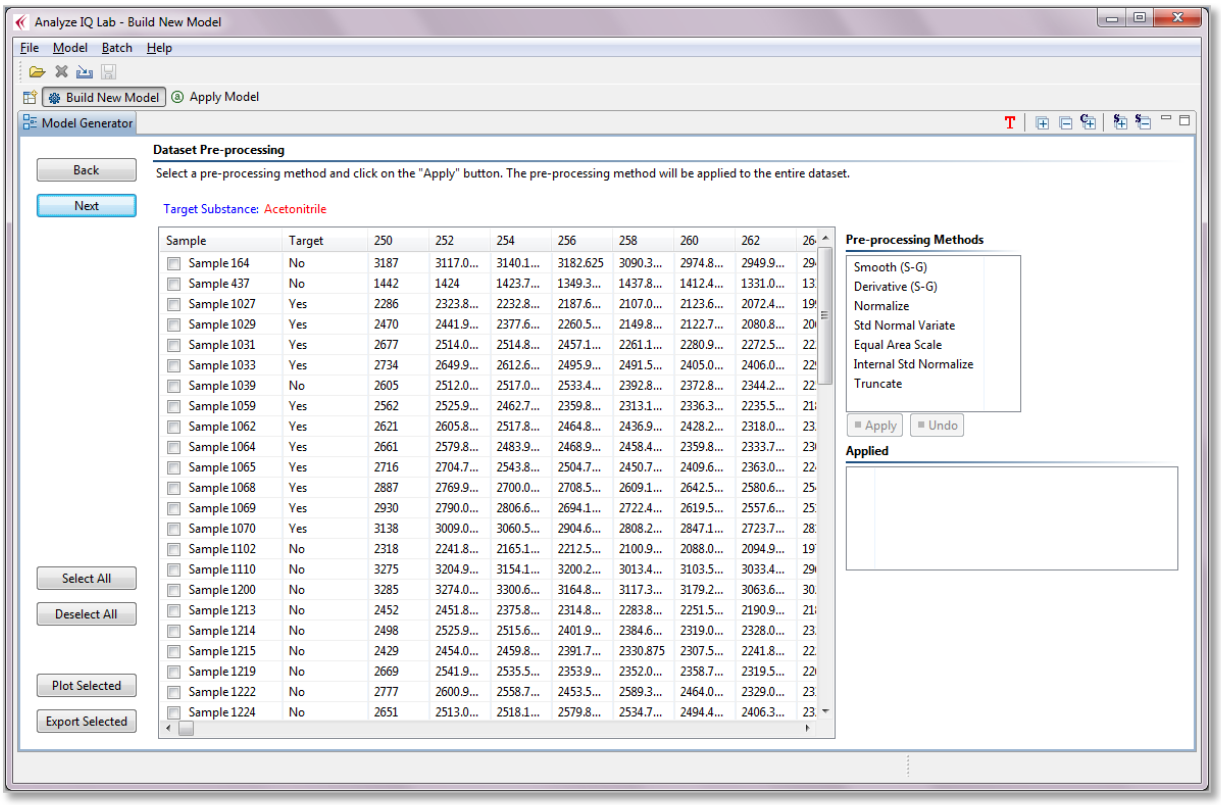


Figure 54: Spectral dataset prior to applying pre-processing

To apply normalization, select the *Normalization* option in the list of pre-processing methods and click on the *Apply* button, as shown in Figure 55. Note that there are no parameter settings for normalization.

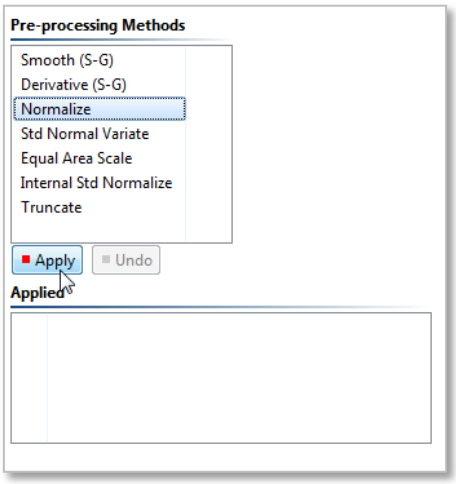


Figure 55: Applying the Normalization pre-processing method

After applying any pre-processing method, the contents of the spectral dataset shown in the table are updated and the list of already applied methods is also updated; Figure 56 shows the current dataset after applying normalization.

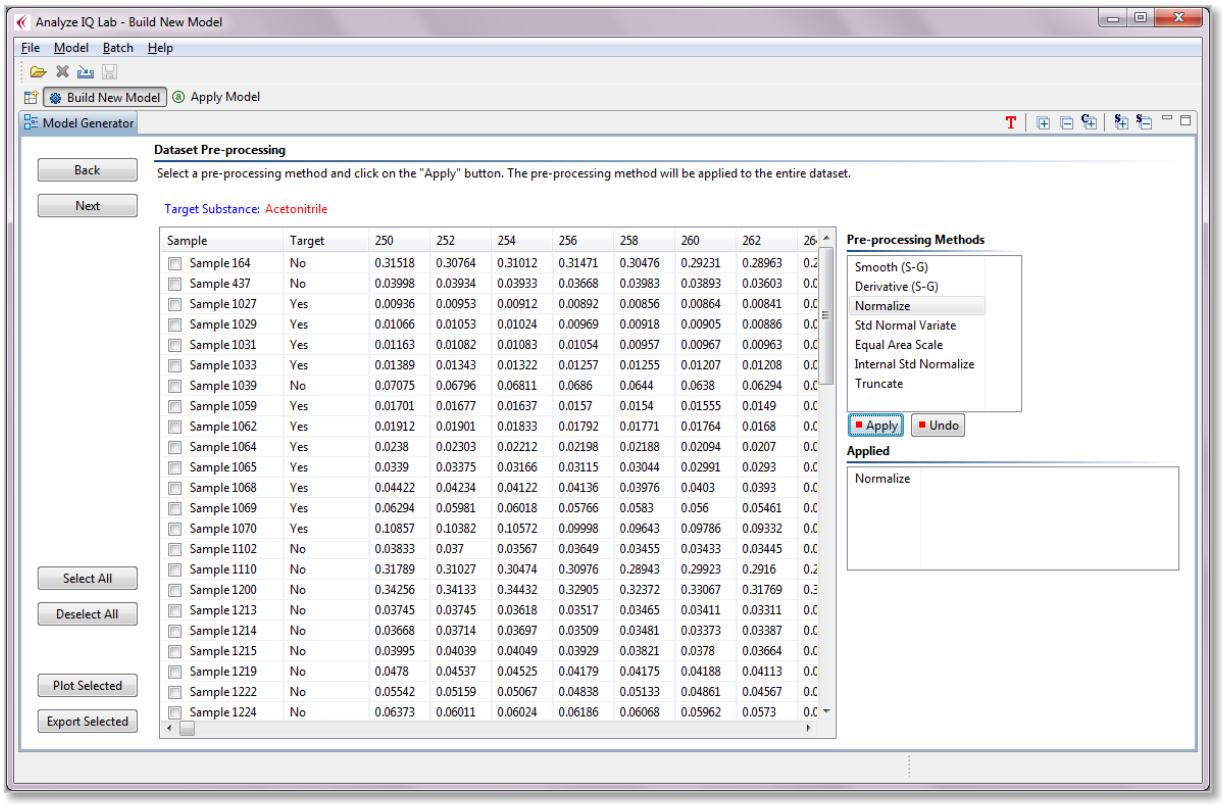


Figure 56: After applying normalization to the dataset (note that numeric values have been scaled)

Any applied pre-processing step may be undone by clicking on the *Undo* button, as shown in Figure 57. The most recently applied step is undone first.

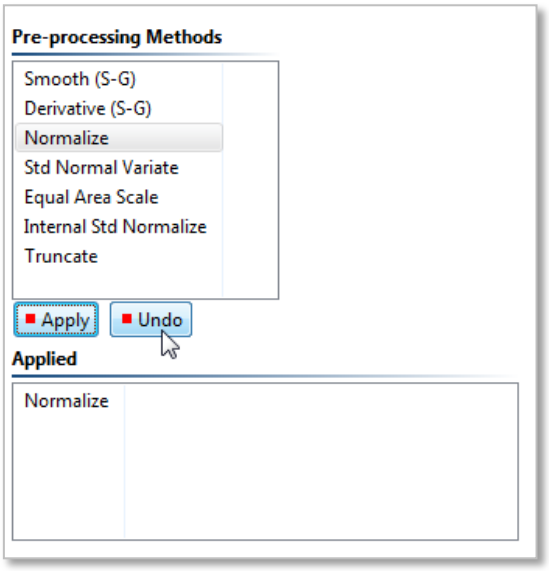


Figure 57: Undoing the Normalization pre-processing

When the pre-processing has been completed, click on the *Next* button to move the selection of classification or quantification method, described in Section 5.5. Alternatively, before proceeding to the next step, you can plot or export spectra, as described next.

5.4 Plotting and Exporting Spectra

5.4.1 Selecting Spectra for Plotting or Exporting

When you are at the Data Pre-processing page, you can select spectra and plot them and/or export their data.

To select spectra, use the check-boxes beside their sample names, as shown in Figure 58. You can also use the *Select All* and *Deselect All* buttons.

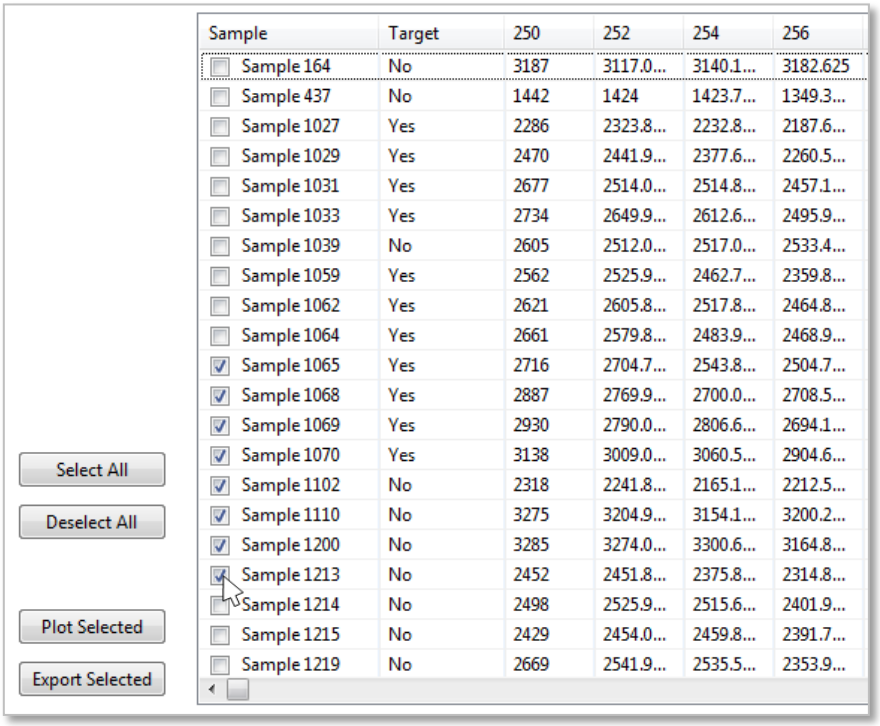


Figure 58: Selecting spectra for plotting/exporting

After selecting spectra, you can use the plot or export them, as described next.

5.4.2 Plotting Spectra

The *Plot Selected* button opens the window shown below in Figure 59. The spectra are plotted with the currently-selected pre-processing options. Note that you can resize this window to make it bigger if you wish, by dragging its corners.

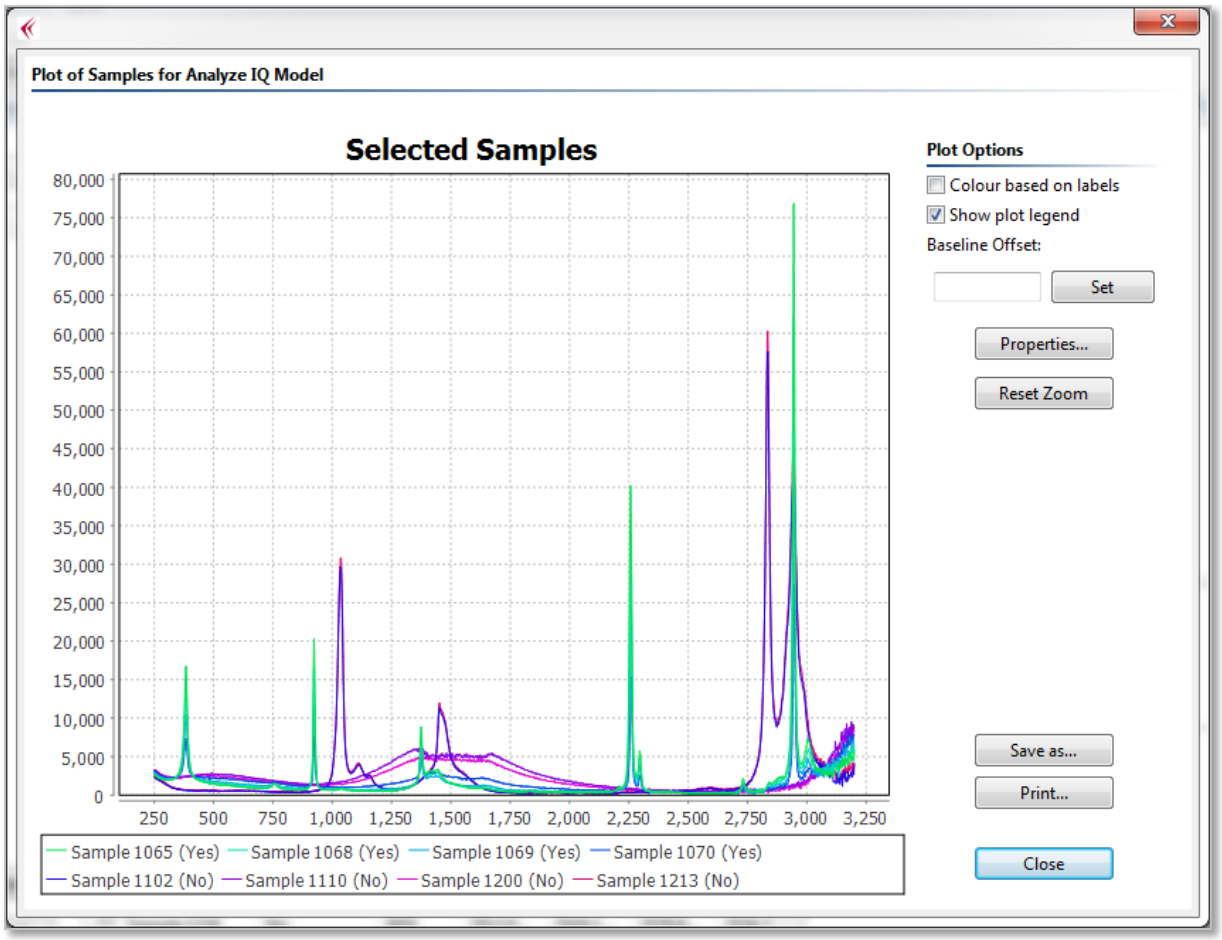


Figure 59: Plot of samples selected from a dataset

As the figure shows, there are several controls beside the plot to change its appearance:

- *Color based on labels:* For a qualitative model, this colors all spectra labeled 'Yes' in blue and all spectra labeled 'No' in red. For a quantitative model, spectra are assigned a color along the blue/red spectrum, with values close to 0 being colored blue and those close to 100 being colored red. An example is shown in Figure 60.
- *Show plot legend:* This allows you to turn on/off display of the legend below the plot.
- *Baseline offset:* This allows you to specify a baseline offset for spectra in the plot, so that different spectra can be seen more easily. In Figure 60, an offset of 5000 has been applied. Naturally, you should choose a baseline offset appropriate to your data.
- *Properties:* This opens a dialog box that allows you to change many properties of the plot display, including changing the title, specifying X and Y axis labels, entering X and Y ranges, and even changing fonts. These properties are spread over 3 tabs, as shown in Figure 61.
- *Reset Zoom:* You can zoom in on an area of the plot by clicking and dragging on it. The *Reset Zoom* button returns to the original zoom setting.

- *Save As:* This allows you to save the graph as an image in PNG format.
- *Print:* This allows you to print the graph.

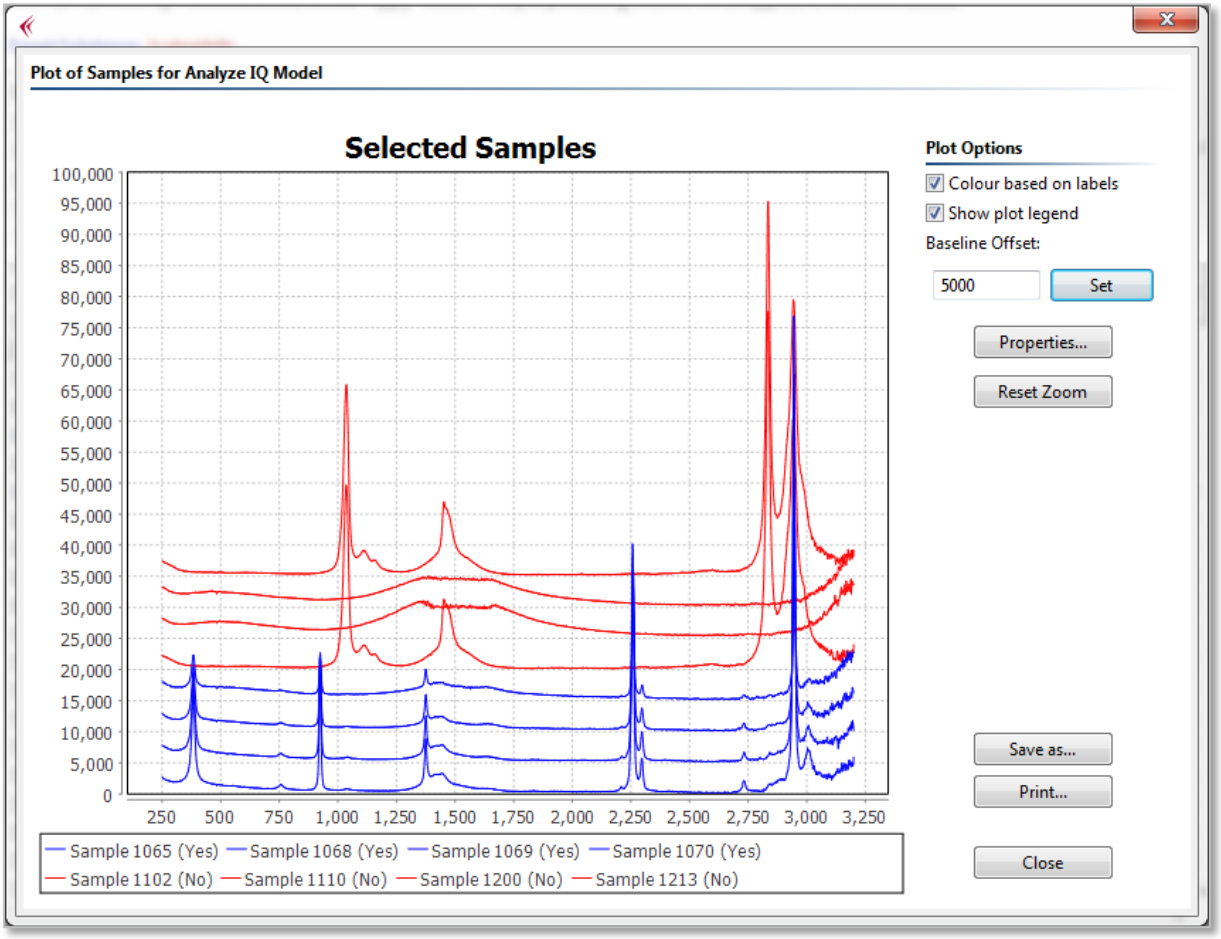


Figure 60: Plot of spectra with 'Colour based on labels' option and a baseline offset

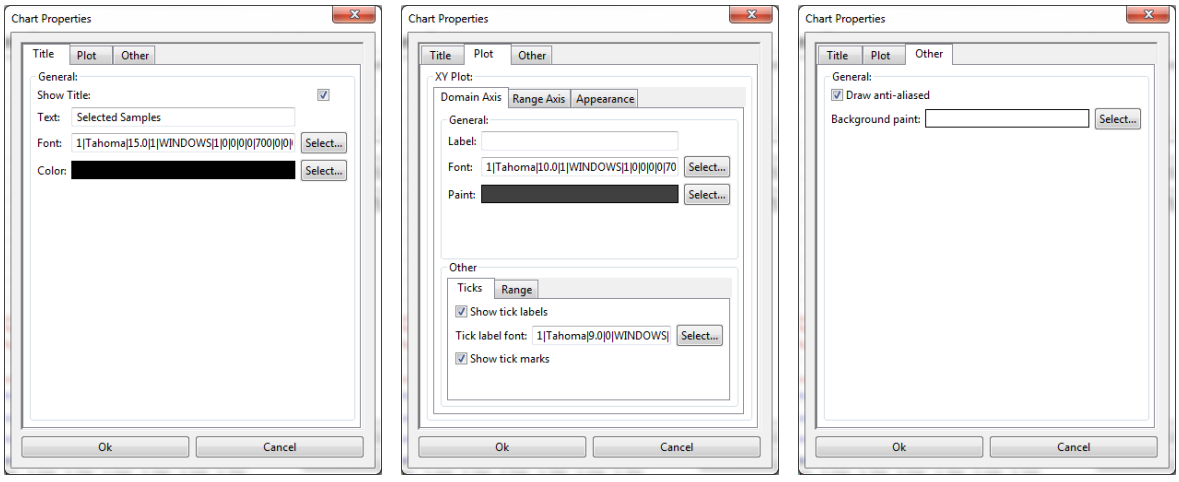


Figure 61: Plot Properties are organized into 3 tabs

5.4.3 Exporting Spectra

The *Export Selected* button prompts you for the name of a CSV (comma-separated values) to which the spectral data are written. The spectral data are stored with the currently-selected pre-processing options.

The CSV files exported here can be opened up in spreadsheet packages such as OpenOffice Calc and Microsoft Excel. They can also be used to build future Analyze IQ models, using the *Import from Single File* option (refer back to Section 5.2.2).

5.5 Select Classification/Quantification Method

After clicking *Next* in the Dataset Pre-processing page, the set of available classification or quantification methods, along with their corresponding parameter settings are displayed; Figure 62 shows the list of methods available for building classification models. Note that the decision as to whether classification or quantification model is being built is determined in the dataset preparation step (see Section 5.2).

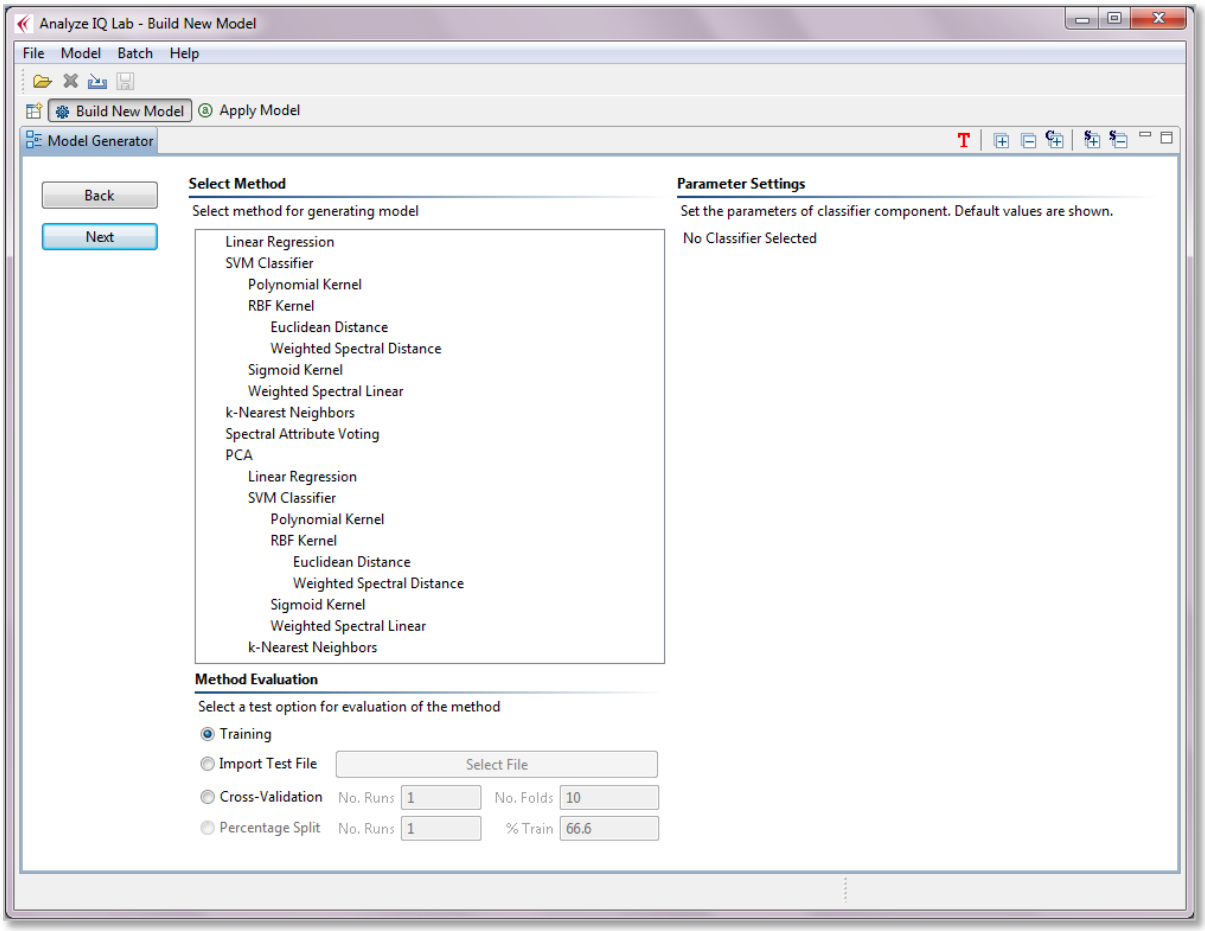


Figure 62: Select classification method

Some methods can have different variants selected, e.g. the kernel of the SVM Classifier or the base classifier of the PCA method. To select variants of a method, choose an entry from the sub-tree of the classifier. In the example shown in Figure 63, a SVM Classifier with Weighted Spectral Linear kernel has been selected.

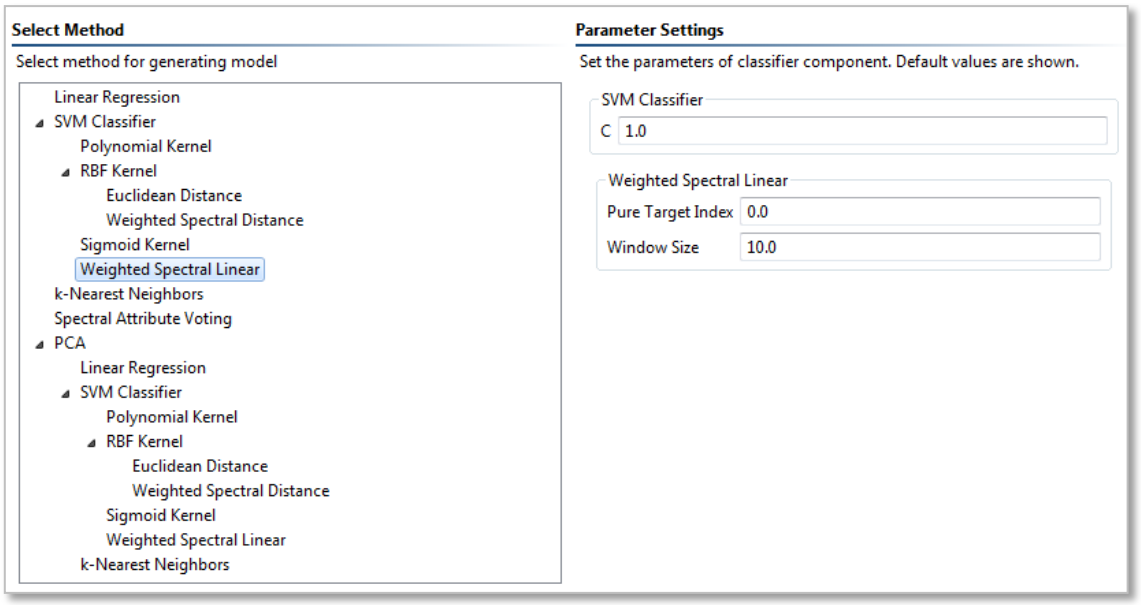


Figure 63: Selecting SVM Classifier with Weighted Spectral kernel

Note that many classification and quantification methods have parameter settings that can be adjusted by the user. Every combination of parameter settings will result in a different final model. For example, the SVM C parameter and Weighted Spectral Linear window size parameter could be changed in Figure 63. For this example, the default values are being used.

Refer to Section 9 for details of all of the **Analyze IQ Lab** classification and quantification methods and their associated parameters.

5.6 *Evaluate Selected Method*

5.6.1 Evaluation Procedure

After selecting the method that you will use to build the model, you must select one of the following three methods[3] for evaluating the method (see Figure 64):

[3] A fourth method, Percentage Split, it not currently activated.

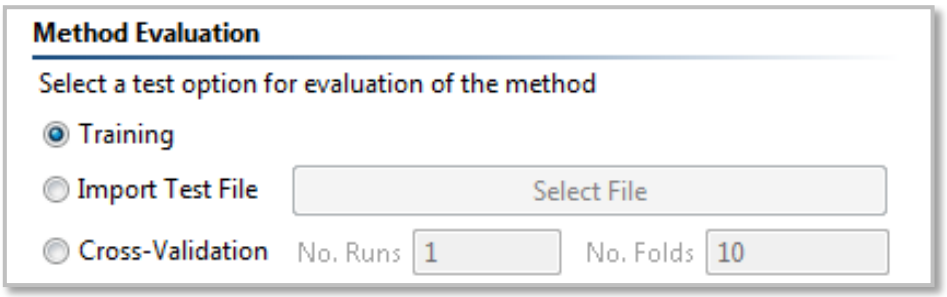


Figure 64: Methods for evaluating chosen classification or quantification method

1. ***Training:*** the entire dataset is used to build a model with the selected method. The resulting model is then tested by using the model to classify or quantify each sample of the training dataset in turn.
2. ***Import Test File:*** the entire dataset is used to build a model with the selected method. The resulting model is then tested by using the model to classify each sample of a separate dataset, selected by the user. For this type of evaluation, select the *Import Test File* option and click on *Select File* button to import the dataset. Note that only files in the *Sample Per Row* format are allowed for this import step. See Section 5.2.2.2 for details on the format of this file. Figure 65 shows the window that is opened after a file been selected for evaluation. As this window indicates, the test dataset is pre-processed using the same pre-processing techniques that have already been applied to the training dataset. The training dataset is the dataset prepared in Section 5.2.1. As can be seen in Figure 65, the functionality for selecting and plotting spectra is available here as it is when preparing the training dataset; refer back to Section 5.4 for details.
3. ***Cross-validation:*** the entire dataset is divided into *k* segments or folds. A model is built using data from folds 2 to *k* and this model is tested on the data from the first fold. Then the second fold is left out of the training (which is now based on folds 1 and 3 to *k*) and used as the test dataset. This process is repeated until all folds have been held out as a test set. In this way, the method is tested on samples of the dataset in a single cross-validation run. The number of folds, *k*, is specified by the user (default number is 10). To carry out a more robust evaluation, a number of cross-validation runs may be used, in which the division of the dataset into folds differs between each run. For example, in Figure 66, a 10-fold cross-validation has been selected and this is to be repeated 5 times (abbreviated as 5x10-fold cross-validation).

Section 5.7.1 has some additional comments on the implications of choosing between these evaluation procedures.

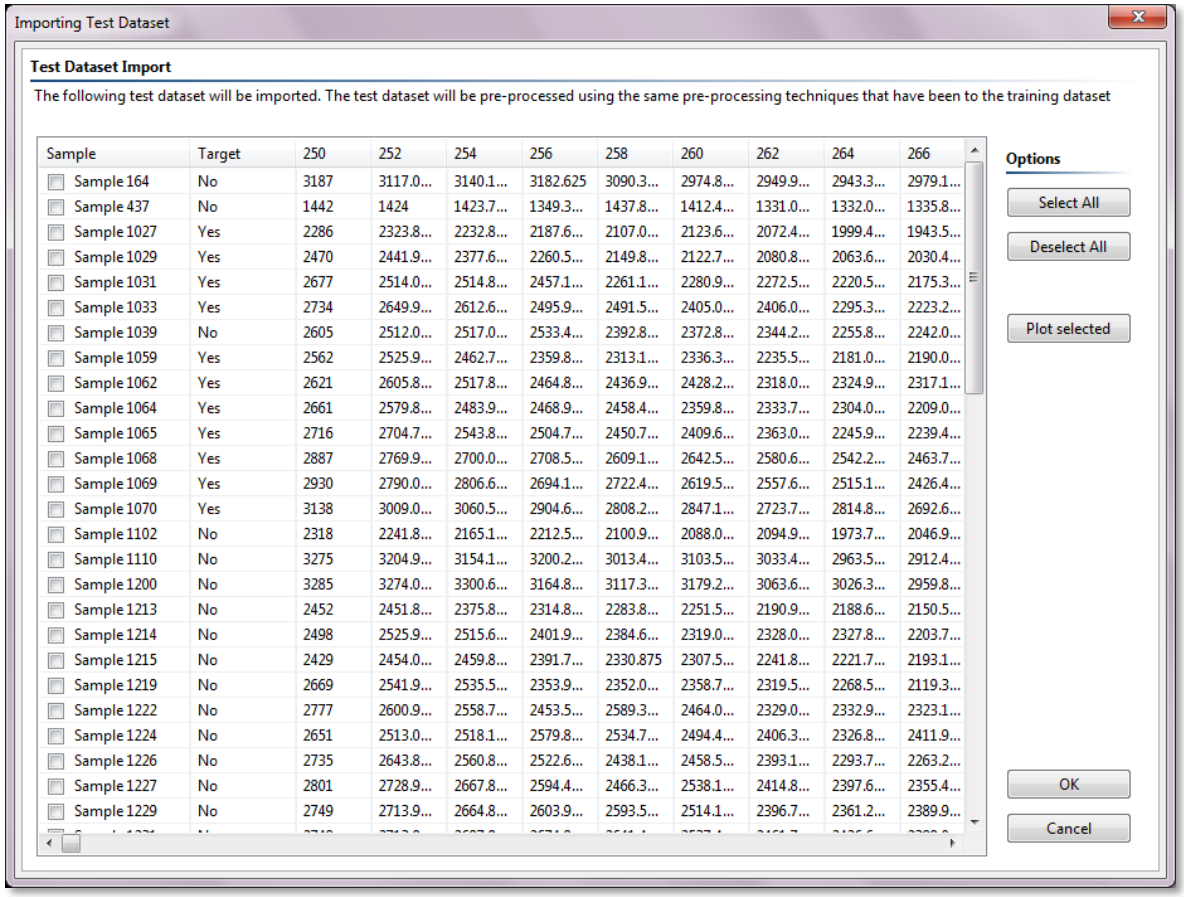


Figure 65: Import Test File evaluation method

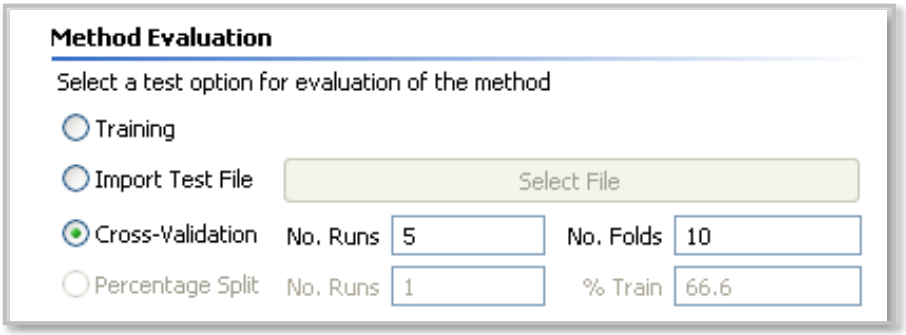


Figure 66: 5x10-fold cross-validation evaluation

After selecting the classification or quantification method and the method of evaluation, click on the *Next* button to run the evaluation.

The results of the evaluation are then displayed. Figure 67 shows the results of evaluating the SVM classifier with Weighted Spectral Linear Kernel using a single 10-fold cross-validation test. In the classification setting, the actual class of each sample (whether it contains the target or not) is shown alongside the class predicted by the method under evaluation. A summary of the results is also shown at the top of the page, including details of the method under evaluation, the number of samples in the

training dataset and the accuracy achieved by the evaluated method. In Figure 67, the classifier achieved an error rate of 0% (i.e. 100% accuracy).

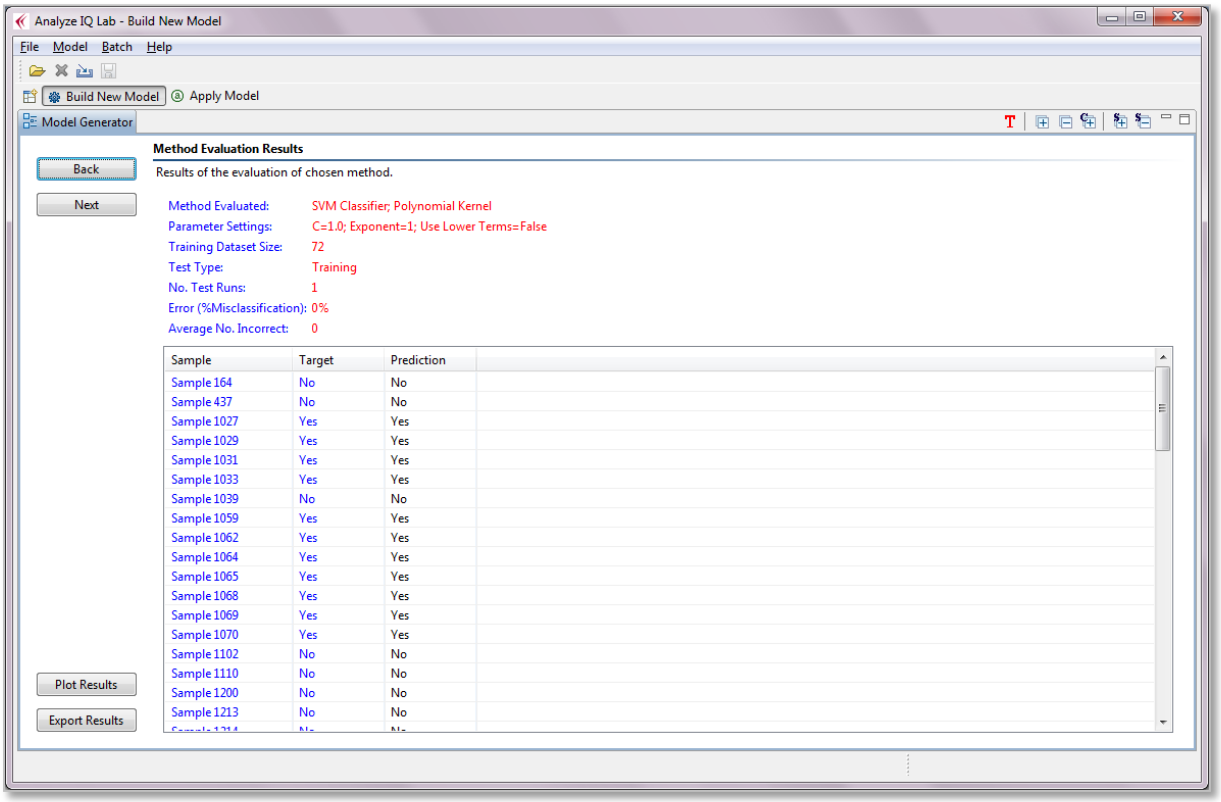


Figure 67: Method evaluation results window

5.6.2 Plot Classification Model Evaluation Results

Pressing the *Plot Results* button opens the window shown in Figure 68. For a classification model, the results are summarized in a bar chart with 4 segments. As indicated in the figure, they are counts of the following categories of results:

- True Positives: samples that are predicted to have the target present (positive), and this prediction is *true*
- False Positives: samples that are predicted to have the target present (positive), and this prediction is *false*
- True Negatives: samples that are predicted *not* to have the target present (negative), and this prediction is *true*
- False Negatives: samples that are predicted *not* to have the target present (negative), and this prediction is *false*

In the example shown, since there are no errors, the bars for False Positives and False Negatives have heights of 0.

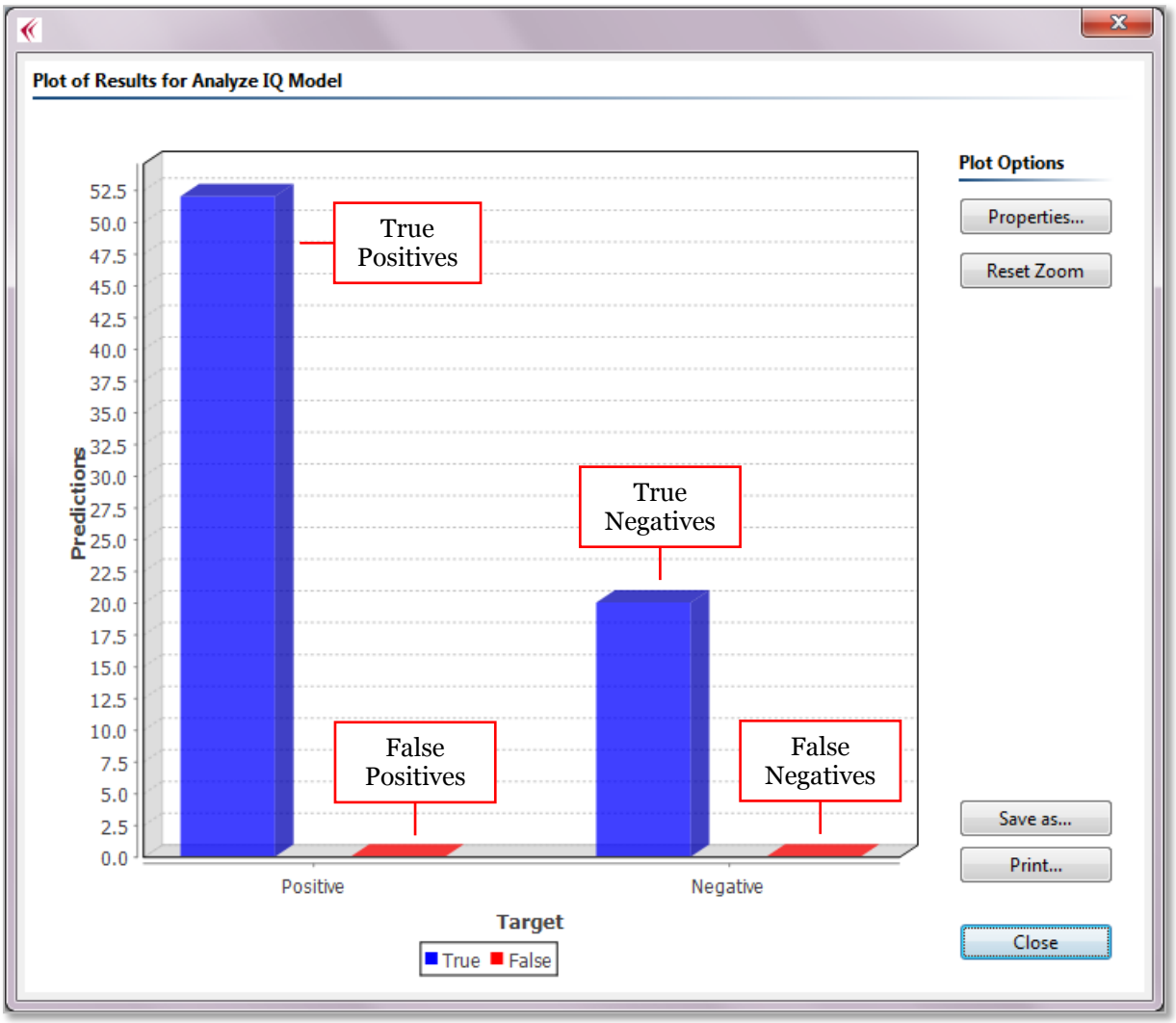


Figure 68: Plot summarizing the results of evaluating a model

This plot window has several options:

- *Properties*: Configure various properties of the plot
- *Reset Zoom*: If you have zoom in on an area, reset to the default view
- *Save As*: Save the graph as an image in PNG format
- *Print*: Print the plot to a printer
- *Close*: Close the window

Refer back to Section 5.4.2 for further explanation of these controls.

As will be discussed in Section 6.2, the results of quantitative analyses are presented differently from this.

5.6.3 Export Model Evaluation Results

As indicated in Figure 69, the *Model Evaluation Results* page has a button labeled *Export Results*.

When you press this, you are prompted to save the results in a CSV file, which is suitable for opening subsequently in Microsoft Excel, OpenOffice Calc, Matlab, or other programs. All of the data from the *Model Evaluation Results* page is written to this file.

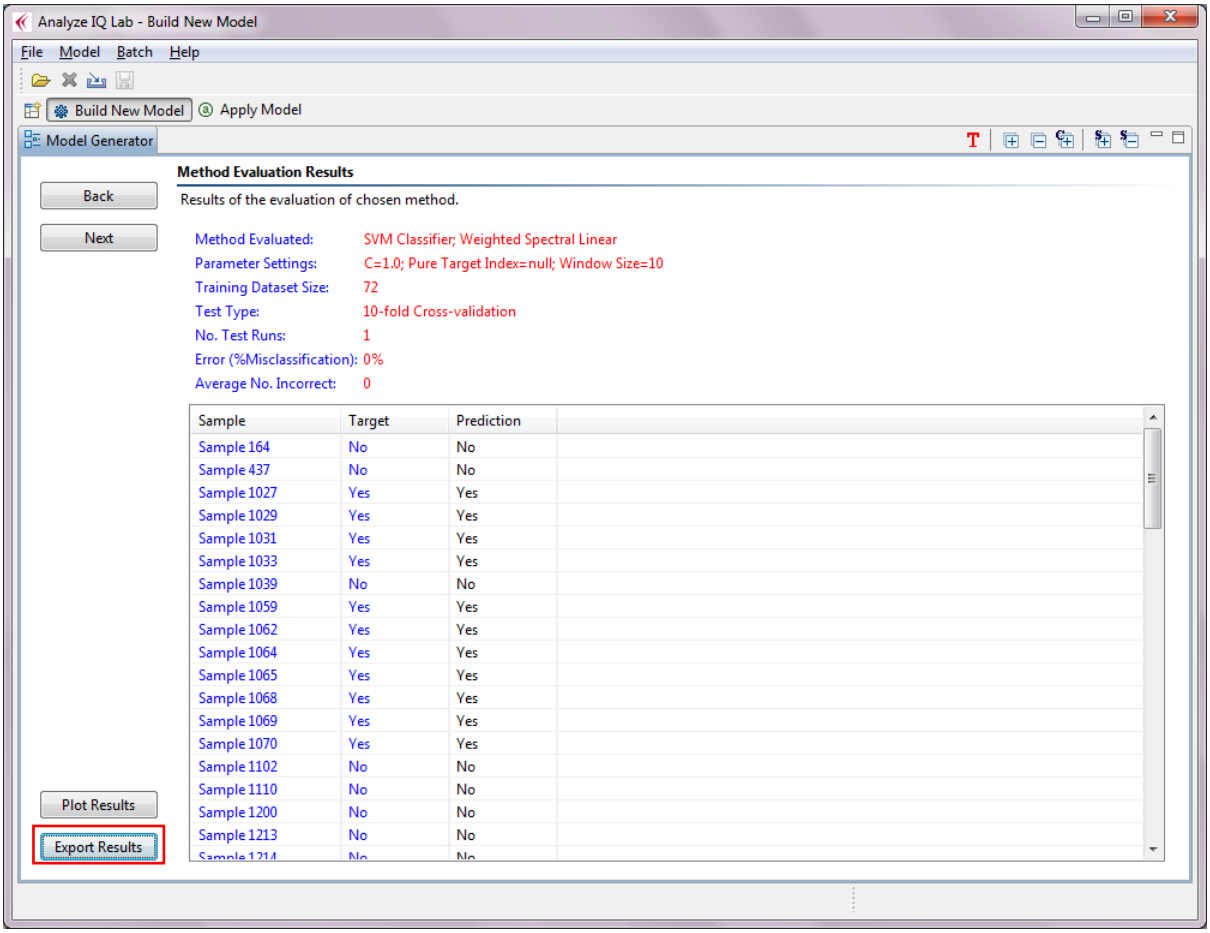


Figure 69: Method evaluation results window, indicating *Export Results* button

5.7 *Understanding and Reducing Classification Error Rates*

5.7.1 Choosing an Evaluation Procedure

The purpose of a classification model is, of course, to determine whether a target substance is present in a given sample. As was described above, one of the steps in building a new model is to evaluate it on a set of test data, where we know the correct

answers. By measuring the model's performance on test data, we gain insight into how reliably we can expect the model to perform in the future, when it will be used to determine whether the target is present in substances for which the correct answer is unknown.

Just like the dataset that you use to build a model (see Section 5), the test dataset must consist of spectra for which you know in advance whether or not they contain the target substance.

In fact, you can evaluate the model simply using the training dataset (see Section 5.6.1), though this is not sufficient on its own, as it will produce an overly optimistic measure of model performance, and cannot identify models that are overfitted to the data. However, it can be useful as a first step in trying out a model-building method: if the method performs badly on the training dataset, it will perform even worse when evaluated more rigorously.

The preferred alternative is to use repeated cross-validation (also described in Section 5.6.1), to get an unbiased estimate of model performance, when building and refining models, and to perform a final evaluation on a holdout set, which you can specify with the *Import Test File* option that is also described in Section 5.6.1.

5.7.2 Understanding Results of Evaluation

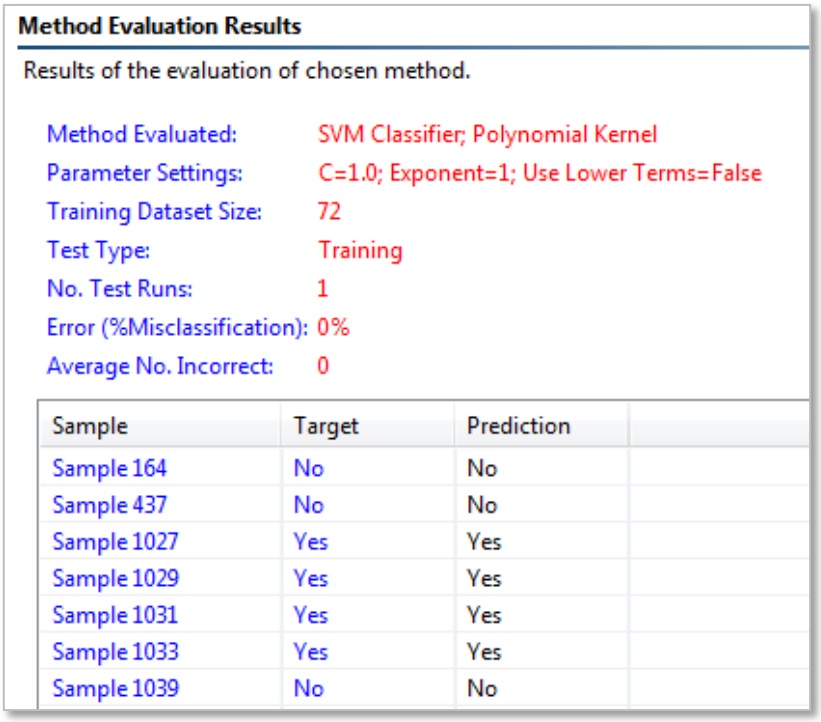
The evaluation procedure is as follows:

- For the purposes of the test, the correct answers (i.e. whether or not the target substance is present) are withheld, and all samples in the test dataset are presented to the newly-built model as if they were unknown.
- The model reports as to whether or not it has detected the target in each test sample.
- It is an error if the model detects that a target is present when it is not (false positive), or if it does not detect that a target is present when it is (false negative).
- The number of errors is counted. If repeated cross-validation runs are performed, the average number of errors over all runs is computed.
- The percentage of errors is computed from the number of errors divided by the number of test samples.

Figure 70 shows a close-up view of a section of the evaluation report from Figure 67. On the lower part of the figure, there is a list of test spectra. Under the heading of 'Target', the correct answers are listed: 'Yes' means that the target is actually present in a sample and 'No' means it is not actually present. Similarly, under the heading of 'Prediction', the model's answers about whether it detected the presence of the target are listed. In the case shown, all of the model's answers are correct, which is why 0% error rate is reported: "Error (% Misclassification): 0%".

Note that the percentage accuracy of a model may be computed from:

Accuracy% = 100% - Error%



Method Evaluation Results

Results of the evaluation of chosen method.

Method Evaluated:	SVM Classifier; Polynomial Kernel
Parameter Settings:	C=1.0; Exponent=1; Use Lower Terms=False
Training Dataset Size:	72
Test Type:	Training
No. Test Runs:	1
Error (%Misclassification):	0%
Average No. Incorrect:	0

Sample	Target	Prediction
Sample 164	No	No
Sample 437	No	No
Sample 1027	Yes	Yes
Sample 1029	Yes	Yes
Sample 1031	Yes	Yes
Sample 1033	Yes	Yes
Sample 1039	No	No

Figure 70: Close-up of evaluation results

5.7.3 Reducing Evaluation Error Rates

In the example shown in Figure 70, the error rate is 0%, and so cannot be reduced. However, it is possible that more stringent evaluation, using repeated cross-validation, would yield a non-zero error rate.

There are several strategies for generating improved models with lower error rates:

1. *Improve the training data set.*
   This is the most important strategy: since models are built through analysis of the training dataset, a model can only be as good as the dataset with which it was built. If you require a model that works well at a large variety of concentrations and with a large variety of diluents or contaminants, then you should include appropriate samples of these, so that the model will be constructed to deal with them. Conversely, if your dataset contains data that is labeled inaccurately or collected sloppily, then it will be difficult to generate a high-quality model from it.
2. *Use appropriate pre-processing on the data set.*
   This is another way of improving the dataset. For example, smoothing can help to reduce the effects of random noise, normalization can help to improve consistency if samples were collected with different exposures, and first-order derivatives can help to reduce the effect of baseline offsets.

3. *Select a different model building method.*
   Analyze IQ Lab provides a range of different model building methods. No one method is best in all situations, although the SVM Classifier with Weighted Spectral Kernel (see Section 9.1.1) and the Spectral Attribute Voting (see Section 9.1.4) are specifically designed for the task of spectral data analysis, and are therefore the recommended choices.
4. *Adjust the parameters of the chosen model building method.*
   Just as selecting a different model building method can make a difference, adjusting parameters can improve performance.

One final point: it is possible to reduce the reported error by testing on the training set or eliminating problematic spectra from the testing set. Needless to say, these actions may make scores look better, but do not improve the underlying model.

5.8 Model Details

After evaluating the selected classification or quantification method, click on the *Next* button to move to the next step, in which the new model details will be shown, as shown in the example in Figure 71.

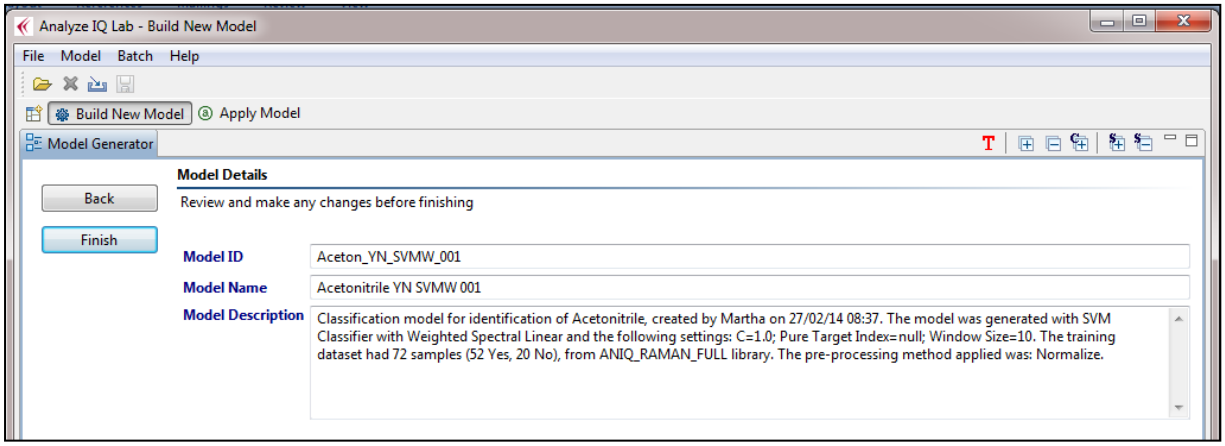


Figure 71: Enter model details

Analyze IQ Lab automatically proposes a model ID, model name and description for you, but you are free to edit these details as you wish. The details are:

- ***Model ID:*** This is the unique identifier for the new model. If a model already exists with the same Model ID, you will be prompted to enter a different name. Note that the following characters are allowed in the Model ID: 'A'-'Z', 'a'-'z', '0'-'9', '_' and '.' For example "Cocaine_Model1" is valid Model ID, where as "Cocaine Model1" is not, as it contains a space. In addition "WSLinearSVM_Acetonitrile" is not valid, as a model with this ID already exists. In the example in Figure 71, we have gone with the default Model ID, "Aceton_YN_SVMW_001".

- ***Model Name:*** The title of the new model. Note that, unlike the model ID, this name does not have to be unique and it may contain spaces.
- ***Description:*** A short description of the new model. The one automatically generated by **Analyze IQ Lab** includes the target substance, the method used the build the model, the parameter settings, the size of the training data set, and the pre-processing methods applied. Again, however, you are free to change this.

5.9 Generate Final Model

After entering the model details, click on the *Finish* button to generate the model. After the model is generated, the message shown in Figure 72 is displayed. Clicking on *Yes* switches the application to the *Apply Model* page.

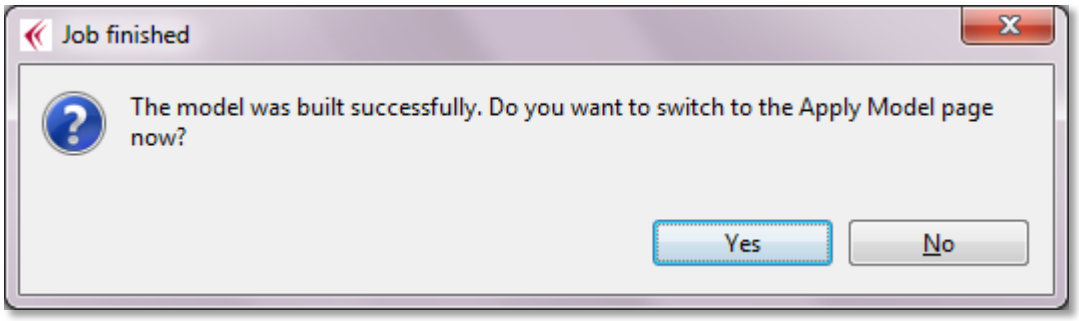


Figure 72: Model generation complete

When you switch to the *Apply Model* page, a new tab is added to the bottom centre window of the page, showing the new model (see Figure 73). Refer back to Section 4.2 for details on how to use this or other models for the analysis of spectra.

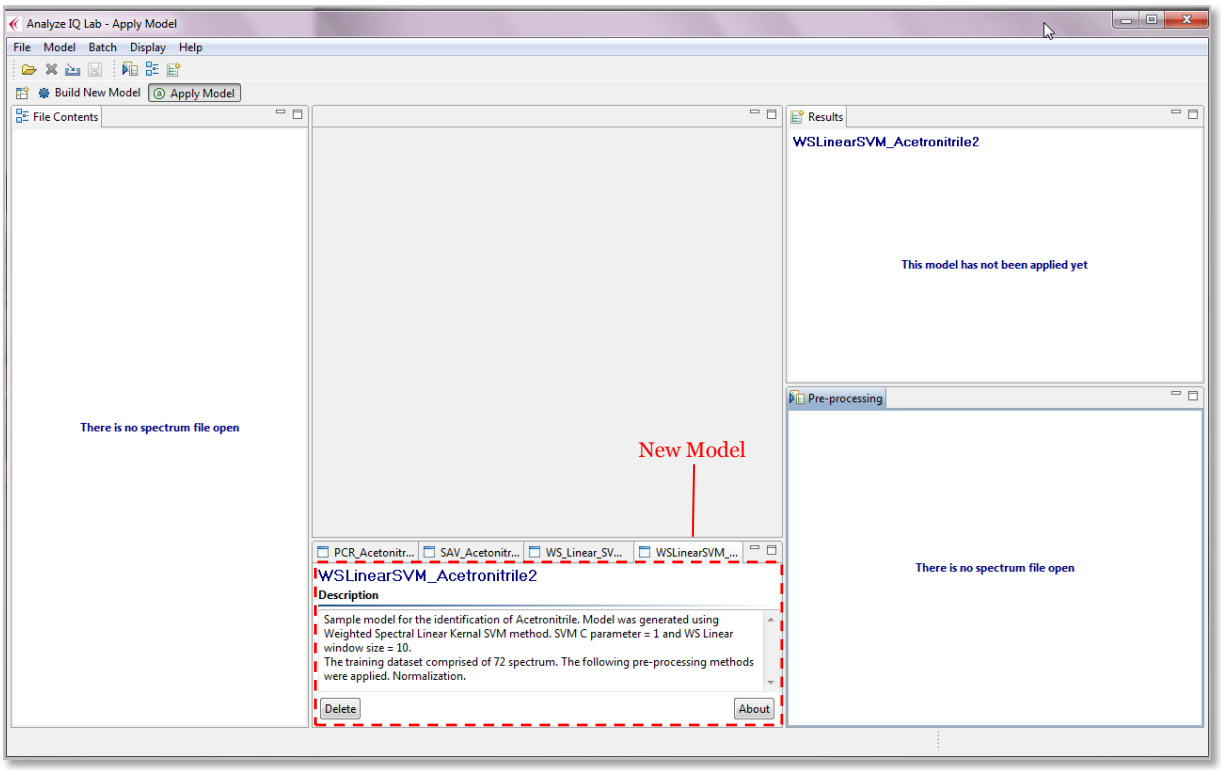


Figure 73: New model added to the *Apply Model* page

5.10 Deleting a Model

It is possible to delete that you have previously built, for example if it is no longer needed or if you are replacing it with a newer one. To do this, simply press the 'Delete' button on the lower left side of the model window which is highlighted in the figure below.

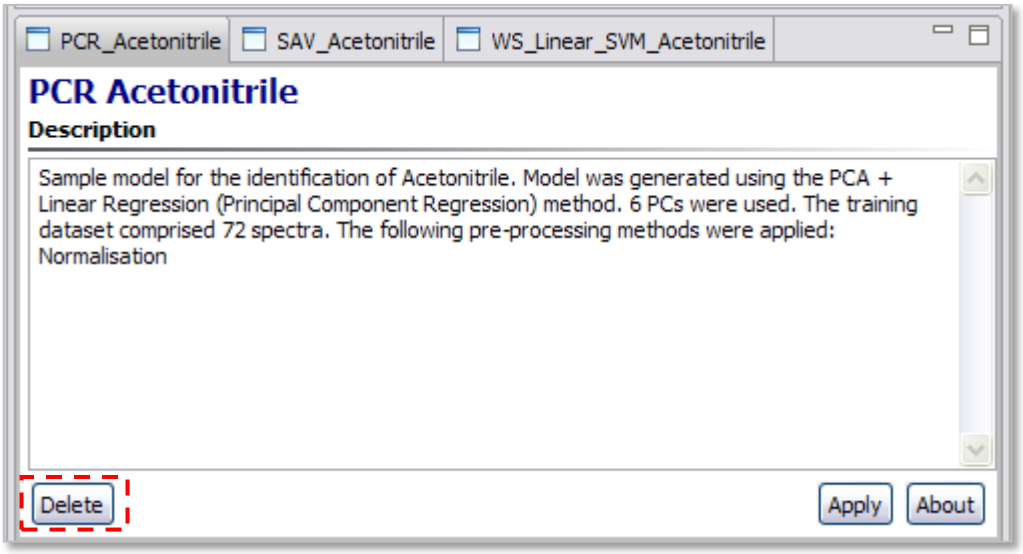


Figure 74: Deleting a model

The *Model Manager*, which will be described later in Section 8, provides you with additional functionality for backing up, restoring, and deleting models.

6 Build and Apply a Quantification Model

The detailed description of the analysis and model building steps in Sections 4 and 5 have used classification examples. The procedure for building and applying models for quantification in **Analyze IQ Lab** is very similar. This section will give a step-by-step example of the generation and application of a quantification model.

6.1 Building a Quantification Model

For this example, the dataset is drawn from ***Spectra Manager***.

1. Switch to the *Build New Model* page (see Figure 75)

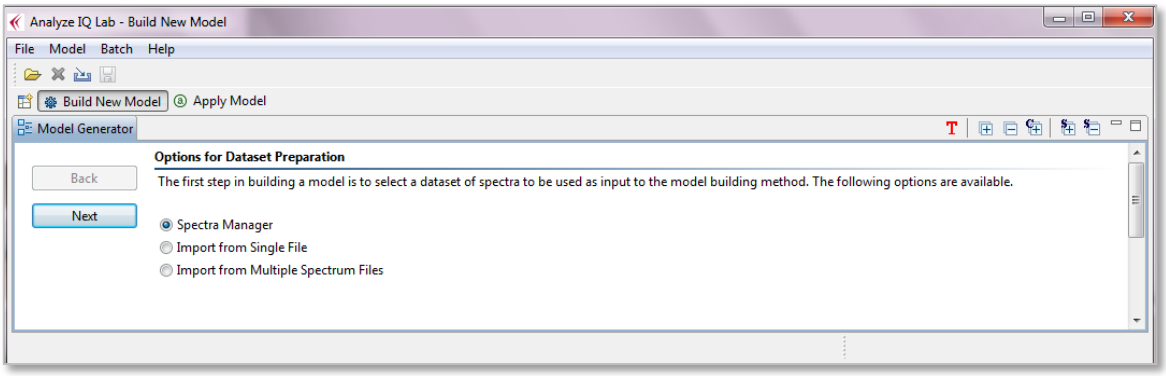


Figure 75: Building a quantification model: switch to Build New Model page

2. Select *Next* and click on the *Connect* button (see Figure 76).

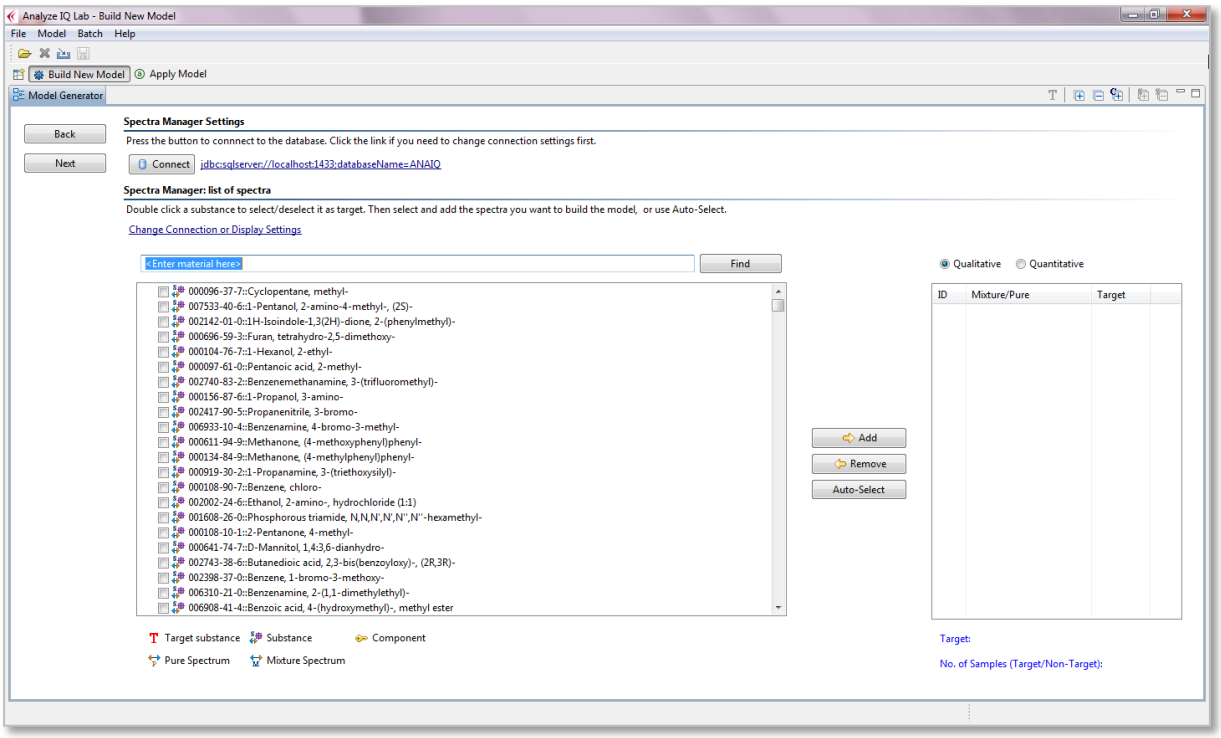


Figure 76: Building a quantification model: connect to database

3. Prepare the Dataset:

    a. Enter target substance (acetonitrile) into the text box beside the *Find* button and click *Find*
    b. Select acetonitrile as the target substance by double-clicking on the acetonitrile substance entry in the database viewer
    c. Select *Quantitative* option in top right-hand area of the page
    d. Click *Auto-Select* to generate a dataset (see Figure 77; note that concentration information for each sample is shown under the *Target* column of the current dataset table, rather than Yes/No classification labels).

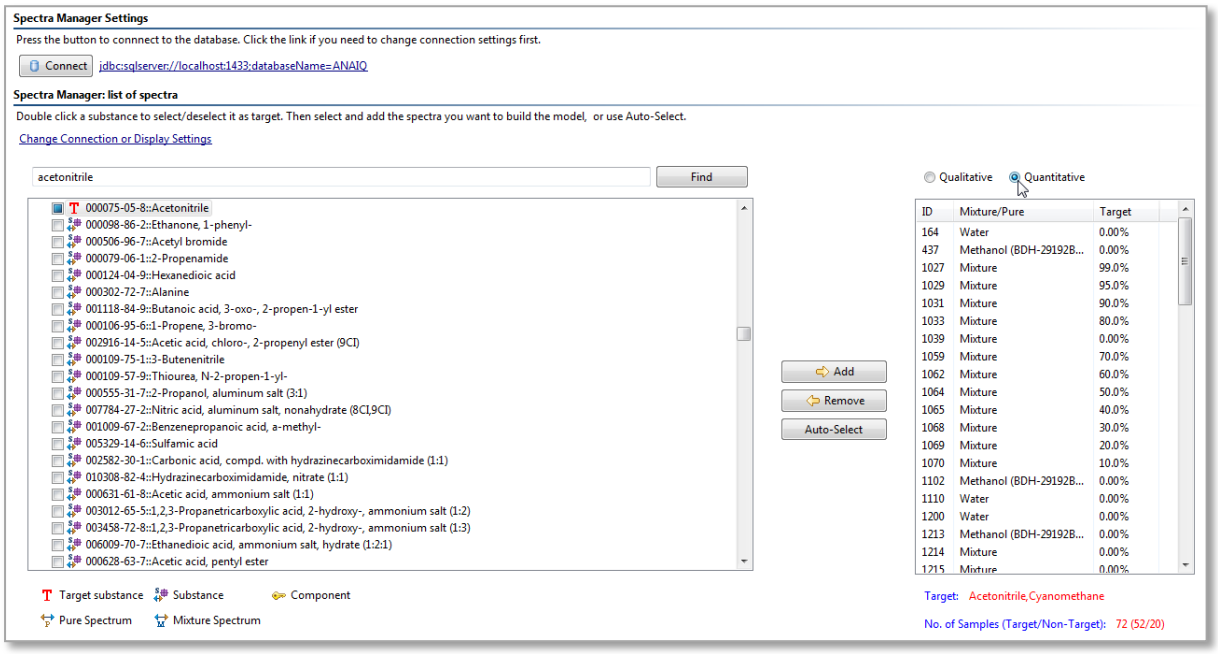


Figure 77: Building a quantification model: prepare dataset

4. Click on *Next* to view the spectral samples of the dataset. As shown in Figure 78, for this example we will apply the Standard Normal Variate transform. (Refer back to Section 4.3.1 for details of the pre-processing methods.)

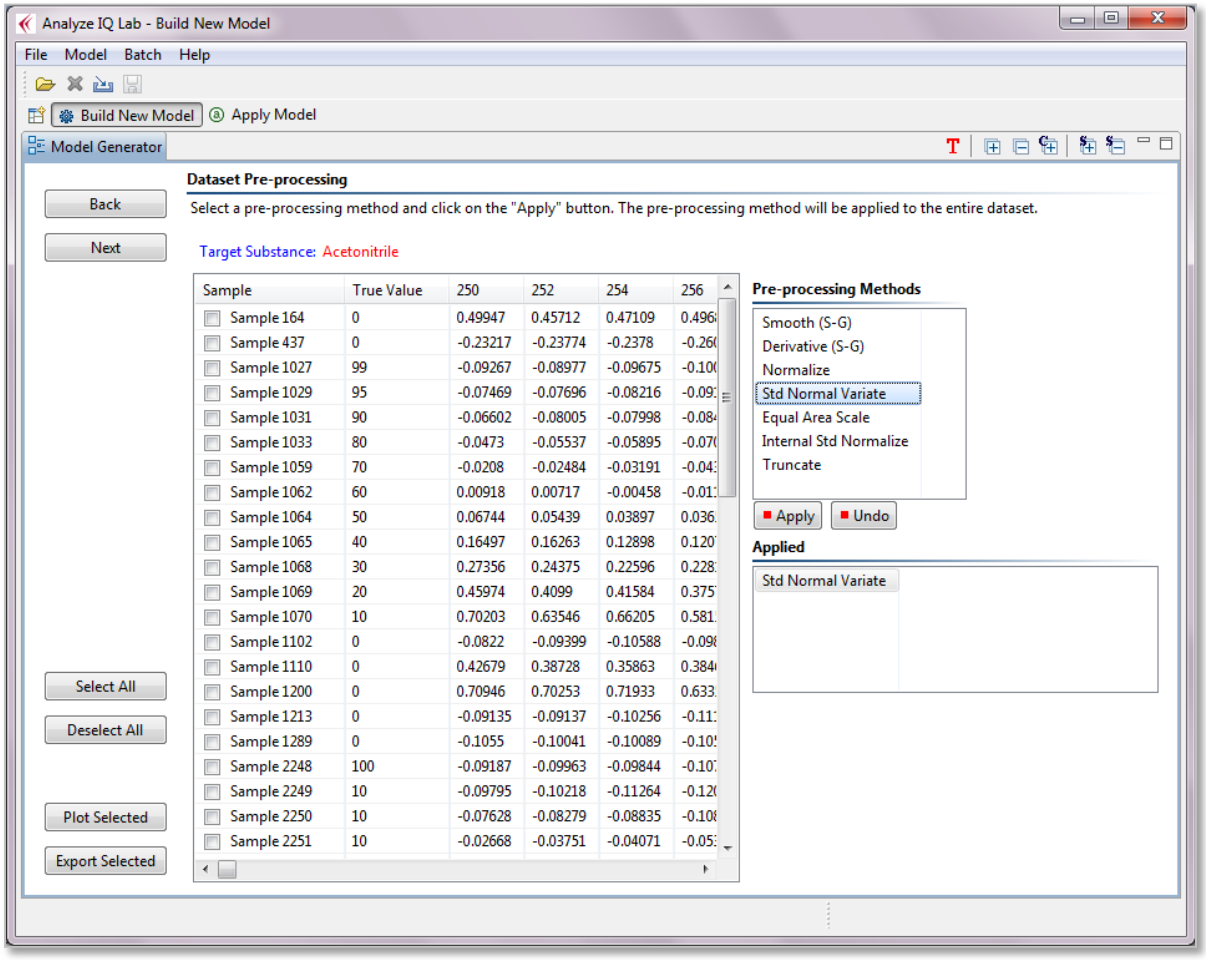


Figure 78: Building a quantification model: dataset pre-processing

5. Method Selection & Evaluation:

    a. Click on *Next* to move to the method selection and evaluation step
    b. Select a quantitative analysis method: here we choose SVM with Polynomial Kernel, as shown in Figure 79
    c. Enter the parameter settings: here we set the parameter C=10, and use a second-order polynomial, i.e. Exponent=2 and we choose to Use Lower Terms (1 means Yes); some preliminary experiments indicated that these settings work reasonably well
    d. Select an evaluation method evaluation method: in Figure 79 we select *Training*, as we are performing a preliminary analysis

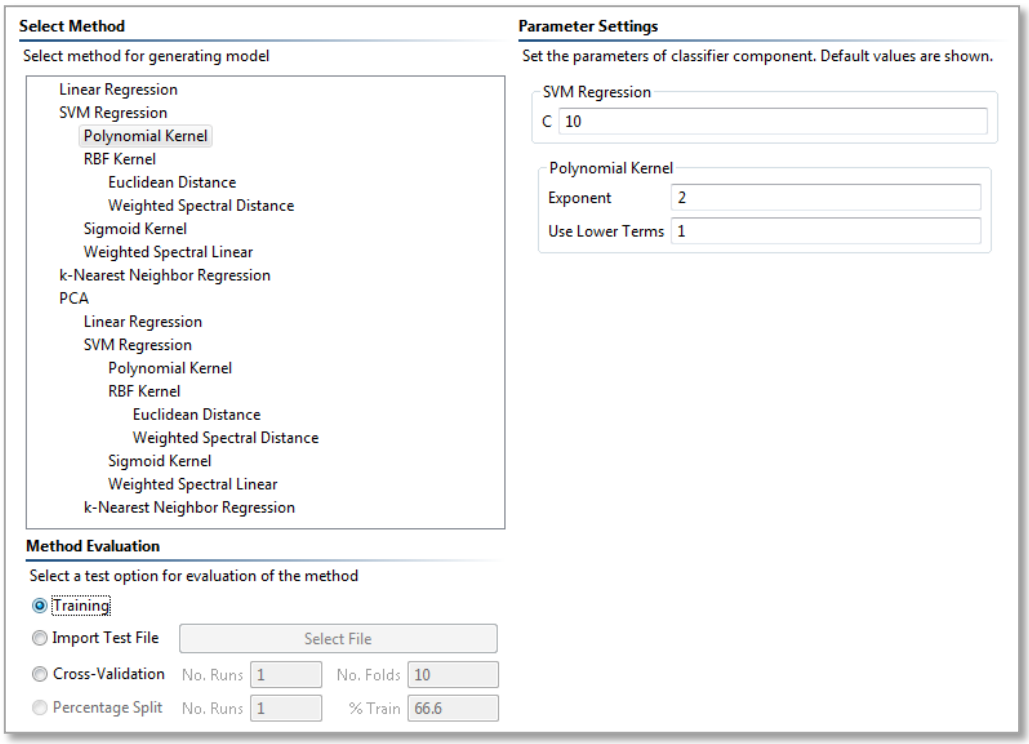


Figure 79: Building a quantification model: selecting method and evaluation

6. Click on *Next* to carry out the evaluation. The results of this evaluation are shown in Figure 80.

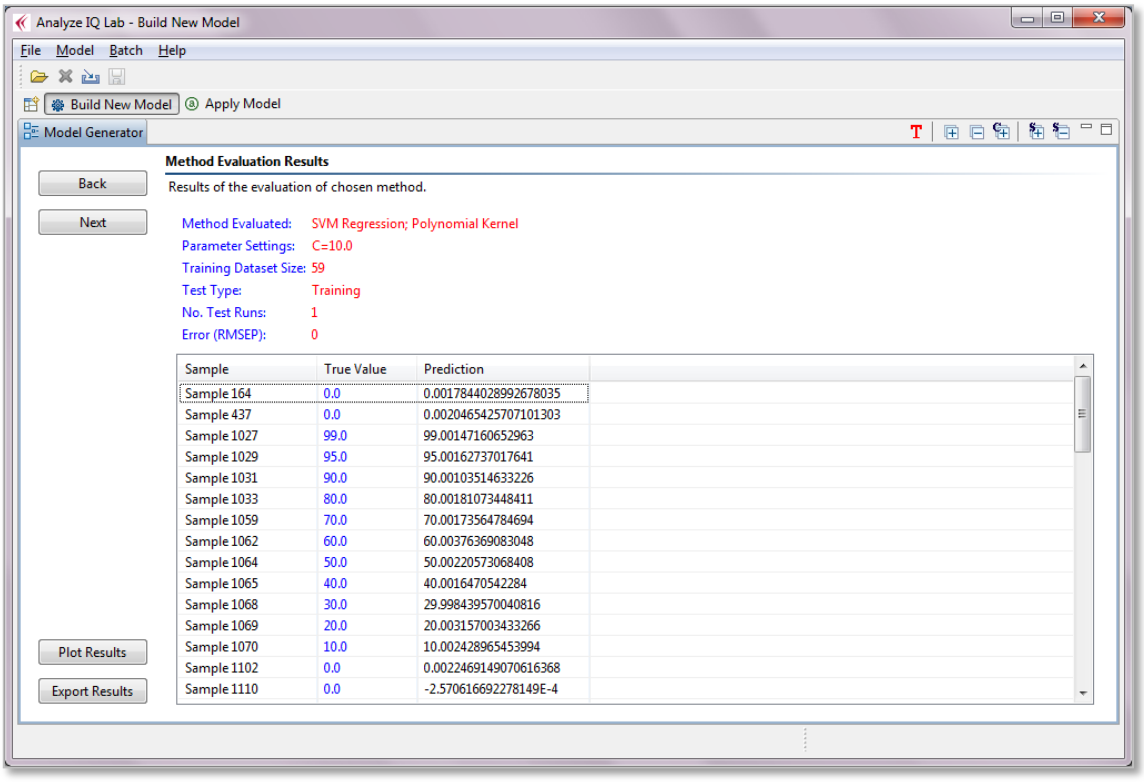


Figure 80: Building a quantification model: results of evaluation

6.2 Plot and Export Quantification Results

6.2.1 Plot Quantitative Model Evaluation Results

On the Model Evaluation Results page, as can be seen in Figure 80, there are buttons labeled *Plot Results* and *Export Results*.

When you press the *Plot Results* button, a plot window as shown in Figure 81 is opened.

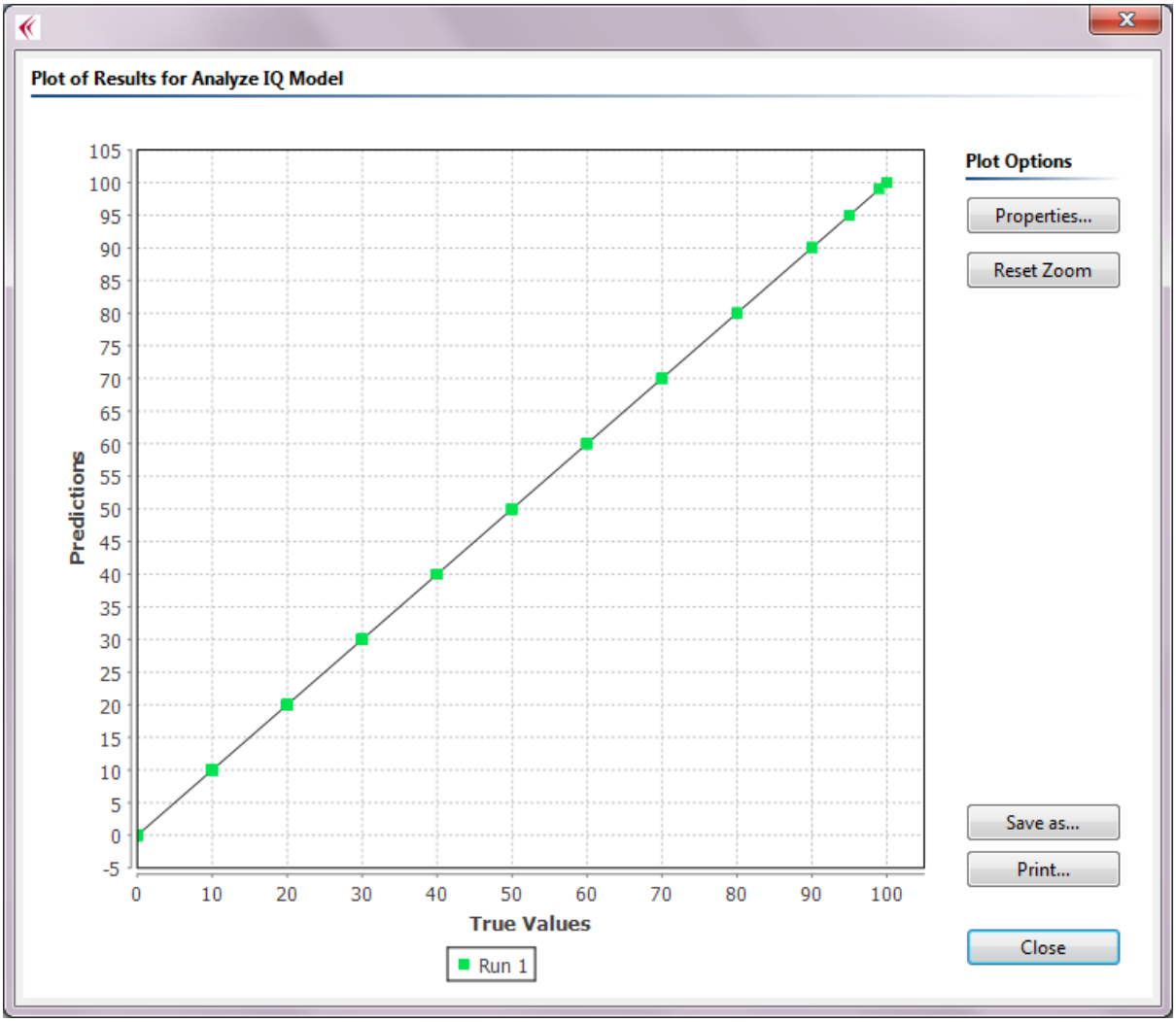


Figure 81: Plot of results of evaluating a quantitative model

This plot provides a visual comparison of true values (plotted on the horizontal axis) for the test data versus values predicted by the model (plotted on the vertical axis). For reference, a black diagonal line corresponding to X=Y is also included in the plot. In the ideal case, the model's predicted values should match the true values, and so all points on the plot should lie on the diagonal.

In the case of the plot shown in Figure 81, there is very good agreement. Looking back at the table of results in Figure 80, it can be seen that this is indeed the case. Of course, these results are from testing the model on its training data, so it is not surprising that they are good. For further validation of the model, it should be tested using repeated cross-validation runs, and/or a hold-out dataset.

As before, this plot window has several options:

- *Properties*: Configure various properties of the plot
- *Reset Zoom*: If you have zoom in on an area, reset to the default view
- *Save As*: Save the graph as an image in PNG format
- *Print*: Print the plot to a printer
- *Close*: Close the window

Refer back to Section 5.4.2 for further explanation of these controls.

6.2.2 Export Model Evaluation Results

As before, the *Model Evaluation Results* page has a button labeled *Export Results*.

When you press this, you are prompted to save the results in a CSV file, which is suitable for opening subsequently in Microsoft Excel, OpenOffice Calc, Matlab, or other programs. All of the data from the *Model Evaluation Results* page is written to this file.

6.3 Completing the Model

When you are satisfied with the model that you have built, click *Next* and enter new model details (see Figure 82).

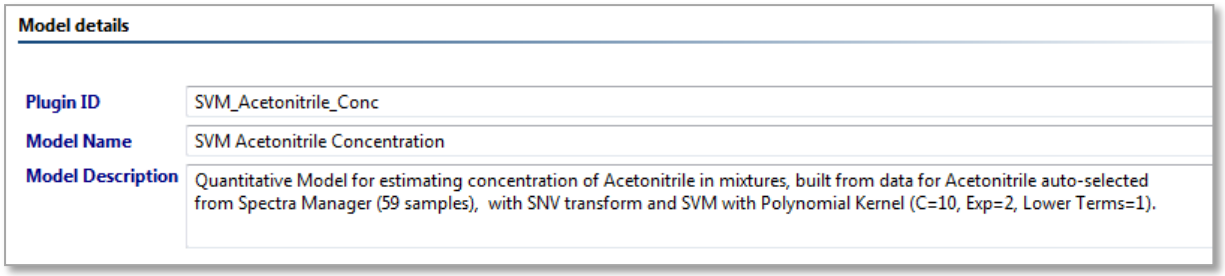


Figure 82: Building a quantification model: enter model details

7. Click *Next* and switch to the *Apply Model* page t (see Figure 83).

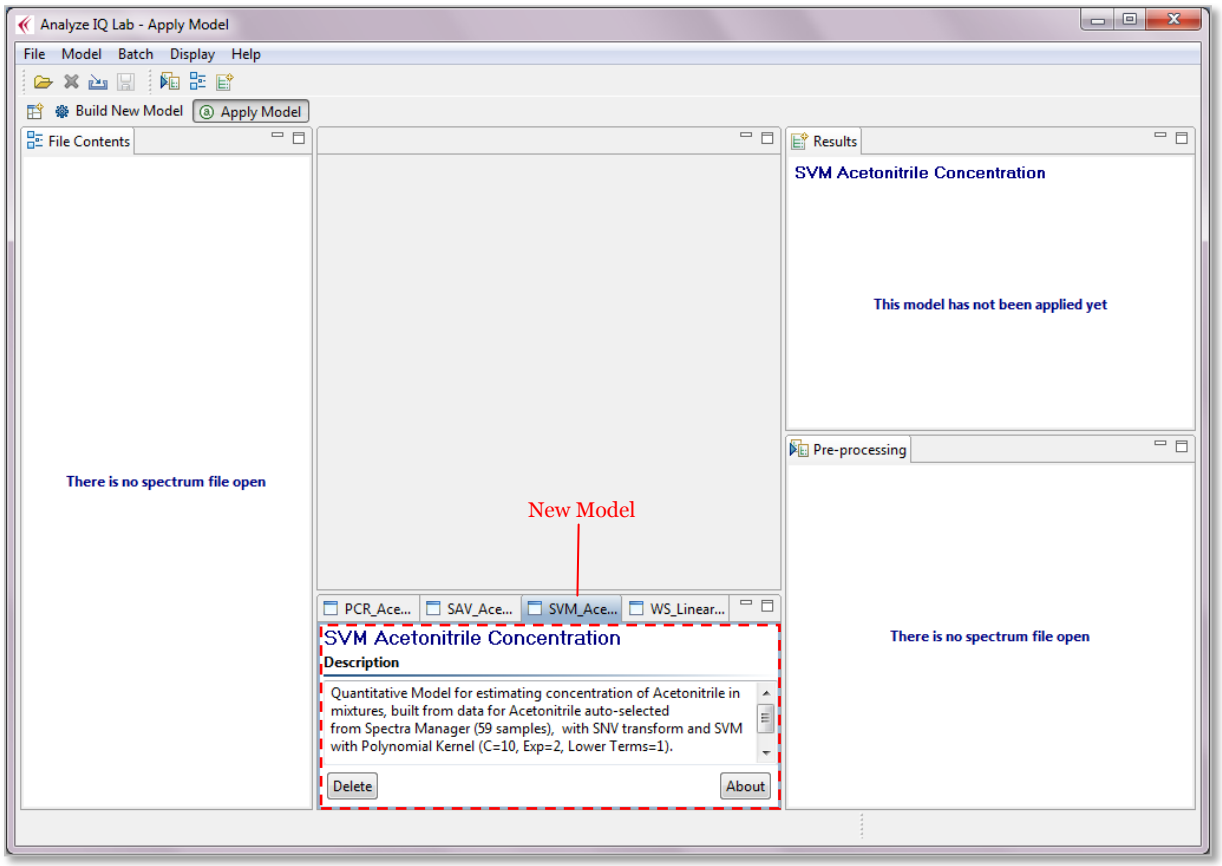


Figure 83: New quantification model added to Apply Model page

6.4 Applying a Quantification Model

Open a spectrum file and click the *Apply* button of the new quantification model (see Figure 84). Figure 84 shows that the new model predicted a concentration of 20.04% acetonitrile, where the sample actually contained 20% Acetonitrile (based on the file name). Given the data set on which this model was constructed, this estimate is good.

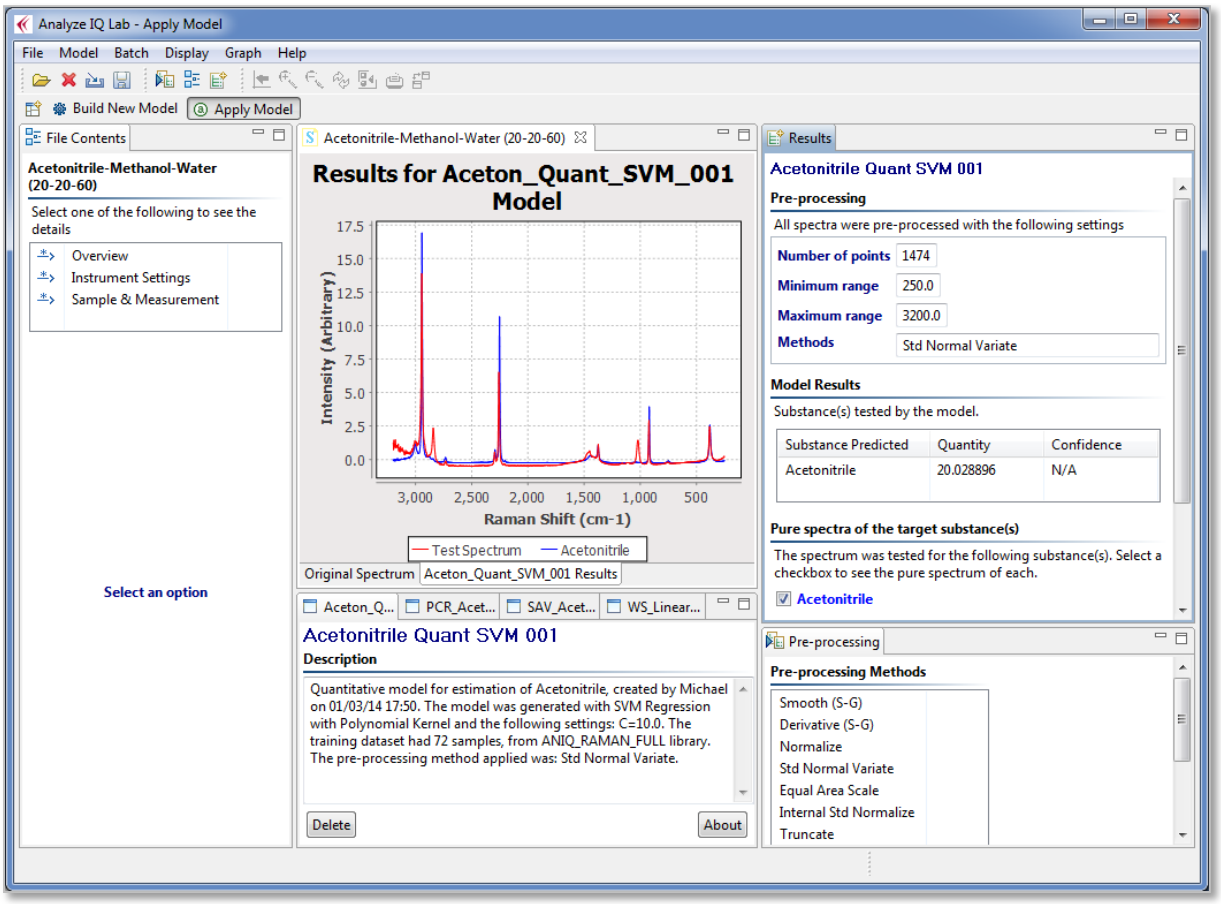


Figure 84: Applying a new quantification model to a sample spectrum

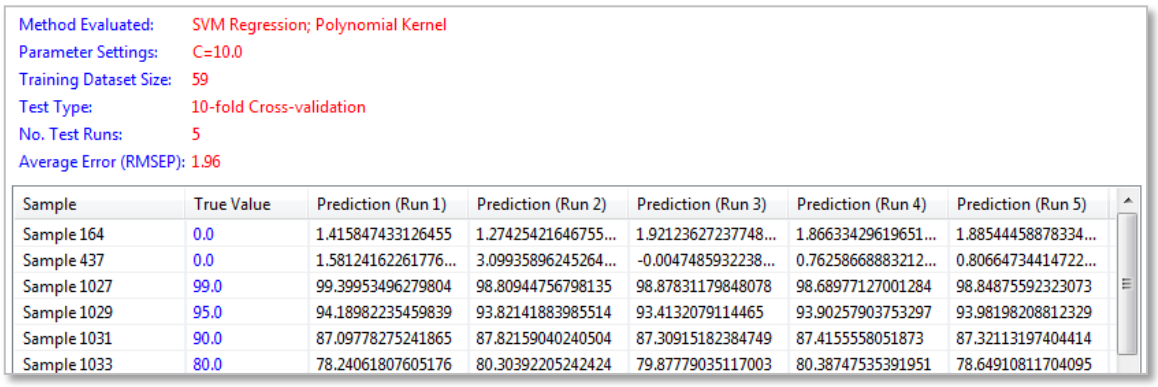
6.5 *Understanding and Reducing Quantification Error Values*

The purpose of a quantification model is of course to estimate the concentration of a target substance in a mixture. By measuring the model's performance on test data, we gain insight into how accurate we can expect the model's estimates of concentration are likely to be in the future, when it will be used to analyse substances for which the true concentration is unknown.

The evaluation results for quantification models are computed using the same procedure as for classification models, as was described earlier in Section 5.7.2: for each test spectrum, the model reports its estimated concentration, and these are compared with the true concentrations. However, for quantification models, the Root Mean Squared Error in Prediction (RMSEP) is computed.

Figure 85 shows a close-up view of a further evaluation that was performed when building this model, in this case performing 5 runs of 10-fold cross-validation. For this example, the RMSEP is 1.96%. This means that for this model, estimates of concentration are expected to be within 1.96% of the true value.

To understand how RMSEP is computed, look to the list of test spectra in the lower part of Figure 85. Under the heading of 'True Value', the correct answers are listed, while under the headings of 'Prediction', the model's answers in each cross-validation run are listed. Thus, for Sample 1027, the true concentration of acetonitrile is 99% and the model predicts a concentration of approximately 99.4%. For this sample, the error is the difference: 99.0% - 99.4% = -0.4%. Since errors may be positive or negative, all individual errors are squared, their mean is computed, and the square root of the mean is computed (hence: Root Mean Squared Error). Since this is done for 'predictions' or estimates on previously unseen data, it is termed the RMSE in Prediction.



Method Evaluated: SVM Regression; Polynomial Kernel
Parameter Settings: C=10.0
Training Dataset Size: 59
Test Type: 10-fold Cross-validation
No. Test Runs: 5
Average Error (RMSEP): 1.96

Sample	True Value	Prediction (Run 1)	Prediction (Run 2)	Prediction (Run 3)	Prediction (Run 4)	Prediction (Run 5)
Sample 164	0.0	1.415847433126455	1.27425421646755...	1.92123627237748...	1.86633429619651...	1.88544458878334...
Sample 437	0.0	1.58124162261776...	3.09935896245264...	-0.0047485932238...	0.76258668883212...	0.80664734414722...
Sample 1027	99.0	99.39953496279804	98.80944756798135	98.87831179848078	98.68977127001284	98.84875592323073
Sample 1029	95.0	94.18982235459839	93.82141883985514	93.4132079114465	93.90257903753297	93.98198208812329
Sample 1031	90.0	87.09778275241865	87.82159040240504	87.30915182384749	87.4155558051873	87.32113197404414
Sample 1033	80.0	78.24061807605176	80.30392205242424	79.87779035117003	80.38747535391951	78.64910811704095

Figure 85: Results of evaluating a quantitative model with 5x10-fold cross-validation

Whether or not an error rate of 1.96% is acceptable depends entirely on the specific application domain. For example, it may be reasonable when testing drug seizures in law enforcement, but not when testing materials on a production line. There are several strategies that may be used to improve the accuracy of quantification models; refer back to Section 5.7.3 for a discussion of them.

7 Batch Analyze

With *Batch Analyze*, you can analyze multiple spectra with multiple models:

- You specify a set of M spectrum files
- You select a set of N **Analyze IQ Lab** models
- **Analyze IQ Lab** automatically analyzes all spectra with all models, and output a table of M×N predictions
- You can optionally provide correct answers, in which case it will also output a table indicating whether each prediction is right or wrong, and overall performance metrics.

7.1 Performing a Batch of Analyses

To begin, choose the *Batch—Batch Analyze* menu item, as shown in Figure 86. This opens the *Batch Analyze* window, as shown in Figure 87.

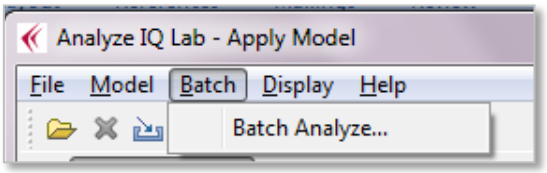


Figure 86: Starting a Batch of Analyses

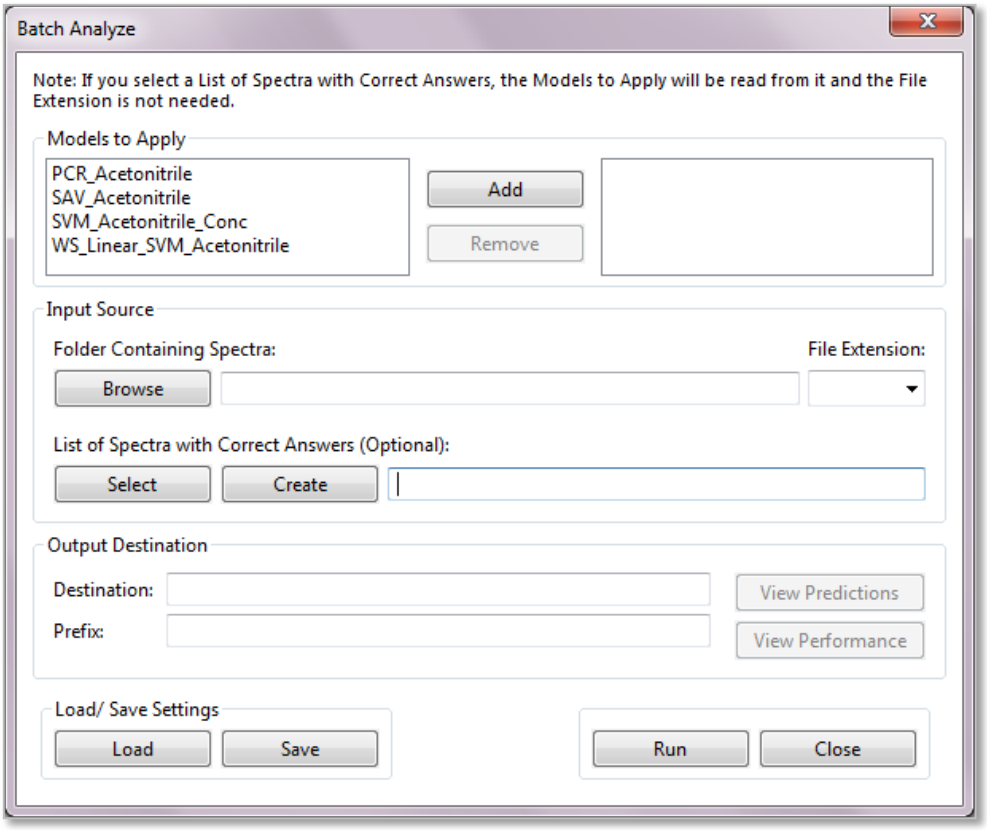


Figure 87: The *Batch Analyze* window

As described in the sub-sections that follow, there are two different modes in which you can run a batch of analyses:

1. Blind testing: correct answers are not provided
2. Model validation: correct answers are provided.

7.1.1 Running Analyses Without Providing Correct Answers

To run *Batch Analyze* without providing correct answers for the analyses, you do the following:

1. Select a set of models
2. Browse to the folder containing the spectra that you wish to analyze, and choose their file extension
3. Provide a prefix that will be used for storing the list of predictions (this defaults to the current date & time)
4. Press the *Run* button

An example of this is shown in Figure 88: four models are selected, the folder containing spectra is set to be the Analyze IQ Lab SampleData folder, the file extension chosen is AIQ, and the default prefix is chosen.

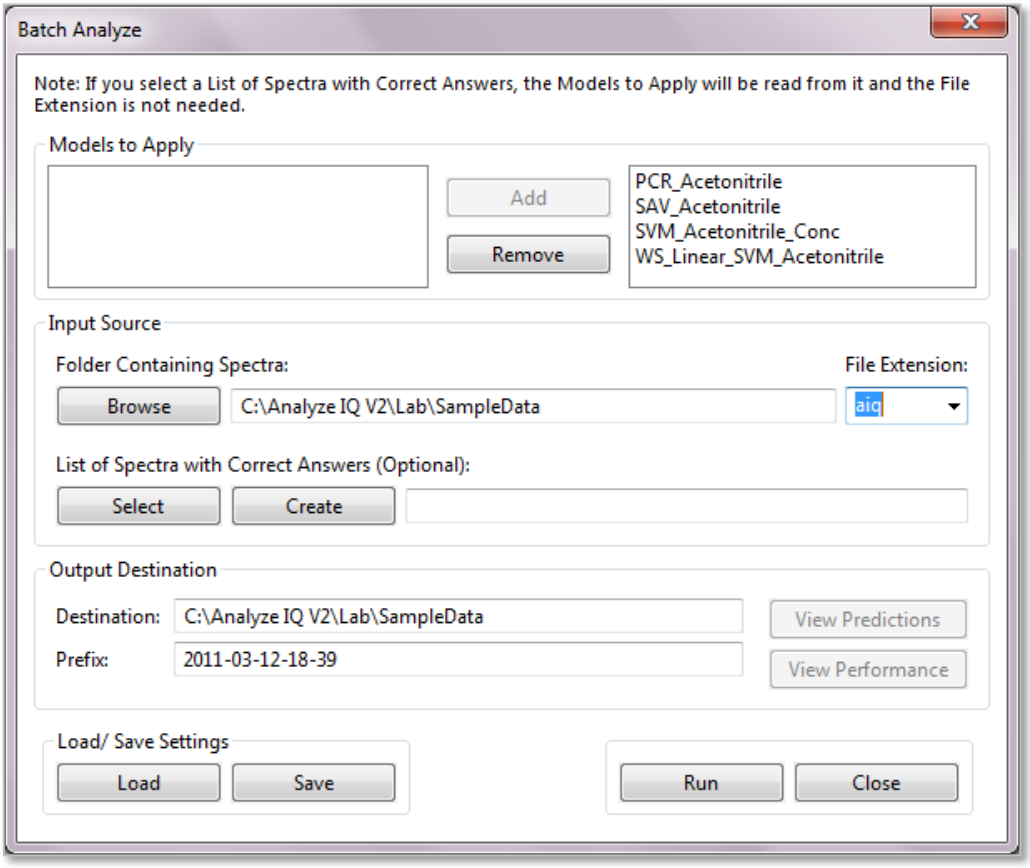


Figure 88: Running *Batch Analyze* without supplying correct answers

When you press *Run*, the models and spectra are loaded, and an information window is displayed, as shown in Figure 89, confirming that all spectra have been loaded and all models are ready for the analyses. If there were any problems, they would be displayed here.

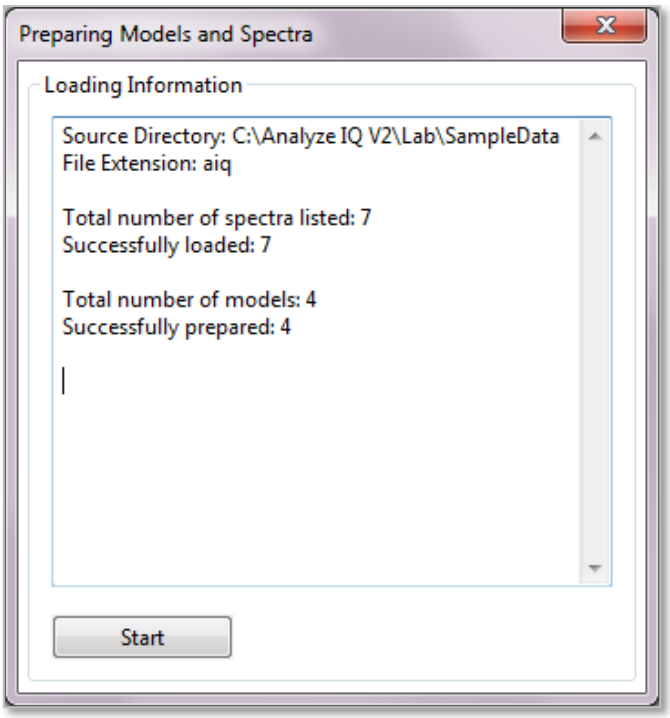


Figure 89: Confirmation that all models and spectra are ready for the analysis

If there are no problems, press the *Start* button to start the batch of analyses.

As the analyses are preformed, a progress bar is displayed (though if the number of models and files is not big, it will appear only briefly). When the analyses are completed, a message is displayed as shown below:

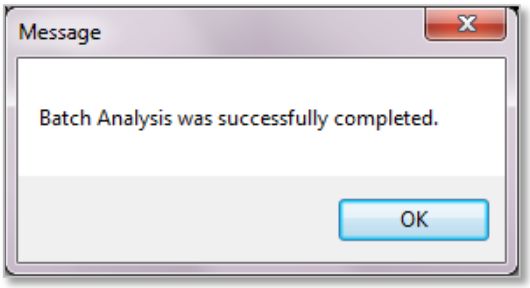


Figure 90: Message displayed when batch of analyses is complete

7.1.2 Viewing the Results

When the analyses are performed, the predictions generated by all models on all spectra are written to a file determined by the *Destination* and *Prefix:* it is stored in

the folder specified by *Destination*, and its name is ***Prefix*_Predictions.csv**. Therefore, for example, if the *Destination* is C:\Program Data and the Prefix is 2011-01-01, the Predictions.csv file will be C:\Program Data\2011-01-01_Predictions.csv.

The Predictions.csv file is structured like the following:

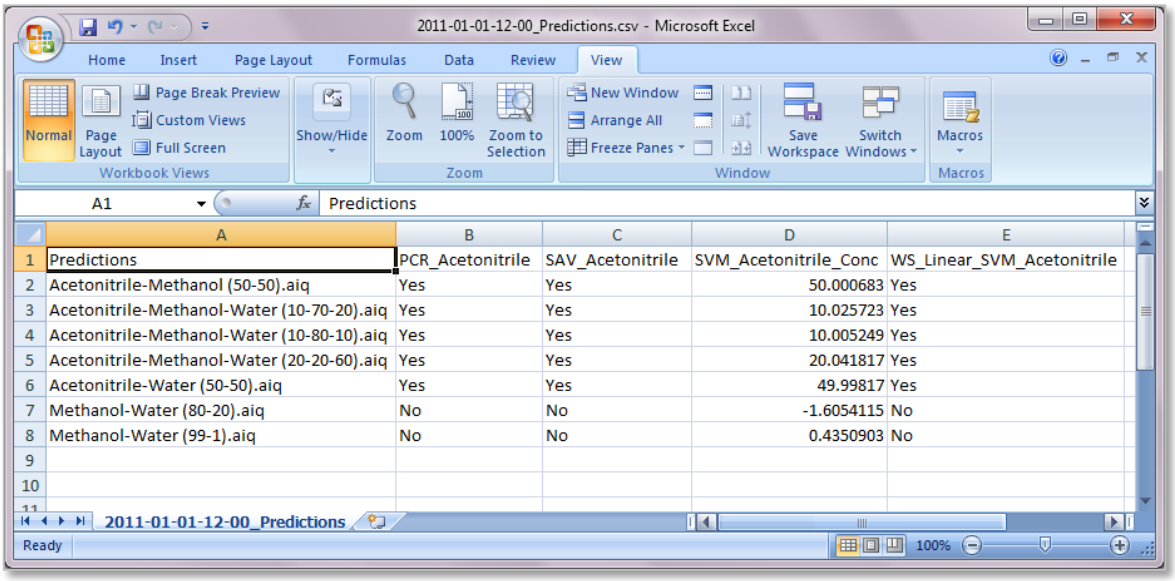
Predictions	Model1	Model2	Model3
Filename1.spc	Yes	Yes	51.0
Filename2.spc	No	Yes	80.5

Figure 91: Structure of Predictions.csv

Here, each Qualitative model produces a Yes/No result, and each Quantitative model produces a numerical result.

After the analyses are complete, you will notice that the *View Predictions* button, which was disabled before the run, becomes enabled (in fact, it is enabled whenever the *Destination* and *Prefix* specify a file that exists).

Clicking on *View Predictions* opens the Predictions.csv file in whatever program you normally use to edit CSV files (typically OpenOffice Calc or Microsoft Excel). This is shown in Figure 92. As can be seen in the figure, for each spectrum there is a result for every model. For qualitative models such as *PCR_Acetonitrile*, the result is 'Yes' or 'No', whereas for quantitative models such as *SVM_Aceotonitrile_Conc*, the result is a number.



	A	B	C	D	E
1	Predictions	PCR_Acetonitrile	SAV_Acetonitrile	SVM_Acetonitrile_Conc	WS_Linear_SVM_Acetonitrile
2	Acetonitrile-Methanol (50-50).aiq	Yes	Yes	50.000683	Yes
3	Acetonitrile-Methanol-Water (10-70-20).aiq	Yes	Yes	10.025723	Yes
4	Acetonitrile-Methanol-Water (10-80-10).aiq	Yes	Yes	10.005249	Yes
5	Acetonitrile-Methanol-Water (20-20-60).aiq	Yes	Yes	20.041817	Yes
6	Acetonitrile-Water (50-50).aiq	Yes	Yes	49.99817	Yes
7	Methanol-Water (80-20).aiq	No	No	-1.6054115	No
8	Methanol-Water (99-1).aiq	No	No	0.4350903	No
9					
10					

Figure 92: Predictions.csv open in Microsoft Excel

7.1.3 Running Analyses With Correct Answers Provided

If you provide correct answers, *Batch Analyze* will compare its predictions to the correct answers, and output a second file listing the performance of all models on all spectra, including overall error rates for each model.

To run Batch Analyze with correct answers provided, you do the following:

1. Browse to the folder containing the spectra that you wish to analyze
2. Select a list of spectra with correct answers (usually called Correct.csv, but you can give it any name)
3. Press the *Run* button

Note that in this use-case, you don't have to select the models to apply: they are read from the Correct.csv file that you specify. Figure 93 highlights the controls needed to supply the inputs in this case.

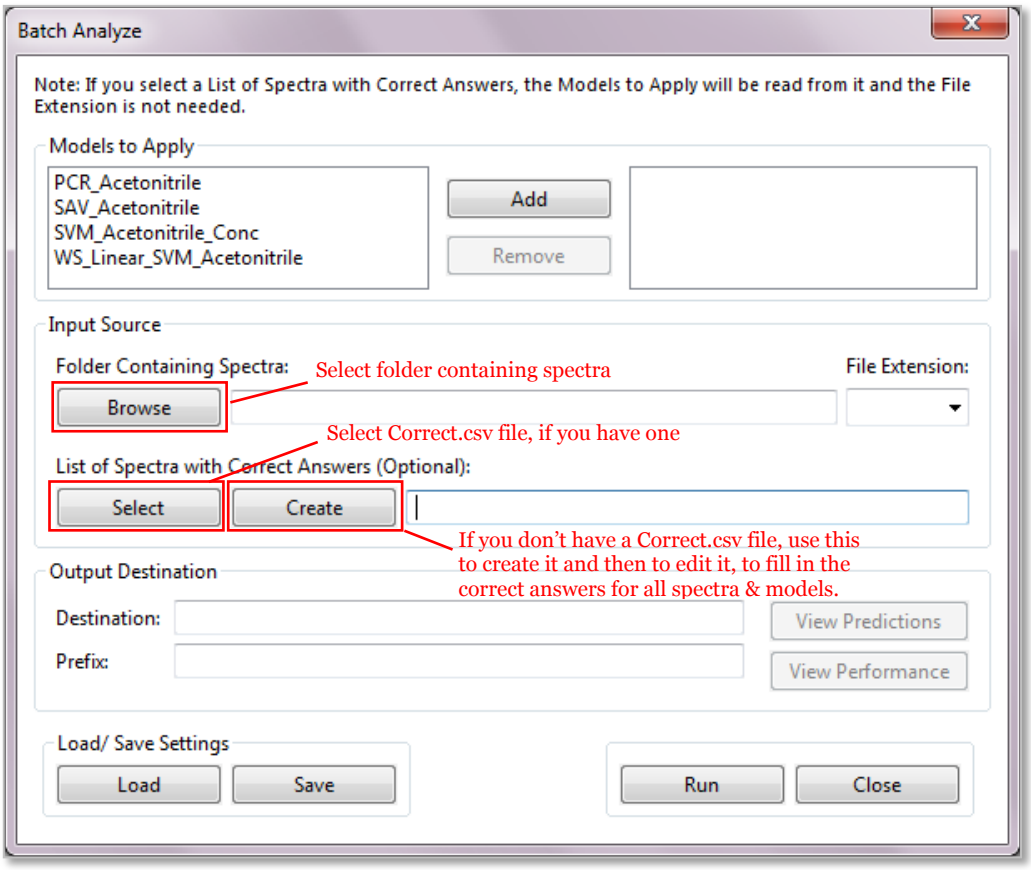


Figure 93: Controls needed for a run where you are providing correct answers

The structure of the Correct.csv file is identical the structure of the Predictions.csv file, as was shown in Figure 91. The only difference is that Correct.csv must contain the correct answers for all cases, whereas *Batch Analyze* will create Predictions.csv with model predictions which may or may not be correct.

If you do not have a Correct.csv file, you can select a list of models and press *Create* to create one, as indicated in Figure 93. The lists of models and files are written into it, but you must edit it to fill in the correct answers. After it is created, the *Create* button changes to *Edit*. When you press this, it is opened in the program that you usually use to edit CSV files (e.g. Microsoft Excel or OpenOffice Calc). When you are finished editing it, make sure to save it again in CSV format, **not** in Excel's XLS, XLSX or any other format.

Once you have the appropriate data selected, you run the batch of analyses by pressing the *Run* button, as described before. An information window is displayed to let you know how many files were parsed and if there were any problems in parsing them. If there are no problems, proceed by pressing *Continue*.

7.1.4 Viewing the Results

As was described in Section 7.1.2, when the batch of analyses is performed, the predictions generated by all models on all spectra are written to a file named ***Prefix*_Predictions.csv**.

In this case, an additional file is created, ***Prefix*_Performance.csv**. It is structured as follows:

Performance	Model1	Model2	Model3
Filename1.spc	Right	Wrong	0.01
Filename2.spc	Right	Right	-0.02
Model Error	0.0%	50.0%	-0.0158
Error Type	% Error	% Error	RMSE

Figure 94: Structure of Predictions.csv

The differences between Performance.csv, shown above, and Predictions.csv, which was shown in Figure 91, are:

- For each result, Performances.csv has an entry saying whether it is Right or Wrong (if qualitative) or giving the difference between the predicted and true values (if quantitative)
- Performances.csv has two extra lines of entries at the end: the first is the overall error rate for each model across all spectra tested, and the second is the type of error; for qualitative models, the percentage error is computed, and for quantitative models, the root mean square error is computed.

After the analyses are complete, the *View Predictions* button and the *View Performance* buttons are both enabled. As before, these open the corresponding files in your default program for editing CSV files.

8 Model Manager

This section describes the Model Manager feature in **Analyze IQ Lab.** There are three functions performed by the Model Manager:

1. Backing up Models
2. Restoring Models
3. Deleting Models

You can use these functions to archive models when you are finished a project and restore them later. You can also use them to transfer models between computers, because Analyze IQ model backup files can be copied and emailed.

You can pick any of the three functions from the Model menu on the top left hand side of the application window (see Figure 95).

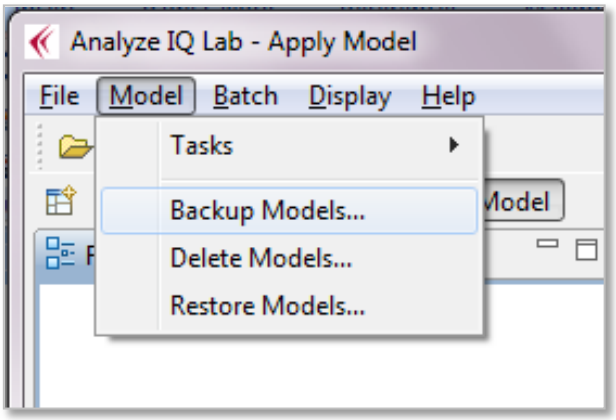


Figure 95: Model Manager menu items

8.1 Backing up and Deleting Models

The *Backup Models* and *Delete Models* menu commands both open the same window, because these two options are closely linked. If you wish to delete models from **Analyze IQ Lab**, it is recommended that you back them up first.

8.1.1 Overview of the Backup/ Delete Window

The window for backing up and deleting models, as shown in Figure 96, has three main elements:

1. Short help messages on the left side
2. List of available models within **Analyze IQ Lab** in the centre
3. The Information Area on the right side

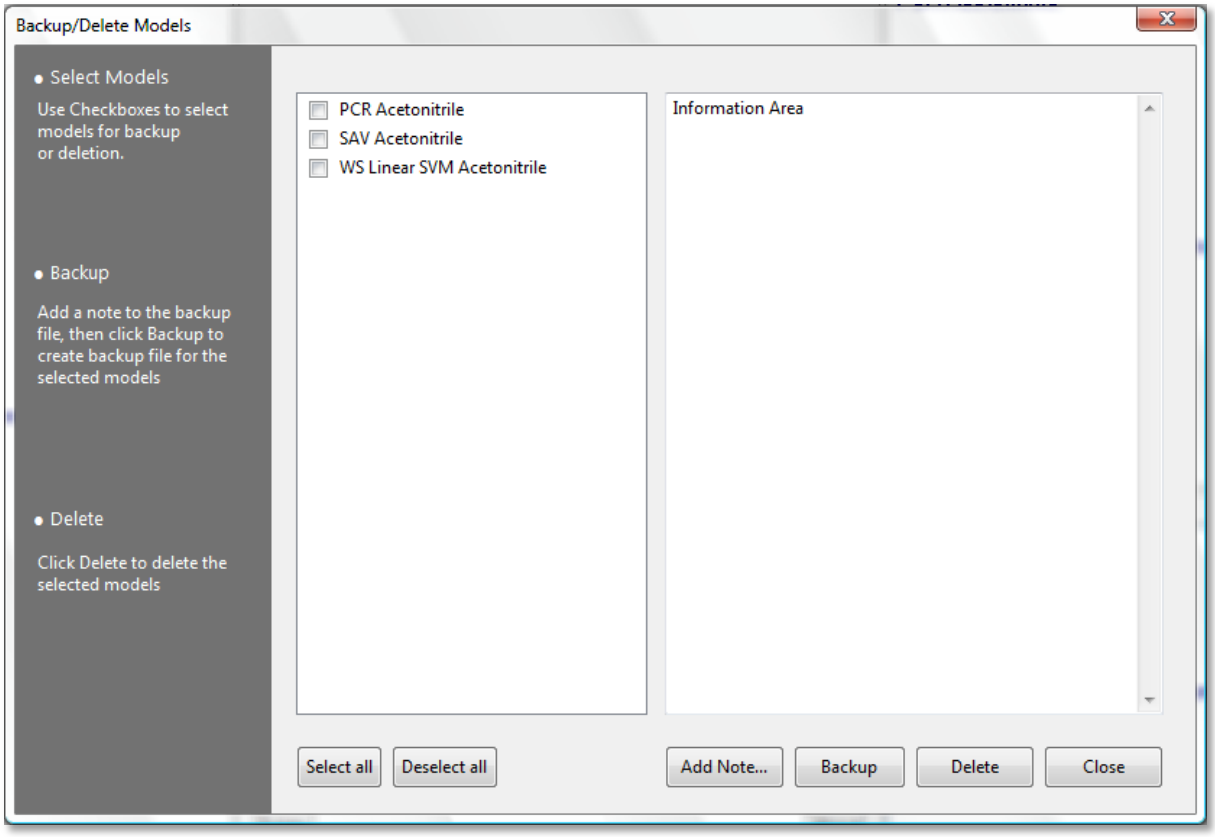


Figure 96: Backup/Delete window

Help messages

The help messages are a brief description of how to back up and delete models. They are there to assist you in the steps to take.

List of available models

The list of models in the centre of the window is loaded on starting the Model Manager. This list consists of all current models in your copy of **Analyze IQ Lab.**

The Information Area

This area displays the information associated with a model. If you click on any model in the list, its Description, Pre-Processing and Substances to be identified information will be displayed in the information area, so that you are aware what this model does and if you would like to back it up or delete it. Figure 97 shows this.

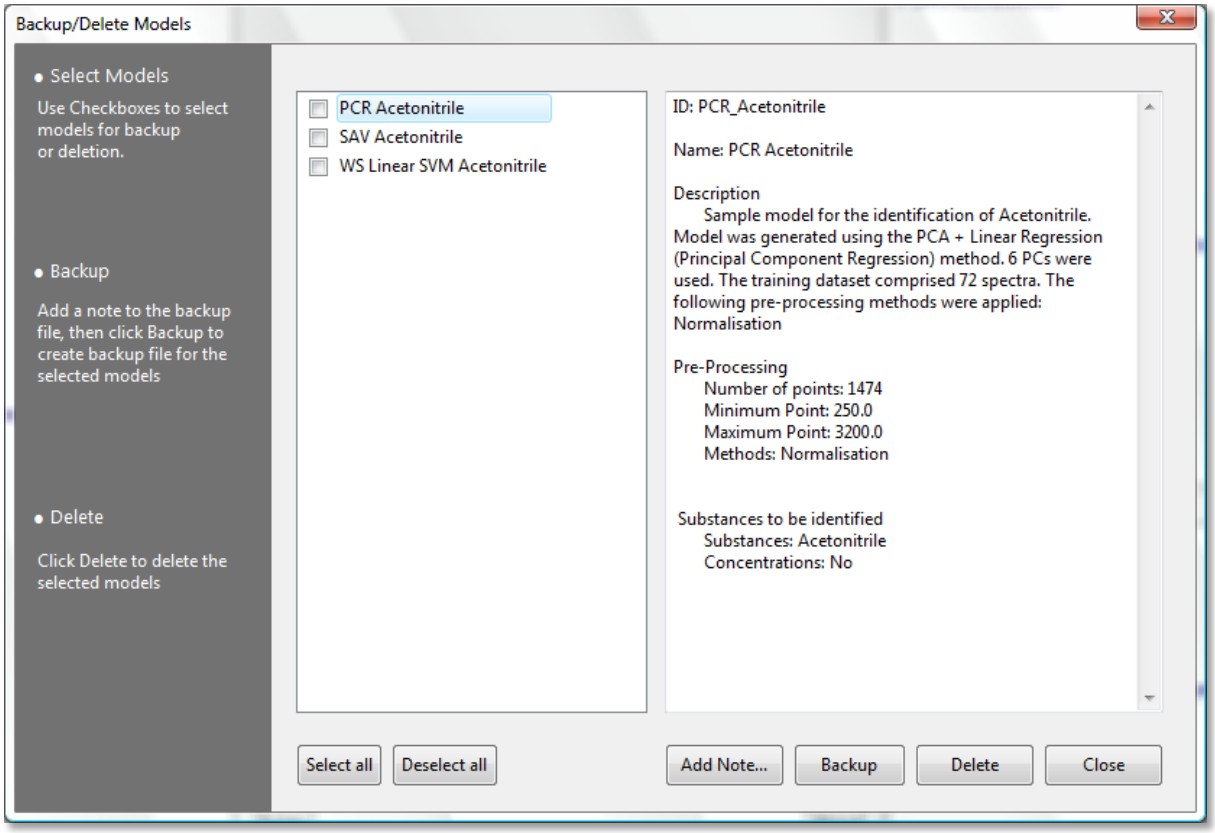


Figure 97: Model details displayed in the Information Area

8.1.2 Selecting Models to Backup or Delete

You can select models using the tick boxes to left of each model name (see Figure 98), and you can also use the *Select all* and *Deselect all* buttons. You can select as many models as you wish.

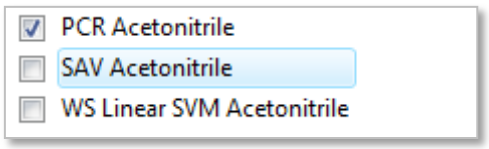


Figure 98: Checked and unchecked models

8.1.3 Adding a Note

If you are making a backup, you have the option to add a note to the backup file. To do this, click on the *Add Note* button; this opens a new window for you to enter your note, as shown in Figure 99. It can be helpful to add a detailed note to each backup, for future reference when restoring these models.

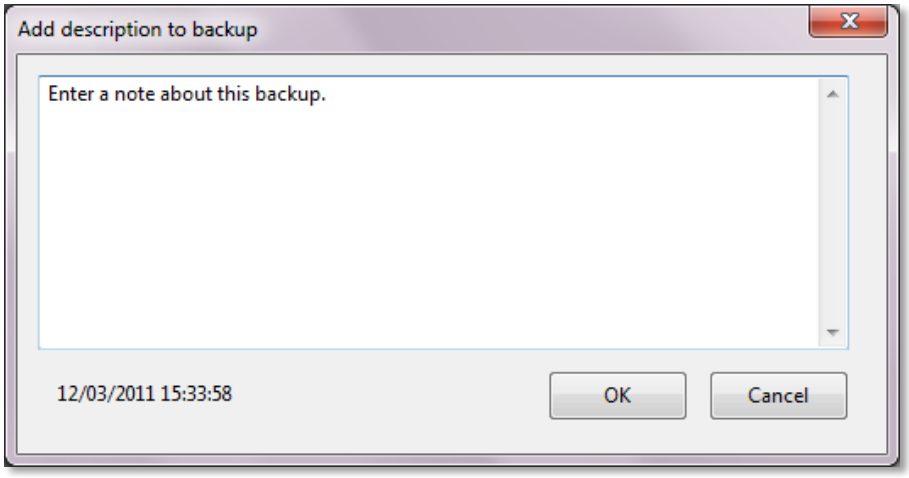


Figure 99: Adding a note to the backup file

8.1.4 Backing up selected models

Once you have selected the models you want to backup, and optionally added a note for the backup file, click on the *Backup* button. This opens a standard *Save As* dialog box to allow you to specify the name and location of your backup file (see Figure 100).

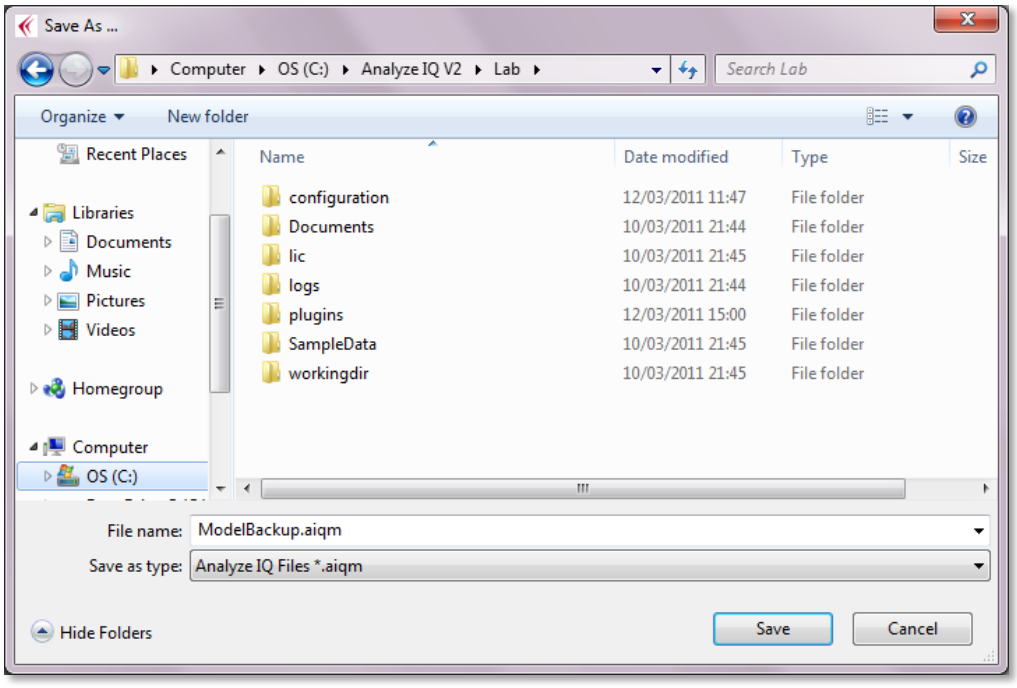


Figure 100: Picking a location to save your backup file

After doing this, a model backup file will be created, containing your selected models. When the backup is finished successfully, a message is displayed as shown in Figure 101.

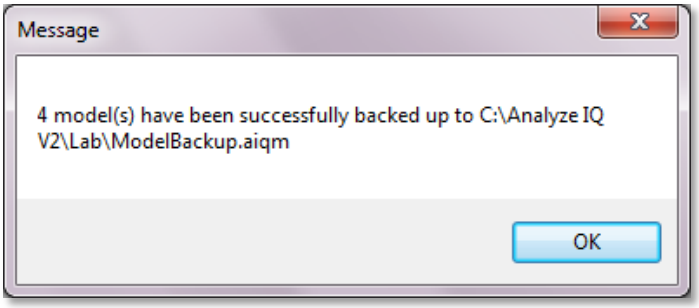


Figure 101: Notice to show the job was successful

8.1.5 Deleting selected models

Once you have selected the models you wish to delete, click on the *Delete* button. To ensure that you do not accidentally delete a model, you will be presented with a dialog box to ask you to confirm that you want to delete the models. If you delete models without having backed them up immediately beforehand, you will be asked if you want to delete them without backing them up first, as shown in Figure 102.

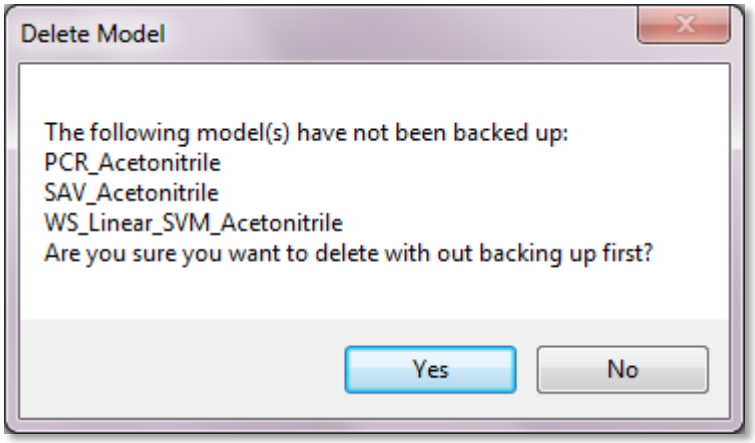


Figure 102: Notice that models have not been backed up

After you click Yes to the confirmation message, the models will be deleted and removed from **Analyze IQ Lab**, and a message will be displayed (see Figure 103).

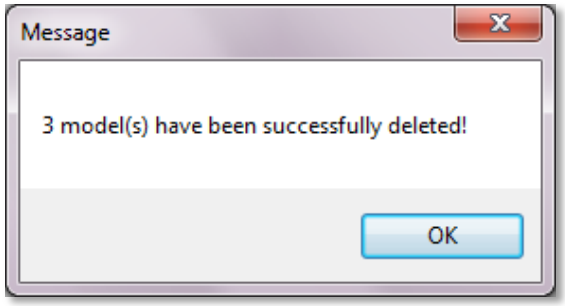


Figure 103: Notice that deletion was successful

8.2 Restoring Models

8.2.1 Overview of the Restore Window

The Restore window consists of four main elements (see Figure 104):

1. Short help messages on the left side
2. An area at the top to select a backup file
3. List of models contained in the selected backup file, in the middle
4. An Information Area that displays the note in the selected backup file

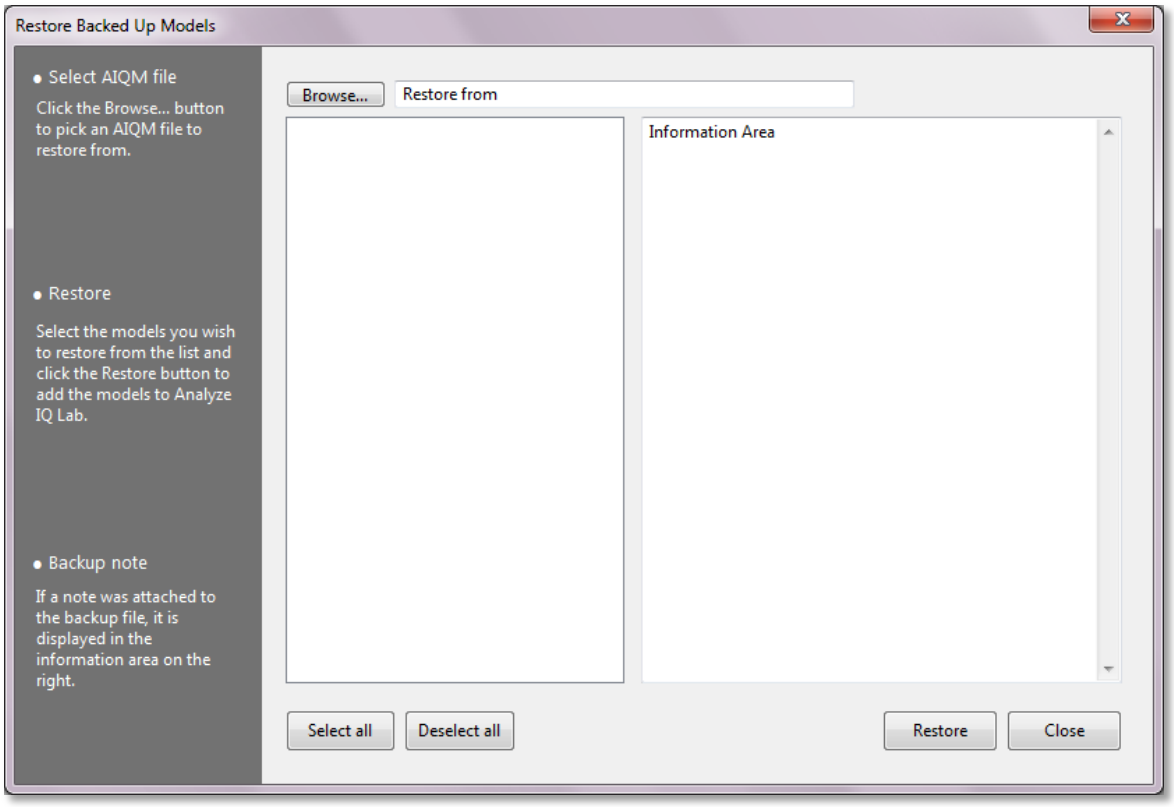


Figure 104: Restore models window

Help messages

The help messages are a brief description of how to restore models. They are there to assist you in the steps to take.

Picking your backup file

To the top of the window, you will see the Browse button with a line of text beside it. To pick a backup file to restore models from, click the Browse button. A window will open for you to pick the appropriate .aiqm file (see Figure 105).

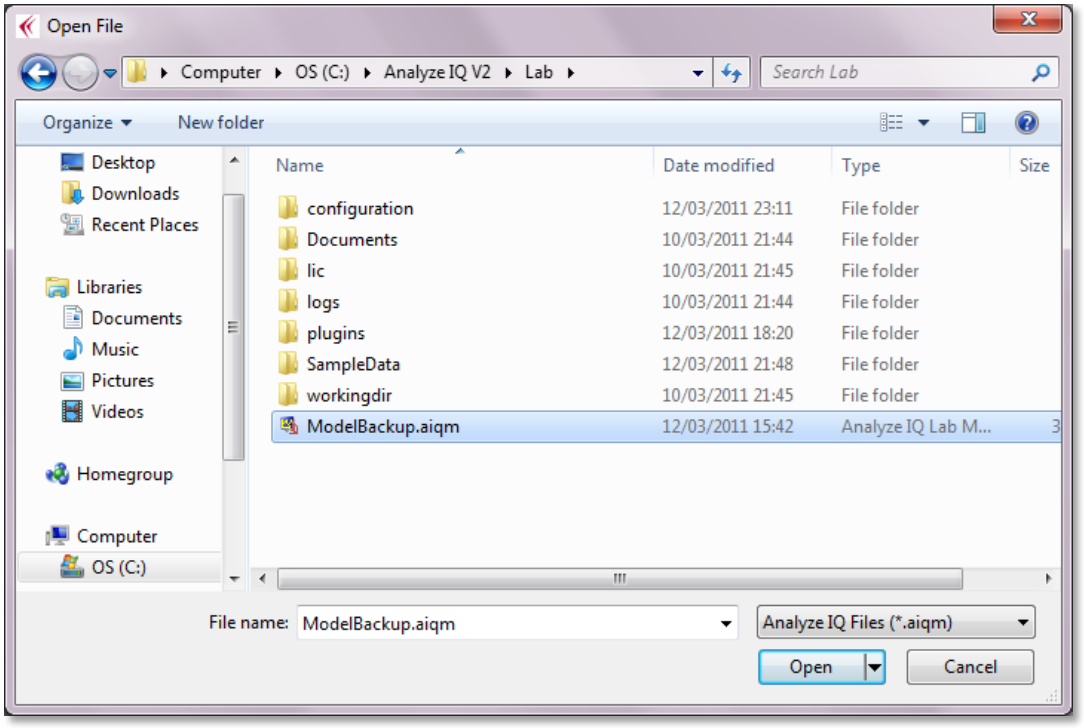


Figure 105: Selecting a backup file to restore from

List of models in a backup file

Once you have selected a backup file to restore from, the models contained in it are shown in the central pane.

Displaying the note from the backup file

When you select the file you want to backup file, as well as the list of models being displayed, the note that was added during backup is shown in the information area on the right side of the window.

8.2.2 Restoring the models

Using the checkboxes, you can choose to select some or all models to be restored, and then click on the *Restore* button. The selected models will be added into **Analyze IQ Lab**.

If you attempt to restore a model that has the same ID as a model already in **Analyze IQ Lab,** you will receive an error message to alert you to the problem. If the model is successfully added to **Analyze IQ Lab**, you will get a message to acknowledge that the selected models have been added.

9 Classification & Quantification Methods

This section describes the classification and quantification methods available in the current version of **Analyze IQ Lab**, along with a description of their corresponding parameters.

9.1 Classification Methods

9.1.1 Support Vector Machine (SVM)

The support vector machine is a powerful machine learning tool that is capable of representing nonlinear relationships and producing models that generalize well to unseen data. For binary classification, a linear SVM (the simplest form of SVM) finds an optimal linear separator between the two classes of data. This optimal separator is the one that results in the widest margin of separation between the two classes, as a wide margin implies that the classifier is better able to classify unseen spectra. To regulate overfitting, SVMs have a complexity parameter, C, which determines the trade-off between choosing a large-margin classifier and the amount by which misclassified samples are tolerated. A higher value of C means that more importance is attached to minimizing the amount of misclassification than to finding a wide margin model.

To handle non-linear data, kernels (e.g. radial basis function, RBF, polynomial or sigmoid) are introduced to map the original data to a new feature space in which a linear separator can be found. A kernel is effectively a function that compares two spectral samples and returns a value that indicates how similar the two spectra are. The SVM uses a kernel in the building of a model and when classifying test spectra. Different kernels may have a number of parameters associated with it.

In addition to the standard kernels, **Analyze IQ Lab** provides the unique Weighted Spectral Kernel (WS Kernel). The WS Kernel is a custom kernel that is designed specifically for the comparison of spectral samples. The WS Kernel utilises information about the spectrum of a target material (in pure form) that is to be identified or quantified. Modifications of the standard Linear and RBF kernels have been developed using this approach:

1. Weighted Spectral Linear Kernel: Select *SVM Classifier* and *Weighted Spectral Linear*.
2. Weighted Spectral RBF Kernel: Select *SVM Classifier*, *RBF Kernel* and *Weighted Spectral Distance*.

See Cristianini & Shawe-Taylor [2] for details on the SVM and the standard kernels (RBF, Polynomial and Sigmoid).

9.1.1.1 Parameters

C: Complexity parameter. A higher value of C means that more importance is attached to minimizing the amount of misclassification on the training set than finding a more generalized or wide margin model.

Kernel: Select one of the following kernels, each of which has its own set of parameters. (Default = Polynomial).

- ***Polynomial Kernel:***
  - ***Exponent:*** the degree of the polynomial kernel. (Default=1, which is specifies a linear kernel).
  - ***Use Lower Terms:*** If equal to zero, lower terms are not included in the polynomial kernel equation. If equal to one, the lower term is included. (Default = 0.)

- ***RBF Kernel:***
  - ***Sigma:*** the width of the RBF kernel. (Default = 0.01.)
  - ***Distance:*** the distance calculation used. A *Euclidean* distance or *Weighted Spectral Kernel* distance may be selected. (Default = Euclidean.) If *Weighted Spectral Kernel* distance is selected, the *Window Size* and *Pure Target Index* must be set; see Weighted Spectral Kernel below for a description of these parameters.

- ***Sigmoid Kernel:***
  - ***Phi:*** the offset value added to the inner product in the hyperbolic tangent function used in a Sigmoid kernel. (Default = 0.)
  - ***Gamma:*** the inner product coefficient in the hyperbolic tangent function used in a Sigmoid kernel. (Default = 0.75.)

- ***Weighted Spectral Kernel:***
  - ***Pure Target Index:*** the index, beginning at zero, of the pure target spectrum in the training dataset. (Default = 0, i.e. the first spectrum in the dataset corresponds to the sample containing 100% of the target substance.)
  - ***Window Size:*** when determining the similarity of two spectra, segments of this length are compared in order to perform on overall comparison of the profiles of the two spectra. (Default = 10.)

9.1.2 Principal Component Analysis (PCA) Classifiers

Two characteristics of spectral data that can lead to poor accuracy in prediction are the high dimensionality and high collinearity of the data. PCA is a classical statistical method for transforming attributes of a dataset into a new set of uncorrelated attributes called principal components (PCs). The key idea is that this allows reduction of the data to a smaller number of dimensions, with low information loss, simply by discarding some of the PCs. Each PC is a linear combination of the original inputs and each PC is orthogonal, which therefore eliminates the problem of

collinearity. **Analyze IQ Lab** uses the Non-linear Iterative Partial Least Squares (NIPALS) PCA method, which is described by Geladi & Kowalski [3].

Principal Component Regression (PCR) is a chemometric technique that is widely used for the classification of spectral data. PCR is a two-step multivariate regression method, in which PCA of the data is carried out in the first step. In the second step, a multiple linear regression between the PC scores obtained in the PCA step and the predictor variable (the Yes/No label) is carried out.

To select PCR for classification in **Analyze IQ Lab**, go to the *PCA* section and select *Linear Regression* under it(see Figure 106).

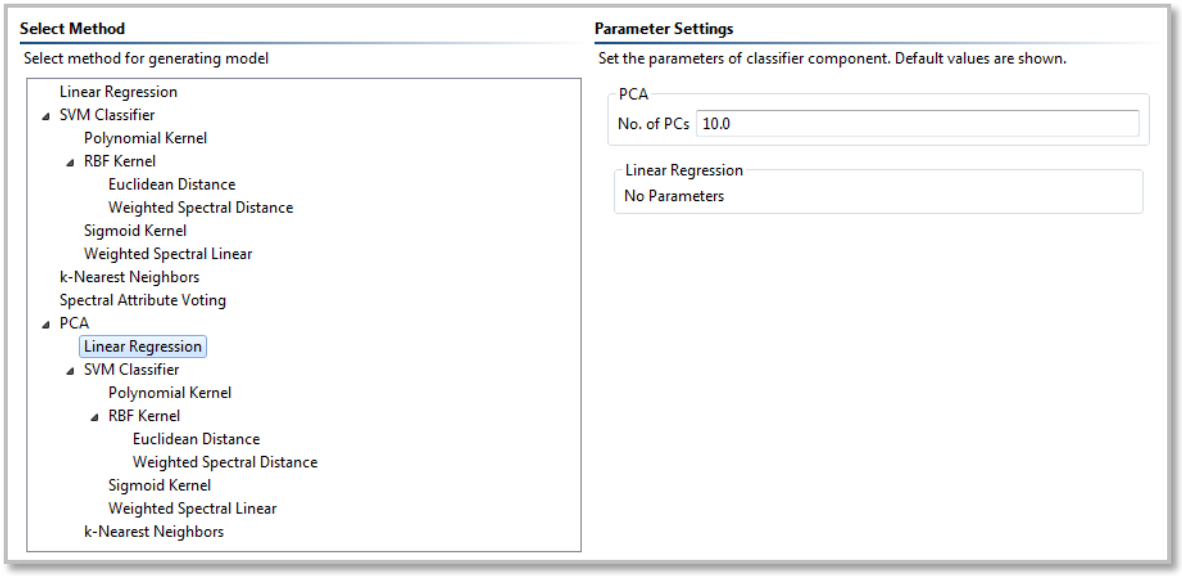


Figure 106: Selecting PCR (PCA + Linear Regression) in Analyze IQ Lab

Analyze IQ Lab also allows for the combination of the PCA transformation technique with two other classification methods: the SVM Classifier, including multiple variants thereof (see Section 9.1.1 for details); and k-Nearest Neighbors (see Section 9.1.3 for details). As with PCR, two steps are involved: reduce the data to a set of principal components; build classification model using the principal component scores as input to the selected base classifier. When using a PCA-based classifier, the *No. of PCs* parameter must be set in addition to any parameters that are required for the base classifier.

9.1.2.1 Parameters

***No. of PCs*:** The number of principal components retained for building the classification model. (Default = 10.)

Base Classifier: Select Linear Regression, k-Nearest Neighbors or SVM Classifier. (Default = Linear Regression.)

Base Classifier Parameters:

- No parameters for Linear Regression.
- See Section 9.1.1 for SVM classifier
- See Section 9.1.3 for k-NN

9.1.3 k-Nearest Neighbors

k-Nearest Neighbors (k-NN) is a learning algorithm that classifies a test sample by firstly obtaining the class of the *k* samples that are the closest to the test sample, where *k*, the number of neighbors, is a value chosen by the user. The majority class of these nearest samples (or the nearest single sample when k=1) is returned as the prediction for that test sample. The Euclidean distance measure is used to compute nearness. In practical terms, each spectrum is compared to every other spectrum in the dataset. At each spectral data point, the difference in intensity between the two spectra is measured (distance). The sum of the squared distances for all the data points across the full spectrum gives a measure of how close the spectra are.

9.1.3.1 Parameters

No. of Neighbors: The number of nearest neighbors used in the determination of the class of a test sample. (Default = 3.)

9.1.4 Spectral Attribute Voting (SAV)

SAV is a proprietary Analyze IQ machine learning technique developed specifically for spectral analysis. It is an ensemble-based approach that accounts for the locally correlated nature of spectra and also uses the shape of the spectral profile (i.e. peaks and troughs) in the generation of models.

9.1.4.1 Parameters

No Parameters.

9.1.5 Linear Regression for Classification

With this technique, a multiple least squares linear regression between the variables (spectral intensities) of the spectral dataset and the predictor variable is carried out. The predictor variable is a value that is chosen to represent the presence or absence of the target in a sample. **Analyze IQ Lab** uses the value 1 to represent target presence and −1 to represent target absence. When classifying a test sample, if the regression model returns a value that is closer to 1 than −1, the sample is predicted to contain the target; otherwise, the test sample is classified as non-target.

9.1.5.1 Parameters

No Parameters.

9.2 Quantification Methods

9.2.1 Linear Regression

Given a target value (such as the concentration of a compound in a mixture) and a set of attribute variables (such as spectral intensities), the goal of Linear Regression is to find an equation that describes the target in terms of the attributes. For a given training set, attribute weights are calculated which minimize the squared error between the predicted and actual target values. If an attribute has a weight of zero then it will have no effect on the outcome. The regression model returns a prediction of the target value for an unknown sample. Using Linear Regression on its own can perform very poorly if the number of samples is **not greater than** the number of dimensions per sample. Therefore, in spectral analysis, Linear Regression is usually used with dimension reduction techniques such as PCA (see Section 9.2.4).

9.2.1.1 Parameters

No Parameters.

9.2.2 Support Vector Machine (SVM) Regression

The SVM approach can be applied to regression, maintaining the maximal margin aspect of the SVM classifier. SVM regression is a form of linear regression that is derived from the SVM classification technique. As with standard least squares regression, SVM regression minimizes the sum of the square errors between the predicted and actual values (e.g. concentrations) for each sample in the training dataset. In addition, SVM regression ignores errors that fall within a certain band around the target function, in order to produce a regression model that better generalizes to unknown data. See Cristianini & Shawe-Taylor [2] for more details on the SVM algorithm.

9.2.2.1 Parameters

See SVM Classifier parameters (Section 9.1.1).

9.2.3 k-Nearest Neighbours Regression

This algorithm works as described in Section 9.1.3, except that in this case it is not the majority class that is being predicted, but the target value of the nearest neighbors, and it is computed by averaging the corresponding values of the nearest neighbors. For example, if we are using the 3 nearest neighbours of the test sample

and their target values are 50, 60 and 70 respectively, the prediction that is returned will be an average of these values. The algorithm will return 60 as the target value prediction for the example above.

9.2.3.1 Parameters

See k-Nearest Neighbors parameters (Section 9.1.3.1)

9.2.4 Principal Component Regression (PCR)

In this method, two steps are involved. Firstly, the data is reduced to a set of principal components, as described in Section 9.1.2, using PCA. A regression model is then built using the principal component scores as input to the selected base regression algorithm.

9.2.4.1 Parameters

***No. of PCs*:** The number of principal components retained for building the regression model. (Default = 10.)

Base Regression Algorithm: Select Linear Regression, SVM Regression or k-Nearest Neighbors Regression. (Default = Linear Regression.)

Base Regression Algorithm Parameters:

- No parameters for Linear Regression.
- See Section 9.2.2 for SVM Regression
- See Section 9.2.3 for k-NN Regression

9.3 Choosing Algorithms and Selecting Parameters

There is no "one size fits all" algorithm available for carrying out chemometric analysis. The success of a particular algorithm will always depend on the structure of the data in question and appropriate parameters being set. We will now take a closer look at the specific details of some of the algorithms available in **Analyze IQ Lab**.

9.3.1 Linear Regression

Linear Regression can give very poor predictions if the number of samples is smaller than the number of dimensions per sample. This is often the case in chemometrics and therefore this algorithm is not often used on its own. It can, however, be worth using if the dataset has relatively low dimensionality. Its other main drawback is that it assumes linearity. It has the advantage of being a very simple method with no hidden complexities and with no parameters to set. It can provide good results when used in conjunction with a dimension reduction technique such as Principal Component Analysis, provided that the linearity assumption is reasonable for a dataset.

9.3.2 Support Vector Machine (SVM)

SVMs can provide very good results and can handle high dimensional data. However, there is an element of model selection involved. An appropriate kernel must be chosen with the right kernel parameters and an effective value for the complexity parameter C.

The C parameter trades off minimizing errors on the training set versus finding a classifier that has better generalization. This value will depend inversely on the scale of the data. Setting the value to 1 is a good starting point if the data are normalized. This value should be reduced it if the scale of the data is larger. The value should be adjusted in powers of 10, for example: 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000. Increasing C will reduce the level of misclassification on training data where as reducing it will increase the generality of the model that is generated.

It is always worth considering simpler forms of a model before moving on to more complex ones. A simple measure of complexity could be to check how many parameters there are to be set. However, you should be aware that some types of kernel can have greater or less complexity depending on their parameters. In particular, increasing the exponent in the Polynomial kernel for SVM increases the complexity of the decision boundary that it can represent. Therefore, the exponent should first be set to 1 when experimenting with a variety of C values before moving onto setting the exponent to 2 etc.

9.3.2.1 Setting Kernel Parameters

Polynomial Kernel	
Exponent	This parameter corresponds to the degree of the polynomial. If this is set to 1 it means that it is linear. The usual range for this parameter is from 1 to 3. Increasing the polynomial will increase the complexity of the model. It will be able to fit more complex data but there will a greater risk of over-fitting.
Use Lower Terms	This is a binary decision. For an exponent that is greater than 1, it is a decision of whether or not to include lower polynomial terms. It is recommended that these are included.

RBF Kernel	
Sigma	This parameter is the width of the RBF kernel. The 'natural' setting for this depends on the scale of the data in question. It is recommended to begin with a value such as 0.01 for normalized data. This value could then be doubled or halved in order to evaluate what effect it has. A common range would include 5 steps (doubling/halving each time) starting from 0.01.
Distance	The standard distance metric that is used in the RBF kernels is Euclidean. Weighted Spectral Kernel distance can also be chosen.

Sigmoid Kernel	
Phi	This value is the offset added to the inner product in the hyperbolic tangent function used in a Sigmoid kernel. This value should be varied in powers of 10, typically in a range from 10^{-5} to 10^{5}.
Gamma	This value is the inner product coefficient in the hyperbolic tangent function used in a Sigmoid kernel. This value should be varied in powers of 10, typically in a range from 10^{-5} to 10^{5}.

Weighted Spectral Kernel	
Pure Target Index	This value is the index (from 0) of the pure target spectrum in the training dataset.
Window Size	Segments of this length are compared in order to perform on overall comparison of the profiles of two spectra. A good value is approximately the number of points in one peak.

9.3.3 k-Nearest Neighbors (k-NN)

The k-NN algorithm can be particularly relevant when the concept to be classified is not coherent (e.g. "toxic/non-toxic"). The algorithm works best when it is dealing

with lower-dimensional data. For this reason, when carrying out spectral analysis it is recommended that you use k-NN in conjunction with a dimension reduction technique such as Principal Component Analysis (PCA).

k-NN is sensitive to pre-processing of the data. It is important to note that similar spectra will have large distances between them if there is a baseline offset or if the scales of the samples are different. The pre-processing techniques of Normalization, Equal Area Scale or Standard Normal Variate are recommended when using the k-NN algorithm.

There is one parameter to be set which is the number of nearest neighbors to use (No. of Neighbors). Simply selecting a value of 1 corresponds to doing a direct search and using the label of the closest match as the label of the unknown sample. In binary classification problems it is helpful to choose the No. of Neighbors as being an odd number as this avoids tied votes. Typical values for the No. of Neighbors are: 3, 5, 7 or 9. If the No. of Neighbors is a large value, the effect of noise on the classification is reduced, however, this makes the boundaries between the classes (or differing target values) less distinct.

9.3.4 Spectral Attribute Voting (SAV)

This method is only used for classification problems. It is an ensemble method that involves constructing small local classifiers for each section of the spectra. All the classifiers that are better than random become part of the final ensemble in which a vote is carried out to arrive at the final decision. Ensemble methods have good theoretical properties and, as long as individual classifiers are independent and better than random, the ensemble as a whole will be better than the individuals. This is because individual errors tend to cancel each other out. This method works well with high dimensional data and assumes the presence of local correlations (as found in spectral data). There are no parameters to be set so, if this method works well, then it is very easy to use.

9.3.5 Principal Component Analysis (PCA)

This dimension reduction technique finds a transformation of the data that accounts for maximum variance in its first component, and then continues to account for additional variance in subsequent components. There is one component for each original dimension. This method is used as a data pre-treating step before the use of other analysis methods.

PCA should be considered when there is a requirement to capture most of the variance in the data, while greatly reducing the dimensionality. It is particularly useful with methods that do not work well with high-dimensional data, such as, Linear Regression and k-Nearest Neighbors. It should be noted that this technique is 'blind' to the data labels and, in some situations; the major variance may not be associated with the question of interest. For example, it may not work well if it is required to measure small variations in an active ingredient.

10 Technical Support, Sales and Services

For assistance and technical support queries relating to **Analyze IQ Lab**, please go to the **User Area** of the Analyze IQ website, http://www.AnalyzeIQ.com, or send an email message to support@AnalyzeIQ.com.

For information on extending an evaluation license or purchasing a full license for **Analyze IQ Lab**, please contact our Sales Department by sending an email message to sales@AnalyzeIQ.com.

In addition to software sales and technical support, Analyze IQ Limited provides the following services for customers:

- Software training
- Data validation and analysis
- Chemometric model development
- Custom software development, including custom analysis solutions
- Integration of Analyze IQ chemometric models in third-party software
- Development of Instrument Interfaces
- OEM licensing

For information on these services, please contact our Sales Department by sending an email message to sales@AnalyzeIQ.com.

11 References

1. Savitzky, A. & Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36, 1627–1639.
2. Cristianini, N. & Shawe-Taylor, J.S. (2000). An Introduction to Support Vector Machines. Cambridge University Press.
3. Geladi, P. & Kowalski, B.R. (1986). Partial Least Squares: A Tutorial. Analytica Chemica Acta, 185, 1–17.
4. Dhanoa, M.S., Lister, S.J., Sanderson, R. and Barnes, R.J. The link between Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) transformations of NIR spectra, Journal of Near Infrared Spectroscopy. Vol. 2 1994.

12 Patent, Copyright & Trademark Notices

This software is protected by international patent filings.

Analyze IQ, Analyze IQ Lab, and Spectra Manager are trademarks of Analyze IQ Limited. Other product names and brand names are trademarks or registered trademarks of their respective owners.

No part of this manual may be copied or reproduced in any form without prior written consent from Analyze IQ Limited.

Copyright © Analyze IQ Limited, 2008-2014. All Rights Reserved.